編碼：RES-105-03

行政院主計總處委託研究

主計資料大數據分析之研究

受委託單位：臺北醫學大學管理學院大數據研究中心

研究主持人：謝邦昌 教授

協同主持人：江志民 助理教授

中華民國 105 年 12 月

摘要

政府財務之處理，必須透過行政、主計、公庫及會計審核等組織，以達成相互配合及分立制衡之功能，而主計功能的發揮，更是促使國家有限資源合理分配及有效運用的重要關鍵。因此，若能整合歲計、會計、統計資料，運用大數據分析技術，選擇相關研究議題，建構示範案例及資料模型，產生對施政有參考價值之研究結果，增進主計業務相關政策規劃應用能力。本研究目的是期能藉由完成「公務檔案串連跨資料整合、「歲計資料整合應用」、「資料平台跨域整合」及「客製化 Text Mining」等 4 項關鍵核心議題的示範案例，經運用大數據分析技術，整合歲計、會計、統計資料，以及資料樣態特徵萃取後，建構出預測與資料加值應用模型之研究成果及後續研究建議。

基於上述四項關鍵核心議題，分別發展出子議題如下：在「村里常住人口推估模型」議題採用內政部人口資料、主計總處人口普查資料與全國達康常住人口資料，進行村里常住人口推估模型之建立。透過與普查年資料銜接，以函數映射方法估計到一定期間各縣市村里的人口常住資料，掌握調查時點最新人口數量及分布，增進後續分析之可行性與價值。在「歲計估算模型」議題建立歲計預算模型，不僅可以即時充分反映政府整體的財政狀況，減少不必要的資源浪費之情形之外，也可以用來監測政府預算，通過建立基本建模的數據庫來分析預算偏離的情況。在「科技跨域整合資料模型」議題採用第 12 次（民國 100 年）工商及服務業普查做為主資料庫，以及民國 100 年的第三次產業創新調查作為輔助資料庫，用函數映射（Functional Mapping）方式將創新調查資料映射至工商普查的資料庫中，以擴增工商普查資料庫之應用價值及提升科技跨域資料之整合分析能力。在「客製化 Text Mining」議題部份，本計畫針對主計總處需求，開發一套客製化的 Text Mining 平台，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，依業務及政策動態需求，隨時協助提供施政方針之參考。

本研究提出各模組之發展程序與應用雛型，示範如何基於大數據分析技術，整合歲計、會計、統計資料進行主計資料大數據分析之可行作法，將有利於可行方案評估與縮短後續系統開發時程，有助於降低開發成本及嘗試錯誤之風險。最後，本研究彙整說明各類模型未來推行之建議。

關鍵字：大數據、主計、跨域整合、函數映射、歲計估算、常住人口、文字探勘、輿情分析

目錄

表目錄

圖目錄

研究團隊

研究團隊	姓名/職稱
督導研究人員	丁台怡/台北醫學大學大數據研究中心顧問
參與研究人員	陳郁婷/台北醫學大學大數據研究中心秘書
參與研究人員	張葦憶/台北醫學大學大數據研究中心秘書
參與研究人員	曾馨頡/台北醫學大學大數據研究中心秘書
研究助理	蕭凱元/台北醫學大學大數據研究中心研究助理
研究助理	李佳儒/台北醫學大學大數據研究中心研究助理
研究助理	謝明穎/台北醫學大學大數據研究中心研究助理
研究助理	吳格非/台北醫學大學大數據研究中心研究助理
研究助理	王興弘/台北醫學大學大數據研究中心研究助理

第一章、研究背景及目的

第一節、研究主旨與目標

政府財務之處理，必須透過行政、主計、公庫及會計審核等組織，以達成相互配合及分立制衡之功能，而主計功能的發揮，更是促使國家有限資源合理分配及有效運用的重要關鍵。因此，若能從主計領域數字資料中，經過資料分析、洞察，發覺有用的資訊，以作為大數據分析最佳範例，將可有效提升主計業務功能發揮。

大數據分析是一門新的資料探索方式，分析結果有助於瞭解整體趨勢、預測未來，能達洞察機先、防患未然效益，可協助政府進行前瞻施政規劃，優化政府施政。為增進主計業務決策品質與效能，強化施政計畫擘劃能力，爰擬透過本研究結果作為後續制定主計業務相關大數據分析應用之規劃參考。具體而言，本研究計畫目的在於研析大數據在主計領域分析規劃應用之可行性，提出可行性應用架構與推行方法，並完成以下四項目標。

一、蒐集及分析國外先進國家有關主計領域大數據分析規劃應用、分析 方法及技術工具等發展趨勢，以及國內各界資料需求、關鍵核心議題，作為未來發展大數據分析與應用之參考。

二、整合歲計、會計、統計資料，運用大數據分析技術，選擇相關研究議題，建構示範案例及資料模型，產生對施政有參考價值之研究結果，增進主計業務相關政策規劃應用能力。

三、善用大數據分析技術，結合主計總處既有厚實之業務能量與經驗，強化現有施政決策分析方式，有效提升主計業務決策品質與效能。

四、藉由本研究計畫示範案例進行雛型平台實作，並提出可行性作法建

議，研究成果可作為推動大數據應用與發展之依據。

第二節、研究方法與過程

本研究計畫係從大數據（Big Data）思維與角度出發，分析技術則以機器學習、資料採礦 CRISP-DM 及統計等演算法為主，運用資料映射（Data Mapping）技術將政府其他資料庫如：政府歲計會計資訊管理系統（GBA）、臺閩地區工商及服務業普查資料、全民健保資料檔資料、財政部營利事業資料、科技部產業創新調查資料...等，以及外部資料如：全國達康資料、全國意向資料...等，進行跨機關及跨領域資料串連，以利後續資料採礦分析之用。

本計畫的整體研究過程如圖 1 所示，前節已確認四項主要研究目標，將蒐整國外先進國家有關主計領域大數據分析之研究（含企業及政府面向）、規劃與分析方法、技術工具及應用案例等發展趨勢相關文獻，並於第二章針對大數據分析概述、國外主計領域大數據分析研究現況與發展趨勢、大數據分析規劃及方法、技術工具及應用案例現況與發展趨勢分節進行探討評析。

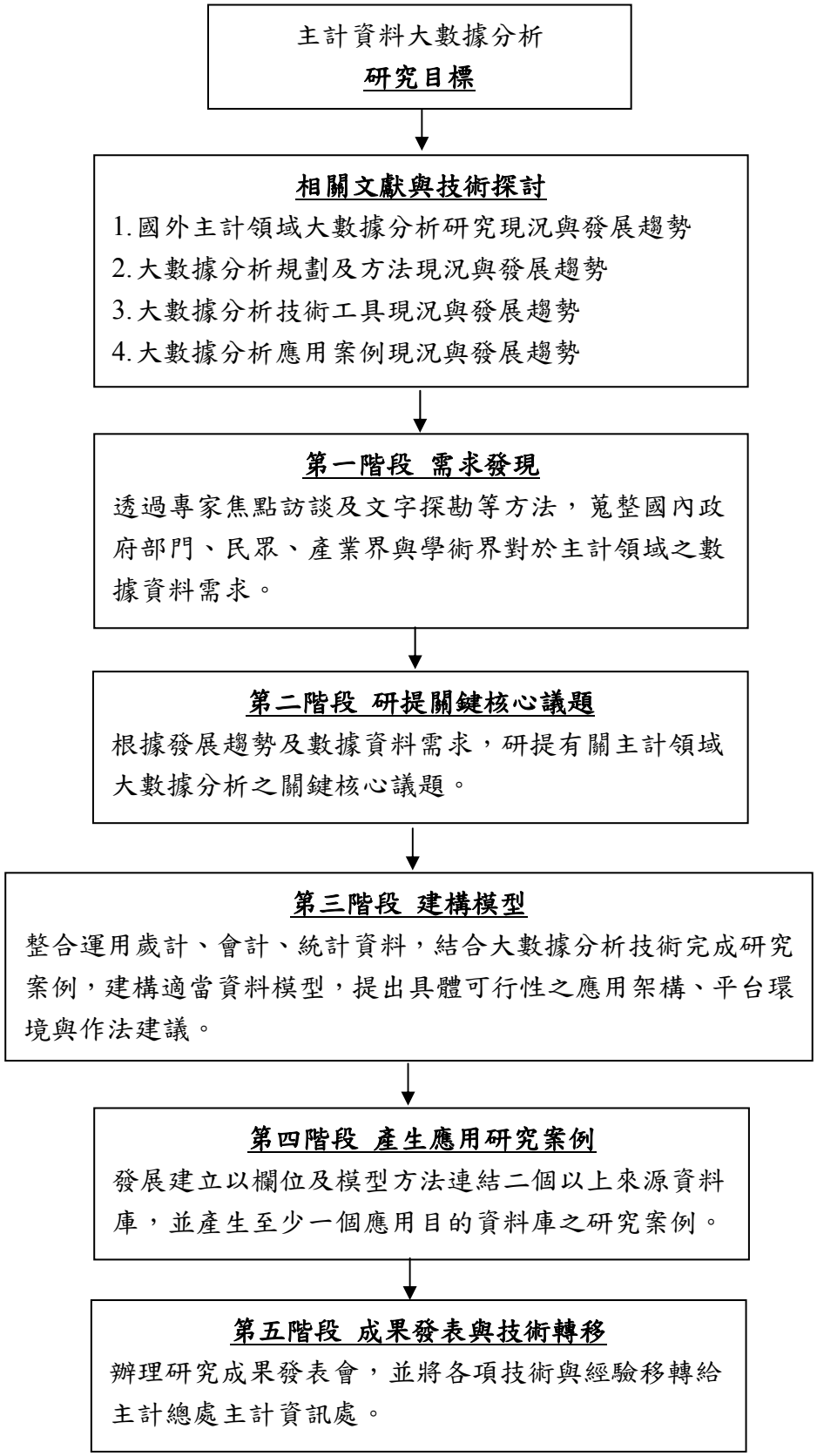


圖 1 本研究案主要研究方法與過程

除了研究目標的確認以及必要的文獻與技術探討以外，本計畫將採用五階段研究方法來達成研究目標，如圖 1 所示。第一階段先發現需求，本研究透過專家焦點訪談、爬蟲技術及文字探勘等方法，蒐整國內政府部門、民眾、產業界與學術界對於主計領域之數據資料需求。第二階段則將依據第一階段結果研提關鍵核心議題，根據前提發展趨勢及數據資料需求，研提有關主計領域大數據分析之關鍵核心議題。第三階段即可依第一階段需求內容與第二階段的關鍵核心議題進行模型的建構，選擇至少一個以上研究議題，整合運用歲計、會計、統計資料，結合大數據分析技術，萃取資料特徵與樣態，建構適當資料模型，並提出具體可行性之應用架構、平台環境與作法建議。完成模型建構後，即可進行第四階段工作，著手產生應用研究案例，並依委辦單位需求發展建立以欄位及模型方法連結二個以上來源資料庫，並產生至少一個應用目的資料庫之研究案例，以展現如何應用大數據分析方法與技術於整合歲計、會計、統計資料服務與服務框架。最後，利用本研究之期中、期末報告，邀集至少五位專家學者，辦理二場專家學者審查會，並提出對主計業務相關施政決策、後續研究建議，以及辦理至少二場研究成果分享發表會。透過本研究在技術領域所累積的知識、各項技術與經驗，移轉給主計總處主計資訊處。

透過上述研究方法與過程，本計畫將可產生對施政有參考價值之研究結果，增進主計業務相關政策規劃應用能力，強化現有施政決策分析方式，有效提升主計業務決策品質與效能。

第三節、交付項目與期程

一、本研究預計交付項目

（一）文獻蒐集與訂定核心議題：將於計畫期中專家審查會及期中報告前蒐集與訂定完成。

（二）大數據資料模型建構：本計畫將依預期目標所列之範圍，完成模型建構，並交付相關系統文件與程式碼。

（三）案例研究：本計畫將依預期目標所列之範圍，完成 4 項關鍵核心議題之研究案例，並交付相關案例說明文件。

（四）研究報告：依本計畫研究重點及「行政院主計總處委託研究計畫作業要點」相關規定格式撰寫期中及期末研究報告，除依需求印製指定數量，並交付相關電子檔案。報告內容包括：計畫背景、相關文獻與國內外最新發展、模型建構說明、應用案例說明、計畫結果總結等。

（五）會議報告：本案進行需要訪談所執行之座談會及個別訪談，研究團隊均將應依下述項目交付報告及紀錄：

1. 深度訪談紀錄：研究團隊將摘要記錄訪談過程及發言重點，會後應彙整提供書面紀錄及電子檔案，並經審閱後列入相關研究報告之附錄。

2. 發表會議紀錄：本案辦理之發表會，研究團隊將摘要記錄會議過程及發言重點，會後應彙整提供書面紀錄及電子檔案，並經審閱後列入相關研究報告之附錄。

二、計畫期程與工作項目列表

本研究計畫期間為民國 105 年 4 月 1 日至民國 105 年 11 月 30 日，共 8 個月。計畫期程與工作項目如表 1 所示；其中，”X”記號代表已於期中報告前完成之工作項目，而”O”記號代表已於期末報告前完成之工作項目：

表 1　本研究案計畫期程與工作項目

月份 / 工作項目	民國105年4月1日至民國105年11月30日								
	3~4	4~5	5~6	6~7	7~8	8~9	9~10	10~11	11~12
第一階段：資料準備期									
1.1 專案展開與組織分工	X								
1.2 文獻彙整與案例分析	X	X							
1.3 訂定核心議題		X	X	X					
1.4 檢視與取得可用資料		X	X	X					
1.5 專家座談審查（期中）					X				
第二階段：資料建模期									
2.1 資料整理與資料映射				O					
2.2 資料建模與評估				O	O				
2.3 模型修正					O	O			
第三階段：資訊分享期									
3.1 撰寫期末報告							O	O	
3.2 專家座談審查（期末）								O	
3.3 產出最終成果報告書								O	O
3.4 成果分享會									O

第四節、工作項目與預期施政助益

一、預計完成工作項目

本研究計畫預計完成項目主要分為三大部分，依序為：

(一)資料準備期

1. 文獻彙整與案例分析：將於履約起始日起 2 個月內完成文獻彙整與案例分析，包括國外主計領域大數據報告、大數據分析方法、技術工具及應用案例現況與發展趨勢。
2. 訂定關鍵核心議題：將於履約起始日起 3 個月內完成對業務單位的訪談，並彙整民眾輿情、產業界與學術界之需求，針對需求進行綜整與研析，訂定關鍵核心議題。
3. 檢視及取得可用資料：將於履約起始日起 4 個月內檢視在資料建模時的可用資料並取得相關檔案。
4. 專家座談審查：將於履約起始日起 4 個月內完成期中審查。

(二)資料建模期

將於履約起始日起 5 個月內完成結合大數據分析技術完成研究案例，建構適當資料模型，提出具體可行性之應用架構、平台環境與作法建議。

(三)資訊分享期

將分別於期中及期末蒐集學者專家交流及審查意見，進行報告修正，於履約起始日起 8 個月內提出期末報告及舉辦成果分享會。

本委託研究計畫應於履約起始日起 4 個月內提出期中報告。期中報告之內容，包括下列各項 （1）研究方法與進度說明；(2）蒐集之資料、文獻與分析；(3） 主計領域需求蒐整；(4）參考文獻。期末報告除依據期中審查意見修正報告架構方向外，並加入關鍵核心議題之應用範例平台建置結果，並於履約起始日起 8 個月內提出期末報告。

二、對相關施政之助益

本研究預期可提供以下助益：

（一）研究結果有助於瞭解國外先進國家有關主計領域大數據分析規劃應用、分析方法及技術工具等發展趨勢，以及國內各界資料需求、關鍵核心議題，作為未來發展大數據分析與應用之參考。

（二）有效整合歲計、會計、統計資料，運用大數據分析技術，選擇若干研究議題，建構示範案例及資料模型，產生對施政有參考價值之研究結果，增進主計業務相關政策規劃應用能力，優化政府施政。

（三）善用大數據分析技術，結合主計總處既有厚實的業務能量與經驗，強化現有施政決策分析方式，有效提升主計業務相關施政決策品質與效能。

（四）藉由本研究計畫示範案例進行雛型平台實作，並提出大數據分析之可行性作法建議，研究成果可作為推動大數據應用與發展的依據。

第二章、文獻探討

大數據 Big Data （又稱海量資料）指的是所涉及的資料量規模非常巨大且相當龐雜，而如此龐大的資料需要透過大數據分析技術協助萃取隱藏在資料庫中的訊息與知識。

第一節、大數據分析概述

一、大數據發展歷程

自 2005 年大數據分析技術 Hadoop 誕生，就預言將對未來十年帶來的影響。當時 Hadoop 提出兩項關鍵服務，一是採用 Hadoop 分布式文件系統（HDFS)的可靠數據存儲服務，另一是運用 MapReduce 技術的高性能平行數據處理服務。此兩項服務的共同目標，是提供一個快速可靠地處理分析結構化且複雜數據的環境基礎，大數據處理分析可從此環境中獲得實現。綜整數據發展沿革如表 2。

表 2 大數據發展沿革

年份	發展歷程
2008 年	●「大數據」得到部分美國知名計算機科學研究人員的認可，發表了一份有影響力的白皮書《大數據計算：在商務、科學和社會領域創建革命性突破》，提出大數據真正重要的是新用途和新見解，而非數據本身。
2009 年	●印度政府建立用於身份識別管理的生物識別數據庫，聯合國全球脈衝項目運用手機和社群網站的數據源，分析預測螺旋價格、疾病爆發等問題。 ●美國政府通過啟動 Data.gov 網站的方式進一步開放資料的大門，該網站向公眾提供各式各樣的政府資料，並收集超過 4.45 萬量資料集；其資料被用於保證一些網站和智能手

年份	發展歷程
	機應用程序，如跟蹤航班、產品召回、特定區域內失業率等應用，此次政策與行動激起世界各國政府紛紛仿效，相繼推出類似舉措。
2010 年	●肯尼斯•庫克爾在《經濟學人》上發表了長達 14 頁的大數據專題報告《數據，無所不在的數據》。 ●科學家和計算機工程師已經為這個現象創造了一個新詞彙：「大數據」。
2011 年	●IBM 的沃森超級計算機每秒可掃描並分析 4TB（約 2 億頁文字量）的數據量。 ●全球知名諮詢顧問公司麥肯錫（McKinsey&Company）全球研究院（MGI）發布《大數據：創新、競爭和生產力的下一個新領域》的研究報告，大數據開始備受關注。
2012 年	●美國奧巴馬政府在白宮網站發布《大數據研究和發展倡議》。 ●美國軟件公司 Splunk 於 4 月 19 日在納斯達克成功上市，成為第一家上市的大數據處理公司。 ●聯合國於紐約發布一份關於大數據政務的白皮書，彙總各國政府如何利用大數據提供更好地服務和保護人民的政策。 ●我國行政院通過「政府資料開放推動策略」，第一階段各部會至少 5 項資料集，第二階段 50 項。
2013 年	●我國訂定「行政院及所屬各級機關政府資料開放作業原則」。 ●建置政府資料開放平台 data.gov.tw。
2014 年	●世界經濟論壇以「大數據的回報與風險」主題發布《全球訊息技術報告（第 13 版）》。報告指出，在未來幾年中針對各種訊息通信技術的政策將會顯得更加重要。

年份	發展歷程
	●我國加速推動「開放資料」，迄年底上架超過 4,000 筆資料集。 ●台灣開放資料聯盟成立，逾 300 名產業代表加入。
2015 年	●我國毛前院長宣示 2015 年為 Open Data 深化應用元年。 ●積極推動大數據與群眾外包。 ●英國開放知識基金會（OKFN）開放資料（Open data）國際評比，2015 年台灣在全球 149 個國家中名列世界第一。

資料來源：本計畫整理

二、大數據分析發展方向

大數據技術是一種新一代技術和架構，其特色是以成本較低、快速的採集、處理和分析技術，從各種超大規模的數據中提取價值。大數據技術不斷湧現和發展，讓處理海量數據更加容易、更加便宜和迅速，成為利用數據的好助手，甚至可以改變許多行業的商業模式，綜合上述，本研究將大數據分析的發展分為八大方向如下說明，另有關 2013-2016 大數據分析發展趨勢如表 3 所示。

(一)大數據採集與預處理方向

最常見的問題是數據的多源和多樣性，導致數據的質量存在差異，嚴重影響到數據的可用性。針對這些問題，目前已有許多公司推出多種數據清整和質量控制工具。

(二)大數據儲存與管理方向

最常見的挑戰是儲存規模大，管理複雜，需要兼顧結構化、非結構化和半結構化的數據。在大數據儲存和管理方向，尤其值得我們關注的是大數據索引和查詢技術、即時處理的發展。

(三)大數據計算模式方向

由於大數據處理多樣性的需求，目前出現多種典型的計算模式，包括大數據查詢分析計算（如 Hive）、批處理計算（如 Hadoop MapReduce）、流式計算（如 Storm）、迭代計算（如 HaLoop）、圖計算（如 Pregel）和內存計算（如 Hana），而這些計算模式的混合計算模式將成為滿足多樣性大數據處理和應用需求的有效手段。

(四)大數據分析與挖掘方向

在數據量迅速膨脹的同時，還要進行深度的數據深度分析和挖掘，並且對自動化分析要求越來越高，越來越多的大數據分析工具和產品應運而生，如用於大數據分析的 R Hadoop 版。

(五)大數據視覺化分析方向

通過視覺化方式來幫助人們探索和解釋複雜的數據，有利於決策者挖掘數據的商業價值，進而有助於大數據的發展。現況已有很多公司也展開相關的研究，試圖把視覺化引入其不同的數據分析和展示的產品中，各種可能相關的商品也將會不斷出現。

(六)大數據安全方向

大數據的安全一直是企業和學術界非常關注的研究方向；目前以通過文件訪問控制來限制呈現對數據的操作、基礎設備加密、匿名化保護技術和加密保護等技術，保護數據安全。

(七)大數據 Fintech 方向

近年來，FinTech（金融科技）浪潮掀起產學各界積極投入分析和研究，成為一門兼具理論與實務的跨領域新科學。在《金融科技

發展策略白皮書》中以「創新數位科技，打造智慧金融」為主軸，其中大數據與用戶的身分認證等都是未來致力的目標。FinTech 重點不是科技，而在於以數據分析與創新模式創造新價值，其價值在於以創新模式填補未被傳統金融業滿足的需求，例如 P2P 網貸、群眾募資、區塊鏈、社群金融、智慧電商、風險預測等，大數據分析都扮演重要角色，藉由分析金融市場中的各種交易數據可開發出智慧型決策系統，瞭解客戶的行為與需求能設計客製化金融商品。

(八)常住人口大數據

未來十年常住人口及人口流動遷移觀察是重要的課題，從發展現況看，隨著工業化、城市化、資訊化加速推進，城鄉發展不平衡、區域發展不協調問題較為突出，大量常住人口變成流動人口，社會結構和利益格局發生變化，政府需要應對的公共事務規模將面臨挑戰。政府部門在長期的人口管理中積累大量的數據，戶籍人口數據、婚姻死亡登記數據、健保檔案、醫療數據、跨縣市流動人口等數據，連同各大研究機構的調查數據，成為數據資源開發共享時代的國家資源，是深入研究新型城市化和流動人口社會問題提供豐富的資源。

表 3　2013~2016 大數據發展趨勢

	2013 年	2014 年	2015 年	2016 年
應用領域	數據資源化	大數據從「概念」走向「價值」	大數據分析成為資料價值化的熱點	視覺化推動大數據平民化
分析方法	大數據分析的革命性方法	深度學習與大數據智慧成為分析主流	1. 計算模式深度學習 2. 視覺化分析與視覺化呈現	深度分析推動大數據智慧應用
技術工具	大數據與雲計算深度融合	大數據架構多樣化模式並存	平台架構與基礎設施	大數據多樣化處理模式
個資安全	大數據隱私問題突顯	大數據安全與隱私	大數據安全與隱私保護	大數據安全與隱私令人憂慮
產業應用	資料共用聯盟	1. 數據商品化與數據共用聯盟化 2. 基於大數據的推薦與預測應用流行 3. 大數據產業成為戰略性產業	與各行業結合，進行跨領域應用	1. 互聯網、金融、健康保持熱度，智慧城市、企業資料化、工業大數據是新亮點 2. 提升社會治理和民生領域應用

資料來源：本計畫整理

三、大數據定義

“Big Data” 一詞最早出現在 1997 年邁克爾考克斯的論文中，直到 2008 年 9 月，《科學》雜誌發表論文”Big Data: Science in the Petabyte Era”，Big Data 一詞才開始廣泛傳播。（謝邦昌，2015）

大數據（Big Data），或稱巨量資料、海量資料，指的是所涉及資料量規模巨大到無法透過目前主流軟體工具，在合理時間內達到擷取、管理、處理、並整理成為幫助企業經營決策更積極目的之資訊。「大數據」是由數

量巨大、結構複雜、類型眾多資料構成的資料集合，是基於雲計算的數據處理與應用模式，通過資料的整合共用形成的智慧資源和知識服務能力。大數據現在不只是數據處理工具，更是一種企業思維和商業模式，因為資料量急速成長、儲存設備成本下降、軟體技術進化和雲端環境成熟等種種客觀條件就位，才能讓資料分析從過去的洞悉歷史進化到預測未來，甚至是破舊立新，開創從所未見的商業模式。(李欣宜，2015)

大數據（Big Data）具備三個特徵：一是大量性（Volume），2020 年全球資料量將成長為 40ZB，企業未來七年內企業需要管理的資料量將成長 4400%（資策會，2015）；二是高速性（Velocity），Google 每分鐘處理 200 萬次以上搜尋 (中國生產力中心，2013)。三是多樣性（Variety），2020 年非結構性資料將占整體資料 80%、行動應用普及帶動各類社群網路、虛擬社團成形，促使多格式(影片、照片、音樂)資料大量產生（謝邦昌，2015）。李欣宜（2015）認為大數據的定義是 Volume（容量）、Velocity（速度）和 Variety（多樣性），但也有人另外加上 Veracity（真實性）和 Value（價值）兩個 V。但其實不論是幾 V，大數據的資料特質和傳統資料最大的不同是，資料來源多元、種類繁多，大多是非結構化資料，而且更新速度非常快，導致資料量大增。

在大數據時代下的研究目的為「探索導向」研究 （discovery-driven）而不是「假設導向」研究（hypothesis-driven）（林顯明，2015）。故大數據真正的意涵即是透過探索、分析、挖掘大數據的資料庫，發現假設與潛在可能性，以產生更多的資料價值。

四、大數據處理

大數據處理資料理念的三大轉變：要全體不要抽樣，要效率不要絕對精確，要相關不要因果。具體的大數據處理方法有很多，可以彙整成大數據處理流程。整個處理流程可以概括為四個步驟，分別是採集、分析、導

入和預處理、以及挖掘。

(一)採集

大數據的採集是指利用多個資料庫來接收使用端的資料，並且使用者可以通過這些資料庫來進行簡單的查詢和處理工作。例如，電商會使用傳統的關係型數據庫 MySQL 和 Oracle 等來存儲每一筆交易資料，除此之外，Redis 和 MongoDB 這樣的 NoSQL 資料庫也常用於資料的採集。在大數據的採集過程中，其主要特點和挑戰是可能會有成千上萬的用戶來進行訪問和操作，如火車票售票網站，在特定時間的同時間訪問量居高，需要在採集端部署大量資料庫才能支撐。

(二)分析

分析主要運用分散式資料庫，或是分散式運算來對儲存於其內的海量資料進行分析和分類匯總等，以滿足大多數常見的分析需求。例如，即時性處理需求會用到 EMC 的 GreenPlum、Oracle 的 Exadata，以及基於 MySQL 的列式儲存 Infobright 等，而批次處理或基於半結構化資料處理的需求可以使用 Hadoop，在此部份特點與挑戰是資料量鉅大且不斷增長，增加分析的複雜度。

越來越多的應用涉及到大數據，而這些大數據的屬性，包括數量、速度、多樣性等都呈現資料不斷增長的複雜性，因此，大數據分析方法在大數據領域就顯得尤為重要，可以說是決定最終資訊是否有價值的決定性因素。大數據分析可以分為五個基本方面：

1. 預測性分析能力（Predictive Analytic Capabilities)

大數據分析技術如統計、人工智慧、機器學習及資料採礦等演算法可以使人更能理解資料，而預測性分析可以根據視覺化分析、人工

智慧、機器學習與資料採礦的結果做出一些預測性的判斷。

2. 資料品質和資料管理（Data Quality and Master Data Management）

資料品質和資料管理是一些管理方面的最佳實踐。通過標準化的流程和工具對資料進行處理可以保證一個預先定義好的高品質的分析結果。

3. 視覺化分析（Analytic Visualizations）

不管是對資料分析專家還是普通使用者，資料視覺化是資料分析工具最基本的要求。視覺化可以直觀的展示資料，讓資料自己說話，讓觀眾聽到結果。

4. 語義引擎（Semantic Engines）

由於非結構化資料的多樣性帶來了資料分析的新的挑戰，因此需要一系列的工具去解析、提取、分析資料。語義引擎需要被設計成能夠從「文檔」中智慧提取資訊。

5. 大數據分析技術（Big Data Analysis）

視覺化是給人看的，大數據分析技術如統計、人工智慧、機器學習及資料採礦等演算法即是給機器看的，可以讓我們深入資料內部，挖掘價值；此外，上述分析技術不僅要處理大數據的量，也要處理大數據的速度。

(三)導入

雖然採集端本身會有很多資料庫，但是如果要對這些海量資料進行有效的分析，還是應該將這些來自前端的資料導入到一個集中的大型分散式資料庫，或者分散式儲存集群環境，並且可以在導入基礎上做一些簡單的清整和預處理工作。導入與預處理過程的特點

和挑戰主要是導入的資料量大，每秒鐘的導入量經常會達到百兆，甚至千兆級別。

(四)挖掘

大數據分析技術如統計、人工智慧、機器學習及資料採礦等演算法，其中人工智慧、機器學習及資料採礦等演算法通常不會預設主題，與統計分析不太相同，主要是將現有資料進行基於各種演算法計算，並得到預測效果，從而實現一些進階資料分析的需求。比較典型演算法有用於聚類的 K-Means、用於統計學習的 SVM 和用於分類的 Naive Bayes 等，主要使用的工具有 Hadoop 的 Mahout 等，該過程的特點和挑戰是用於挖掘的演算法較複雜，並且計算涉及的資料量和計算量較大。

第二節、國外主計領域大數據分析研究現況與發展趨勢

大數據伴隨著雲計算與互聯網的發展，正對全球經濟社會產生巨大的影響。也因為大數據提供新的技術與方法，可帶來主計領域（歲計、會計與統計）的創新思維與機會，將有助於主計領域長遠性的發展與效能的提升。

一、企業面發展現況與趨勢

（一）現況

根據「2014 年資誠（PwC）全球資料分析調查--Gut &Gigabytes 報告」顯示，愈來愈多企業已在制定決策時使用大數據。在受訪的 1,135 家全球企業中，超過六成（64%）表示大數據的應用已對其決策帶來影響。企業現在正面臨艱鉅的生存挑戰。過去，企業只要固守本業與專注少數熟悉國家（主要是歐美），就能搶占市場大餅。但在經濟迅速變遷且市場競爭加劇的今日，商機不僅在不同國家間存有差異，甚至在同一國內亦會隨區域而有所差別，故企業必須更瞭解不同區域的客戶。近日行政院科技內閣希望開放更多有價值的資料，讓產業可應用；其實國內企業現在就有不少公開可取得的財務資料，搭配 XBRL 等工具的應用，可讓企業監控競爭同業及供應鏈客戶財務資料，通過對外部非結構性資料的比較分析，企業可以更直接地洞察它們的財務結構競爭力與風險管理指標等。利用大數據，管理會計也已經可以更好地發揮事前預測的功能，在預測的基礎上，制定實施相關戰略或策略。

1. 對大型建設項目投入產出的預測

大型建設項目前期的建設成本直接影響到後期的營運成本。例如：國內機場或港口的建設，前期的建設成本特別大，而在真正使用

後，根本沒有那麼多的業務量，從而導致虧損（或巨額虧損）。如果這些機場或港口建設的前期，利用可以得到之多個來源數據進行利潤預測，就可能避免資源投入的浪費。與以往不同，是通過電信營運商提供的位置服務，可以較準確地知道在某一地區活動的人數，再參考其他的數據，如收入水準、經濟發展、天氣變化等數據，可以較準確地對建設項目進行分析。

2. 預測市場，確定資源配置

對於企業而言，過去在預測市場需求是相對比較困難的事，而現在則可利用大數據技術對一個地方的市場進行預測。例如：某建商在佈局選地決策之前，通過電信手機運營商，在同一時間點檢測當地在網使用人數的數據，來判斷實際人口規模，並結合新屋供應量、房價均價、購房人群年齡結構等眾多數據，進行市場和區域預判，並最終判斷當地是否擁有巨大的購房群體支持。

3. 預測機器的運行或使用狀態，避免停機帶來的更大成本

機器停用的成本一般來說要大於維修成本，智慧儀器儀表和各種傳感器的使用，使企業能夠即時地收集機器或設備運行狀況的數據，通過對這些數據的分析，判斷機器或設備運行狀態的好壞以及是否需要修理，從而避免機器停用所帶來的更高的恢復成本。例如：美國的快遞公司 UPS，通過在每輛車上安裝傳感器，檢測車輛的運行狀態，通過大數據的分析，判斷哪個零件可能要出問題，並即時地進行預防性修理或更換，避免車輛拋錨後發生的延誤或拖車等更多的成本，這項措施已為該公司節省了幾百萬美元。

4. 挖掘客戶需求，預測產品的流行趨勢

如何預測產品的流行趨勢，生產適合客戶需要的產品，是企業非常

關心的問題。目前通過社群網站、商務網站以及搜索網站都可以收集到產品流行的數據，通過對這些數據的分析，可以及時瞭解各種產品的流行趨勢。例如：著名的服裝企業 ZARA 通過網店和實體店兩個管道來收集客戶對服裝的評價數據。在實體店，店員隨時記錄顧客對服裝的評價；在網店上，透過網路收集客戶對服裝的瀏覽記錄、評價記錄和客戶的實際消費數據，這些數據傳到總部進行集中分析，得出客戶對每種服裝的改進意見及流行趨勢，並將其反饋給服裝設計師加以改進，從而能夠對客戶的需求做出快速反應。

5. 評價客戶信用，預測企業風險

一個客戶能否償還所欠貸款或者貨款，企業是否要承擔壞帳風險，壞帳風險有多大，換言之，對客戶瞭解得越充分，企業預測是否出現壞帳的準確度愈高，避免壞帳出現的機率也愈大。目前可以從不同的管道獲得反映客戶不同方面的數據，而不僅僅是財務狀況的數據，這樣可以更真實地反映一個客戶的信用狀況。

6. 預測總體環境的變化，製造消費

大數據技術的應用一方面在挖掘客戶需求，另一方面還可以通過對天氣、各種社會事件、環境等數據的分析，預測一些社會事件的發生，如 Google 成功預測發生在美國地區的流感。根據預測發生的這些社會事件，可以給終端消費者採購商品提供建議，從而由挖掘顧客的需求，進一步變為「創造」消費需求。沃爾瑪正是利用大數據技術的思維，2011 年 4 月以 3 億美元收購了 Kosmix 公司。Kosmix 從事的業務是收集、分析網路上的大數據，並將這些數據賣給企業。Kosmix 為沃爾瑪打造的大數據系統連結到美國著名的社群網站 Twitter、Facebook 等。網站的工作人員從每天熱門消息中，推出與社會時事呼應的商品，創造消費需求。

會計對企業各種經營數據的分析是給企業內部管理層決策使用的，因此不僅應分析出數據，還應分析出原因的數據。過去由於種種取得資料困難或無法取得數據的限制，無法得到數據去解釋導致結果的原因，大數據時代的到來改變了這一切。

1. 收入分析由結果面向轉為過程面向

如果是通過網店進行銷售的企業，從顧客對產品的評價中，就可以得到很多數據（如產品的質量、樣式、材料和服務態度如何、快遞是否及時等），並通過對這些數據的分析來解釋收入增加或減少的原因。另外，也可以從其他的管道獲得數據，分析產品市場的整體變化情況和競爭對手的情況，從各種不同的角度解釋對銷售收入增加或減少的原因。

2. 成本分析由結果面向轉為過程面向

對電信的營運商來說，前期投入的設備屬於沉沒成本，無法改變，而對客戶的服務費用則可以通過改變策略進行改變。例如，利用大數據技術可以得到每個用戶的通話時間及各種諮詢的問題，在手機實名制的情況下，可以分析出不同年齡區間、不同地區客戶通話的時間長短及所諮詢的問題。根據分析的結果，可以針對不同年齡或不同地區客戶重新製定新的服務策略與營銷策略，從而降低服務費用或增加收入。

3. 風險分析由結果結面轉為過程面向

美國一家名為 SCOR 的金融資訊公司，當收到銀行客戶的信用評估申請後，經過客戶同意，調取其在 facebook、twitter 等社群媒體的數據，分析該客戶的行為特點、興趣愛好，甚至會根據該客戶朋友

圈特性對客戶信用進行評估。這些社群數據能真實反映客戶信用狀況，幫助銀行更準確地判斷客戶的違約風險。ZestCash 為大量運用社群分析的網路金融公司，它透過研究其他銀行尚未採用的變量評估客戶，如：電話帳單付款記錄、該客戶在借貸網站逗留多久等訊息，取代傳統銀行核貸評估採用的有限資料－信用報告及信評分數。ZestCash 靠著聰明的運用資料，有效降低客戶違約率，並將之反應在客戶的借貸成本，借款人甚至可以只支付二分之一傳統發薪日貸款費用（註：payday loan，小額短期融資，供借款人在領到薪水前短期調度資金之用）。

另外，平衡計分卡廣泛應用於企業的戰略管理、業績評價中，但在平衡計分卡中，除了最上一層的指標是財務指標外，其他各層都是非財務指標。這些非財務指標能更好、更客觀地反映企業戰略執行的狀態，也能更公正地評價企業內部組織的業績。由於各種技術條件的限制，過去很多非財務指標很難收集到。電子商務的興起使廠家離客戶越來越近，客戶對產品和服務的感受、產品在市場上的表現等數據都很容易收集。如客戶滿意度指標，不同行業的企業採用不同的方法進行收集。電信或銀行等服務業收集客戶對所提供服務的滿意程度；而對於製造業，可以通過網站收集客戶對產品的各種評價數據（如對產品設計、產品質量和對產品採用原料的滿意程度），成為企業對產品的設計、生產和採購等部門業績的評價指標。

(二）發展趨勢

1. 企業會計發展趨勢

(1)從事後的財務報告向即時財務報告發展

當前，只有在企業生產經營以及銷售等業務結束後，會計人員才能夠編制各種財務報告，而且財務報告編寫過程較漫長，例如：年度財務報告常常需要三到四個月時間才能編寫完成，這樣

會嚴重影響會計資料的及時性和有效性。採用事後編報財務報告的方式顯得過於遲緩，尤其是對於大型企業，反映日益頻繁以及複雜的企業經營和管理活動的財務資訊。隨著資訊數據處理技術快速發展，越來越多企業已經意識到即時財務報告的迫切性和重要性，大數據資訊處理技術能將即時財務報告實現。

(2)從會計資訊的過去到預測未來方向發展

長久以來，企業會計報告制度的主要目的是反映過去既成事實，對預測未來非常薄弱，甚至力不從心。今後，企業會計預測結果，已經能夠顯示出比總結過去更為重要作用。在大數據時代，會計人員能夠尋求利用大數據對企業的財務以及運行的未來進行預測、預報，甚至可採取風險規避措施，並明確顯示企業業績增長點。在企業中，會計人員發揮更具企業戰略性和“前瞻性”的領導作用。通過不斷收集企業會計資訊、儲存、傳遞和分析的海量企業會計資料，會計人員會逐漸改變會計核心內容，工作的重心。從資料資訊獲取、分析和挖掘過程中，會計人員將向企業領導提供可靠的預測資訊，為股東和廣大員工利益提供有利於企業做出正確決策正確資訊。

2. 企業會計審核發展趨勢

現今國內外大型會計師事務所都看到 D&A 會計審核技術發展趨勢，不約而同研發有助於數據分析的會計審核工具，並廣泛招募不同領域技術專才，部份領先業者已展開初步的運用及技術實踐。例如 KPMG 台灣所率先選擇若干擁有大量數據的特定產業及資訊化程度高的會計審核客戶為指標案件（Pilot Case），採用最新的數據分析的會計審核工具，試著將數據分析會計審核概念嵌入查核方法中。

自從互聯網技術的出現，正悄悄地改變企業思維和社會經濟行為，同時也催生了更多的新興技術，也在不斷開拓和推動著人類社會快速發展，擴展對資訊與資料的應用範圍。大數據將成為全人類下一個創新、創造的熱點領域，包含各行各業。未來通過對大數據的處理、交換、整合和分析能夠發現新的知識和創造新的價值。

3. 企業統計發展趨勢

隨著互聯網、電子商務、移動互聯網、社群網路、物聯網等技術蓬勃發展，大數據成為這些新一代資訊技術發展的必然產物。超過90%的企業認為分析資料的能力對該企業未來五年的商業模式有關鍵性的影響，這些企業主要著重在其價值鏈內的資料能有效地交換，產出的產品被數位化記錄，以及採用及時資料監控整個生產線。

企業統計發展可以從工業大數據看到趨勢，工業大數據是指在工業領域資訊化應用中所產生的大數據。隨著資訊化與工業化的深度融合，資訊技術滲透到了工業企業產業鏈的各個環節，條碼、二維碼、RFID、工業感測器、工業自動控制系統、工業物聯網、ERP、CAD/CAM/CAE/CAI 等技術在工業及企業中得到廣泛應用，尤其是互聯網、移動互聯網、物聯網等新一代資訊技術在工業領域的應用，工業及企業也進入了互聯網工業的新發展階段，工業及企業所擁有的資料也日益豐富。工業及企業中生產線處於高速運轉，由工業設備所產生、採集和處理的資料量遠大於企業中電腦和人工產生的資料，從資料類型看也多是非結構化資料，生產線的高速運轉則對資料的即時性要求也更高。因此，工業大數據應用所面臨的問題和挑戰並不比互聯網行業的大數據應用少，某些情況下甚至更為複雜。

工業大數據應用將帶來工業企業創新和變革的新時代。通過互

聯網、移動物聯網等帶來的低成本感知、高速移動連接、分散式運算和高級分析，資訊技術和全球工業系統正在深入融合，給全球工業帶來深刻的變革，創新企業的研發、生產、運營、行銷和管理方式。這些創新不同行業的工業企業帶來了更快的速度、更高的效率和更高的洞察力。工業大數據的典型應用包括產品創新、產品故障診斷與預測、工業生產線物聯網分析、工業企業供應鏈優化和產品精準行銷等各個方面。

首先看產品創新的應用。客戶與工業企業之間的交互和交易行為將產生大量資料，挖掘和分析這些客戶動態資料，能夠說明客戶參與到產品的需求分析和產品設計等創新活動中，為產品創新作出貢獻。福特公司即是這方面的表率，該公司將大數據技術應用到福特福克斯電動車的產品創新和優化中，這款車成為一款名副其實的「大數據電動車」。第一代福特福克斯電動車在駕駛和停車時產生大量資料，在行駛中，司機持續地更新車輛的加速度、剎車、電池充電和位置資訊。這對於司機非常有用，但資料也傳回福特工程師那裡，以瞭解客戶的駕駛習慣，包括如何、何時以及何處充電。即使車輛處於靜止狀態，它也會持續將車輛胎壓和電池系統的資料傳送給最近的智慧型電話。這種以客戶為中心的大數據應用場景具有多方面的好處，因為大數據實現寶貴的新型產品創新和協作方式。司機獲得有用的最新資訊，而位於底特律的工程師彙總關於駕駛行為的資訊，以瞭解客戶，制訂產品改進計畫，並實施新產品創新。而且，電力公司和其他協力廠商供應商也可以分析數百萬英里的駕駛資料，以決定在何處建立新的充電站，以及如何防止脆弱的電網超負荷運轉。

第二個應用趨勢是產品故障診斷與預測，這可以被用於產品售後服務與產品改進。無所不在的感測器、互聯網技術的引入使得產品故障即時診斷變為現實，大數據應用、建模與模擬技術則使得預

測動態性成為可能。在馬航 MH370 失聯客機搜尋過程中，波音公司獲取的發動機運轉資料對於確定飛機的失聯路徑起了關鍵作用。以波音公司飛機系統為例，看看大數據應用在產品故障診斷中如何發揮作用。在波音的飛機上，發動機、燃油系統、液壓和電力系統等數以百計的變數組成了在航狀態，這些資料不到幾微秒就被測量和發送一次。以波音 737 為例，發動機在飛行中每 30 分鐘就能產生 10 TB 資料。這些資料不僅僅是未來某個時間點能夠分析的工程遙測資料，而且還促進即時自我調整控制、燃油使用、零件故障預測和飛行員通報，能有效實現故障診斷和預測。再看一個通用電氣（GE）的例子，位於美國亞特蘭大的 GE 能源監測和診斷（M&D）中心，收集全球 50 多個國家上千台 GE 燃氣輪機的資料，每天就能為客戶收集 10GB 的資料，通過分析來自系統內的感測器振動和溫度信號的恒定大數據流程，這些大數據分析將為 GE 公司對燃氣輪機故障診斷和預警提供支撐。風力渦輪機製造商 Vestas 也通過對天氣資料及其渦輪儀錶資料進行交叉分析，從而對風力渦輪機佈局進行改善，由此增加了風力渦輪機的電力輸出水準並延長了服務壽命。

二、政府面發展現況與趨勢

（一）現況

隨著大數據分析運用在企業界蓬勃發展，政府機關也逐漸開始運用大數據以追求良善治理（good governance）（廖洲棚、陳敦源、蕭乃沂、廖興中，2013）。政府可使用最新科技來更好的運用大數據、並藉此提升政府績效、治理能力，而又不會使財政負擔過重（Shindelar, 2014）。從公共治理的角度來看，政府有義務將施政績效與治理能力提昇來達到更好的公共治理，而大數據分析方法的出現正是一個能夠強化政府公共治理能力的利器。對於政府而言，大數據資料可依據資料產生的來源而區分為內部資料與外部資料。以內部資料來說，政府

本身擁有非常大量的資料，舉凡財稅、健康保險、教育、衛生福利等資料都是屬於政府內部設置之系統或設備產生的數位資料，外部資料則是與公共治理相關但非由政府擁有之系統或設備產生的數位資料，例如公開於網際網路的新聞媒體資料、研究機構公開的調查資料、民眾公開於網路上討論分享的數位資料等等。對民主國家而言，為使政府施政能反映民意需求，當有愈來愈多民眾願意透過網路分享或討論公共事務時，這些散佈在網際網路上的民意資料也是另一股重要的政府外部資料的來源，如果政府能夠有效的運用這股新興的技術進行資料的深入分析，勢必能夠大幅提昇政府的施政績效與治理能力。

大數據分析已滲透到美國的各個領域，並帶給美國強大的變化。紐約市長資料分析辦公室（MODA），通過資料分析來提升政府日常運作水準、預防和處置緊急事件，MODA 還和新企業加速服務團隊（NBAT）合作，利用量化分析手段評估政府決策。大數據分析在美國的公共事業領域也大顯身手，美國的教育和醫療等借助大數據分析一展拳腳，學校通過資料分析來調整教學方法、衛生研究院等機構著手研發大數據分析醫療儀器、能源部鼓勵公眾高效利用能源、市政高校合作利用大數據分析提供城市交通解決方案。美國從互聯網滲透到工業，美國建設國家製造業創新網路給其強大背後支援的就是大數據分析，而美國的這種策略會使人工智慧等將在未來越來越普及。有關美國相關聯邦機構及地方政府大數據研發計畫如表 4 所示。

表 4 美國相關聯邦機構及地方政府大數據研發計畫

機構	項目重點
美國國家科學基金會與國家衛生研究院	推動大數據科學和工程的核心方法及技術研究，管理、分析、視覺化以及從大量的多樣化資料集中提取有用資訊的核心科學技術。
國防部以及國防部高	推進大數據輔助決策，集中在情報、偵查、網

機構	項目重點
級研究局	路間諜等方面，彙集感測器、感知能力和決策支援建立真正的自治系統，實現操作和決策的自動化。
美國能源部	改善資料集研究，通過先進的計算進行科學發現，建立可擴展的資料管理、分析和視覺化研究所。
美國地質調查局	給科學家提供深入分析的場所和實踐、最高水準的計算能力和理解大數據集的協作工具，催化在地理系統科學的創新思維。
西雅圖	大數據節能計畫，採用 Azure 平台，目標是降低耗電量 25%。
波士頓	IBM Smart Cities Challenge 計畫，採用 Info Sphere 平台，目標為優化交通。
孟菲斯	Blue CRUSH 預測型分析系統，通過犯罪預測來降低城市整體犯罪率。
紐約	將城市數據綜合分析，並達到政府行為透明以及政府績效考核提升。

資料來源：本計畫整理

英國是大數據的積極推動者，英國的政府、研究機構、企業都紛紛搶佔「資料革命」先機。2011 年 11 月，英國政府就發布對公開資料進行研究的戰略政策，英國政府早有意帶頭建立「英國資料銀行」；英國政府在 2013 年注資 1.89 億 英鎊發展大數據項目，稍後發布的《英國農業技術戰略》更是強調英國今後對農業技術的投資將集中在大數據上，讓英國的農業科技商業化，將英國打造成農業資訊學世界強國。

在亞洲，韓國的「智慧首爾地圖」就是各國智慧城市發展策略中的代表。通過一系列的手機應用，市民可以查詢殘疾人士無障礙設施、首爾市免費無線網路熱點、公廁、餐飲及行政資訊。在 2011 年，首爾就提出了「智慧首爾 2015」計畫，讓首爾成為世界上最方便使用智慧技

術的城市，建成適應未來生活的基礎設施、成為有創造力的智慧經濟都市。

在日本，安倍內閣於 2013 年 6 月發布「創建最尖端 IT 國家宣言」，全面闡述 2013 年至 2020 年間以發展開放公共資料和大數據為核心的國家戰略，強調「提升日本競爭力，大數據應用不可或缺」。

新加坡是世界網速最快的國家之一，在 2011 年 6 月啟用了政府分享公開資料平台，開放了來自 60 多個公共機構的近 9000 個資料庫。利用開放資料，企業和有關部門已開發 100 多項應用，涉及停車訊息、公廁甚至野貓管理等。

我國政府單位在政府統計大數據也頗有成績，在目前主要應用在五個部會，包括有：ETC 資料應用於制訂交通政策、財稅資料應用於稅制政策，改善貧富差距、電子發票用於分析商業情勢、健保資料用於特殊或亞太疾病研究，開發科技新藥與健保制度改進、圖資災防資料應用於國土規劃及救災、聯徵資料用於分析產業發展趨勢及國發會建置「政府物價資訊看板平台」，整合財政部、行政院消保處及主計總處等單位的資訊，可以查到賣場、便利商店近期的食品、飲料、日常用品售價。除了電子發票提供零售價格、銷售品項資訊之外，政府勞保統計、出入境統計也有助於我們了解國內就業、海外就業的情況。近年為了解國人赴海外就業情形，相關單位也開始運用入出境統計進行研判。經常引起各界關注的貧富差距問題，財政部將五百多萬戶家庭的稅檔資訊彙整成十等分位、二十等分位的所得分配統計，成為外界研判台灣貧富差距的重要參考。不論是電子發票、勞保統計、入出境資料、綜所稅檔皆屬公務統計，無需抽樣推估即可取得，有助於我們了解台灣經濟社會，其對政府抽樣調查所取得的相關統計也有相輔相成之效。

(二)發展趨勢

1. 大數據時代的會計、審核六大發展趨勢

面對大數據所帶來的新思維、新技術和方法的變革，政府會計、審核人員需要應時而變來適應思維模式及數據處理模式的變化。大數據對會計、審核發展的影響，主要表現在以下幾個方面：

(1)從事後的財務報告向即時財務報告發展

即時財務報告是資訊技術與大數據技術交叉融合的產物，是資訊化條件下會計技術和方法發展的必然產物。在大數據時代，政府單位要做到即時財務報告，首先要在政府內網中做到會計資訊系統和管理資訊系統的數據集成；其次是將政府內部內網與外部互聯網相連，即時財務報告系統中所用到數據則來源於外部互聯網和內網的中心數據庫。即時財務報告由會計人員對資料庫資訊進行網頁化處理後供用戶瀏覽,通過 ASP 等動態頁面生成技術即時生成所需的財務資訊頁面，為財務報告使用者提供即時的財務會計資訊。

(2)從會計的反映過去向預測未來發展

政府會計人員要做到從反映過去向預測未來發展，將需要做到以下三方面的工作：首先，要制定數據評估的方法和服務，在符合法規且有效管理數據資產方面，發揮其對內控方面的作用。其次，利用大數據提供更具針對性的決策支持，並決定何時與內部和外部利益相關對象分享數據最有效。最後，利用大數據及其相關工具並不只是為了適時識別風險和提高會計服務能力，而是為了評估生產經營活動中所面臨的短期和長期風險和規避。

(3)從財務管理理念向綜合管理理念發展

大數據的出現將顛覆現行財務管理的理念和模式，財務管理將不再局限於傳統的財務領域，而是向銷售、研發、人力資源等多個領域延伸和滲透，對於跟政府業務有關的一切數據的收集、處理和分析將成為財務管理的主要定位和主導任務。大數據時代的財務管理拓展了傳統財務管理的領域和範圍，一些原本不屬於傳統財務管理範疇的業務會進入大數據時代的財務管理視野，可以將其稱為「綜合財務管理」。

(4)從抽樣會計審核模式向總體會計審核模式發展

抽樣會計審核模式，由於抽取樣本的有限性，而忽視大量的業務活動，無法完全發現和揭示被審核單位的重大舞弊行為，隱藏著嚴重的會計審核風險。在大數據時代，數據的跨單位搜集和分析，可以不用隨機抽樣方法，而採用搜集和分析被審核單位所有數據的總體會計審核模式。大數據環境下的總體會計審核模式是要分析與會計審核對象相關的所有數據，使得會計審核人員可以建立總體會計審核的思維模式。

(5)從單一會計審核報告向綜合會計審核成果應用發展

目前，會計審核人員的審核成果主要是提供給被審核單位的會計審核報告，其格式固定，內容單一，包含的資訊較少。隨著大數據技術逐漸在會計審核中廣泛應用後，會計審核人員的審核成果除了審核報告外，還有在審核過程中採集、挖掘、分析和處理的大量的資料和數據，可以提供給被審核單位用於改進經營管理，促進審核成果的綜合應用，提高綜合會計審核成果的應用效果。

(6)從精確的數字會計審核向高效的數據會計審核發展

在大數據環境下，傳統的很多會計審核技術和方法顯得效率低下和無法實施，大數據時代的超大數據量體佔相當比例的半結構化和非結構化數據的存在，已經超越了傳統資料庫的管理能力，必須使用新的大數據儲存、處理和檢索方法。圍繞大數據，一批新興的資料採礦、數據儲存、數據處理與分析技術將不斷湧現。在實施會計審核時，會計審核人員應使用分布式拓樸結構、雲數據庫、聯網會計審核、資料採礦等新型的技術手段和工具，以提高會計審核的效率。

2. 大數據時代的統計發展趨勢

大數據與互聯網的發展相輔相成。一方面，互聯網資料是大數據中重要的資訊與資源。如 Google 等每天有大量使用者瀏覽資訊，並即時記錄關鍵字的搜索密度。隨著電子通訊和媒體技術的發展，傳統媒體報紙、廣播、電視也紛紛進入互聯網路時代，由於互聯網時代資訊傳播的廣域性和互動性，使得媒體資料以更快的速度出現。另一方面，大數據為互聯網的發展提供了更多支撐、服務與應用。大數據是互聯網發展到現今階段的一種表象或特徵，在以雲計算為代表的技術創新襯托下，這些原本很難收集和使用的資料開始變得容易利用。

對於政府統計而言，互聯網資料主要有社群網站資料、媒體資料和搜尋引擎資料三種類型。互聯網大數據在政府統計諸多專業中都具有廣闊的應用前景。如在宏觀層面，互聯網搜索資料能夠為官方統計提供分析、預測與決策支援。

(1)經濟發展

傳統官方統計按月度、季度或年度統計各項經濟指標，以GDP、社會消費品零售總額、固定資產投資額、採購經理指數等各項資料來分析經濟發展趨勢；而互聯網企業可以利用大數據來探索和完善各項經濟指標，及時有效地反映國民經濟運行狀況，提高總體經濟監測的全面性和及時性，為總體經濟部門把握經濟發展趨勢、監控企業景氣狀態提供分析、預測與決策支持。

(2)價格統計

在 CPI 統計方面，電子商務交易資料、互聯網企業資料都是價格統計的新資料來源，這些資料量大、更新快，充分利用這些資料有助於減少調查成本，提高指標發布的頻率。應用大數據進行價格統計的途徑有三種：一是採用搜索方式收集網上交易價格資料；二是與電子商務企業進行合作，獲取交易價格資料；三是建立商場、超市、醫院等實行電子計價的採價點向統計部門報送交易記錄的制度。

(3)批發零售業統計

由於網上電商交易資料量非常大、更新速度快，而且在全社會商品零售貿易中所占比重越來越大。因此，充分利用這些資訊可以為改善傳統的批發零售貿易業統計。

(4)人口統計

傳統官方統計投入大量人力物力財力，進行人口普查，可獲得資料包括全國和地區人口數量、城市和農村人口數量、人口性別比例、人口地域分布、年齡結構、出生率/死亡率等；而利用互聯網，可以快速及時地統計 PC 端和移動端使用者，統計維度包括地域、年齡、性別、學歷等，將來還可以根據民眾行為挖掘出群體的消費力水準、興趣點，更立體地洞察人群特徵。

(5)社會就業

傳統官方統計通過畢業生人數增長情況和勞動力需求增長情況的對比研究就業形勢，而互聯網大數據通過民眾對特定關鍵字的搜索趨勢就可以直觀地分析求職需求和就業壓力。如可以從「找工作」的搜索指數變動情況來瞭解求職需求動向，補充人力資源與社會保障部門資料的不足，輔助瞭解就業趨勢，把握就業需求，支持政策調整。

(6)醫療衛生

傳統官方統計通過醫療機構數量、診療人次等離線資料分析醫療服務情況，而互聯網大數據可以利用使用者線上行為資料研究疾病趨勢。利用民眾的疾病相關搜索資料，建立科學的預測模型，動態預測特定地域未來疾病的活躍指數，並呈現每個城市多種疾病的熱門醫院排名。互聯網搜索大數據能輔助衛生部門監測流行病發展態勢，提前做好預防措施，監督管理醫院。

(7)旅遊管理

傳統官方統計對旅遊人數的統計屬於事後統計，而基於民眾出遊前的網路搜索資料，得到使用者選擇的出行路線，可以預測旅遊趨勢。通過分析旅遊相關關鍵字搜索資料與實際出遊人數之間的密切關係，可以預測各旅遊景點未來的人流趨勢，進而輔助旅遊管理部門預警景點人潮流量。

第三節、大數據分析規劃及分析方法現況與發展趨勢

目前大數據分析的規劃及分析主要涉及資料採礦等一系列的方法，最被常使用的方法包括有針對結構性資料分析的資料採礦、資料映射、時間序列及決策樹預測法等。針對非結構性資料的分析法包括文字探勘法與情緒分析法。深度學習是人工智慧中成長最為快速的領域，可協助電腦理解影像、聲音和文字等資料。現在透過多層級的神經網路法已成為大數據分析重要的發展趨勢，以下將針對每個方法進一步說明。

一、資料採礦（Data Mining）

根據 Berry and Linoff（1997）的定義：「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；Grupe and Owrang（1995）則認為「資料採礦」乃是從現存資料中剖析出新事實及發現專家們尚未知曉的新關係。在 Fayyad et al.（1996）的定義中，對於知識發現（knowledge discovery in database， KDD）與資料採礦的意義則有著不同的解釋。其對知識發現的定義為：它是一個指出資料中有效、嶄新、潛在效益的一個非細瑣（nontrivial）流程，其最終的目標是瞭解資料的樣式（patterns），而在進行知識發現時其主要的步驟可以參考圖 2。

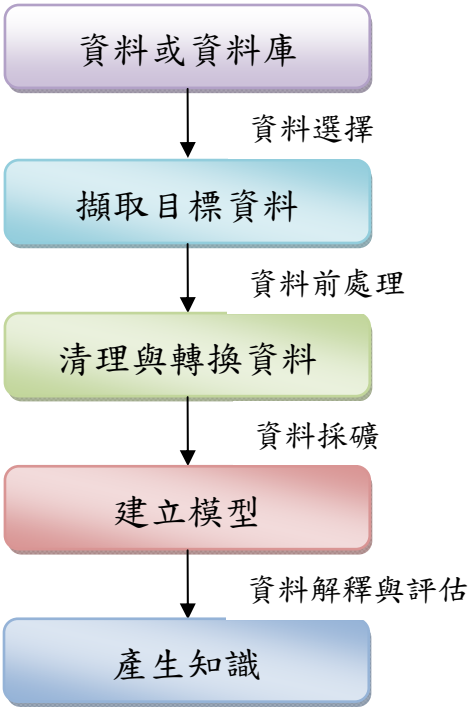


圖 2 從資料到知識的發現流程

（修改自 Fayyad et al., 1996 的 KDD Process）

Chapman 等人（2000）提出可應用於各產業領域的資料採礦標準過程 CRISP－DM（Cross Industry Standard Process for Data Mining），此過程涵蓋資料採礦中最主要的六個階段。此六個階段構成了一個迴圈過程，如圖 3 所示（謝邦昌，2001）。

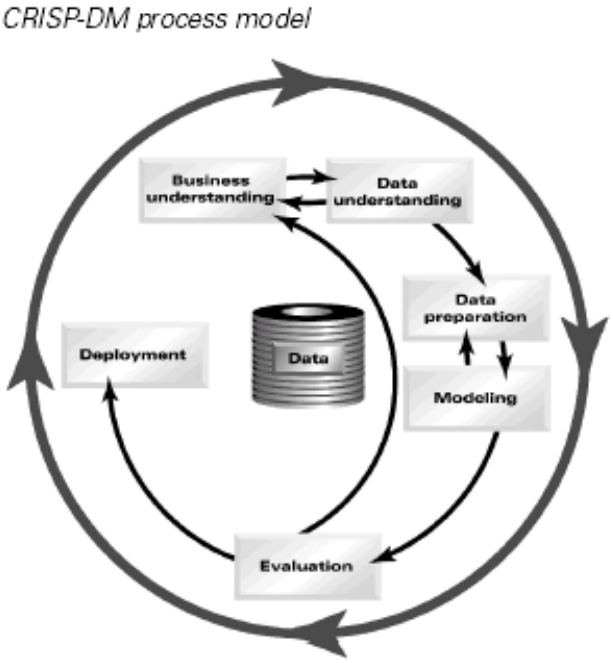


圖 3 CRISP-DM 過程模型（取自 Chapman et al, 2000）

此六個階段涵蓋資料採礦的全部過程，分別是：

（一）專業領域理解（Business Understanding）：

專業領域理解包括決定專業領域的目標，環境評估，決定資料採礦目標及制定一個專案計劃，此為資料採礦中最重要的部分。

（二）資料理解（Data Understanding）：

此階段強調需要清楚資料來源是什麼，其特徵為何。此階段包括收集 原始資料，描述資料，探索資料，及驗證資料的品質。

（三）資料準備（Data Preparation）：

在將資料源分類完成後，需要準備用於後續採礦的資料。準備過程包括選樣、清理、重構、整合及格式化資料等。

（四）建模（Modeling）：

成熟與完善的分析方法與建模，將能夠從資料中提取資訊。這個部分包括選擇模型技巧、產生測試計劃、建模和模型評估。

（五）評估（Evaluation）：

一旦選擇模型，即應準備好對資料採礦的結果進行評估，並了解

是否符合資料採礦最初設定的目標。這個部分包括評估結果、回顧資料採礦過程及確定接下來的步驟。

（六）展開（Deployment）：

這個階段著重於將新知識融入到每天的商業運作過程中，從而解決最初的商業問題。此部分包括計劃發布、監控與維護、產生最終報告及回顧整個專案。

在這個 CRISP-DM 標準過程中，在前五個階段之間存在以非線性方式互相影響的地方。例如，資料準備通常在建模之前，但是在建模過程中進行決策和資訊收集時，也會讓你重新考慮資料準備方面的問題，而這些也會讓你重新考慮建模方式，以此循環往復。相似地，評估過程也會使你重新評估最初的專業領域理解。此時，在新的目標下，可以修訂原來的專業領域理解和後續產生建模的過程及其產生的模型。

一般而言，資料採礦（Data Mining）可包含下列五項功能：分類（classification）、推估（estimation）、預測（prediction）、關聯分組（affinity grouping）、同質分組（clustering）。這些功能的意義及可能使用的技巧簡述如下：

（一）分類：

按照分析對象的屬性分門別類加以定義，建立類組（class）。例如，將創新研發企業依其創新研發能力，區分為高度創新研發企業、中度創新研發企業及低度創新研發企業。

（二）推估：

根據既有連續性數值之相關屬性資料，以獲致某一屬性未知之值。例如按創新研發企業之產業別、企業規模、地區等來推估其創新研發投入經費與能力的情況。使用的技巧包括統計方法上之相關分析、迴歸分析及類神經網路方法。

（三）預測：

根據對象屬性之過去觀察值來推估該屬性未來之值。例如由企業

單位過去之創新研發能力預測其未來之創新研發能力。使用的技巧包括迴歸分析、時間數列分析及類神經網路方法。

(四)關聯分組：

從所有物件決定那些相關物件應該放在一起。例如超市中相關之盥洗用品（牙刷、牙膏、牙線），放在同一間貨架上。在客戶行銷系統上，此種功能係用來確認交叉銷售（cross-selling）的機會以設計出吸引人的產品群組。

(五)同質分組：

將異質母體中區隔為較具同質性之群組（clusters），同質分組相當於行銷術語中的區隔化（segmentation），但是此時假定事先未對於區隔加以定義，而資料中自然產生區隔。常用的技巧包括 k-means 法。

二、資料映射（Data Mapping）

建置關聯式資料庫之時，對於兩個或兩個以上不同的關聯表或資料庫之間都會利用另一個關聯表來串連，此關聯表的作用即是銜接兩個相對應的資料庫。若此關聯表不存在時，於是資料表或資料庫之間就無法直接合併了。本研究的資料來源為跨資料庫，即關聯表自然不存在，所有資料間無法直接串聯形成新的關聯式資料庫。

資料映射（Data Mapping）之研究概念，可分為兩個部分，第一部份在於先判別輔助資料庫與目標資料庫具有相關性的欄位，即虛擬索引（Virtual Indices，簡稱 VI），此步驟目的在找出兩資料庫間的關連欄位，以提供成為欲映射欄位從輔助資料庫轉換到目標資料庫的轉換依據；第二部分則是要以輔助資料庫之資料表的虛擬索引欄位，建立欲映射欄位的推估模型，藉由推估模型與第一部分所找出的兩資料庫虛擬索引的對應，便可將欲映射欄位推估到目標資料庫之資料表中，在目標資料庫中產生出拓增欄位。以圖 4 舉一個簡單的例子，假設欲將目標資料庫之資料表拓增 A′、B′、C′、D′四個欄位，且輔助資料庫之資料表中有此四個欄位（分別為 A、B、C、

D)，則概念上的實行步驟為：

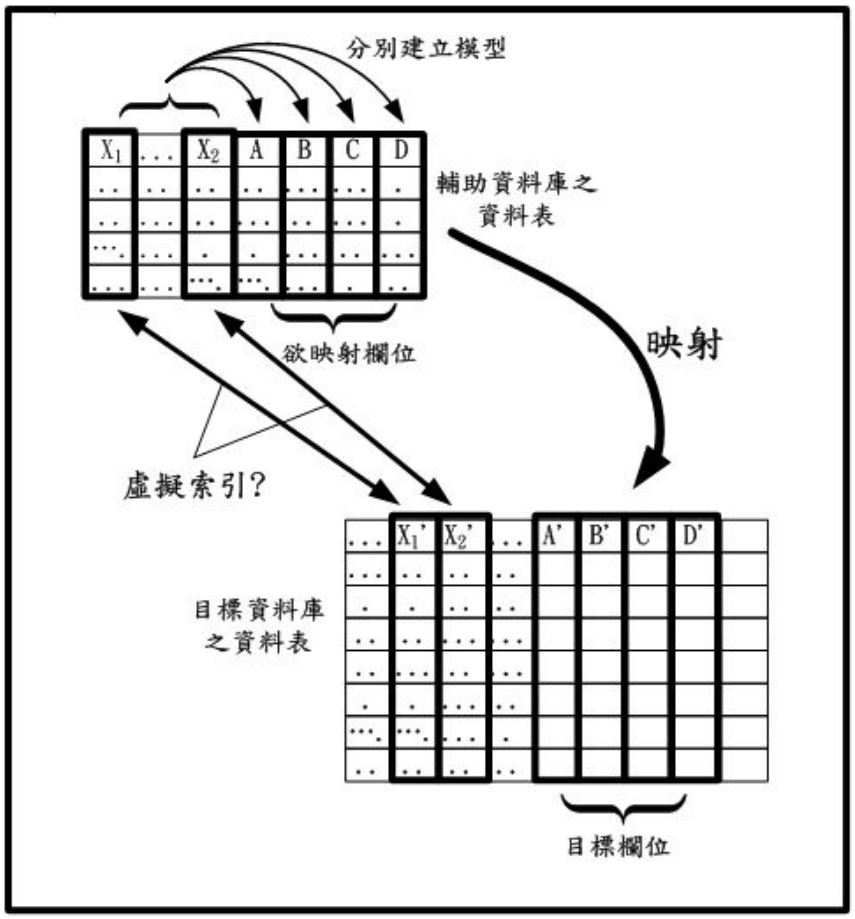


圖 4 資料映射概念圖

(一)找出輔助資料之資料表與目標資料之資料表可對應的欄位，如圖中的 X_1、X_1'與 X_2、X_2'。

(二)判別 X_1 與 X_1'、X_2 與 X_2'之分布是否相同，以判別 X_1、X_2 可否成為虛擬索引。

(三)假設 X_1、X_2 可成為虛擬索引，即以 X_1、X_2 為輸入，A 為輸出進行分類模型的訓練。建立出分類模型後，即可將 X_1'、X_2'輸入此模型，推估出 A′的值。

(四)以 B、C 與 D 取代 A，重複步驟 2、3 的運算，即可推估出 B′、C′與 D′欄位之值。

三、時間序列預測法（Time Series）

時間序列，亦稱時間數列、歷史複數或動態數列。它是將某種統計指標的數值，按時間先後順序排到所形成的數列。時間序列預測法就是通過編製和分析時間序列，根據時間序列所反映出來的發展過程、方向和趨勢，進行類推或延伸，藉以預測下一段時間或以後數年內可能達到的水平。其內容包括：收集與整理某種社會現象的歷史資料；對這些資料進行檢查鑑別，排成數列；分析時間數列，從中尋找該社會現象隨時間變化而變化的規律，得出一定的模式；以此模式去預測該社會現象將來的情況。

時間序列預測法是一種根據動態資料揭示系統動態結構和規律的統計方法，時間序列的獲取是通過對已有資料進行分類彙總後得到的，獲取時間序列資料以後可以對其進行預測，從而準確地預測系統的演進（趙仁義&朱玉輝,2011）。在大數據背景下對學生的學習過程或學習結果進行預測時使用時間序列法可將隨著時間 t 變化的量，即在 $t_1<t_2<\ldots<t_n<\ldots$時所產生的階段性學習結果的觀測值記為 $y(t_1), y(t_2), \ldots, y(t_n)$，…組成離散有序集合，稱為一個時間序列，記作 $\{y(t)\}$。由於不同人在學習過程中的時間序列是由某一隨機過程產生的，因此可將基於時間序列的學習預測記作公式：$y(t) = f(t) + p(t) + x(t)$，其中 $f(t)$ 為趨勢項反映 $y(t)$ 的變化趨勢；$p(t)$ 為週期項，反映 $y(t)$ 不同階段的週期變化；$x(t)$ 為隨機項，反映隨機因素對 $y(t)$ 的影響。

時間序列分析有二大特徵：

（一）時間序列分析法是根據過去的變化趨勢預測未來的發展，其前提是假定事物的過去延續到未來。

時間序列分析，正是根據客觀事物發展的連續規律性，運用過去的歷史數據，通過統計分析，進一步推測未來的發展趨勢。事物的過去會延續到未來這個假設前提包含兩層含義：一是不會發生突然的跳躍變化，是以相對小的步伐前進；二是過去和當前的現象可

能表明現在和將來活動的發展變化趨向。以上意涵就決定了在一般情況下，時間序列分析法對於短、近期預測比較顯著，但如延伸到更遠的將來，就會出現很大的局限性，導致預測值偏離實際較大而使決策失誤。

（二）時間序列數據變動存在著規律性與不規律性

時間序列中的每個觀察值大小，是影響變化的各種不同因素在同一時刻發生作用的綜合結果。

四、決策樹預測法（Decision Tree）

由一個決策圖和可能的結果（包括資源成本和風險）組成，用來建立到達目標的規劃。決策樹建立用來輔助決策，是一種特殊的樹結構。決策樹是一個利用像樹一樣的圖形或決策模型的決策支持工具，包括隨機事件結果，資源代價和實用性。它是一個演算法顯示的方法。決策樹經常在作業研究(Operations Research, OR)中使用，特別是在決策分析中，其幫助確定一個能最可能達到目標的策略。

機器學習中，決策樹是一個預測模型；其代表的是對象屬性與對象值之間的一種映射關係。樹中每個節點表示某個對象，而每個分叉路徑則代表的某個可能的屬性值，而每個葉結點則對應從根節點到該葉節點所經歷的路徑所表示的對象的值。決策樹僅有單一輸出，若欲有複數輸出(為實數的延伸，它使任一多項式方程式都有根)，可以建立獨立的決策樹以處理不同輸出。資料採礦中決策樹是一種經常要用到的技術，可以用於分析數據，同樣也可以用來作預測。

決策樹學習也是資料採礦中之一普通方法。在此，每個決策樹都表述了一種樹型結構，它的分支來對該類型的對象依靠屬性進行分類。每個決策樹可以對原始資料的分割進行數據測試。這個過程可以遞歸式的對樹進

行修剪。當不能再進行分割或一個單獨的類可以被應用於某一分支時，遞歸過程就完成了。另外，隨機森林分類器[1]將許多決策樹結合起來以提升分類的正確率。

五、微軟 Office Excel Add-In 進行關聯分析

若有需要開發複雜的統計或工程分析，使用 Excel [分析工具箱] 可以節省不少步驟和時間。只需要為每個分析提供資料和參數，分析工具就會使用正確的統計或工程巨集函數計算，並在輸出表格中顯示結果。有些工具在產生輸出表格時，還可以同時繪製圖表（如圖 5）。

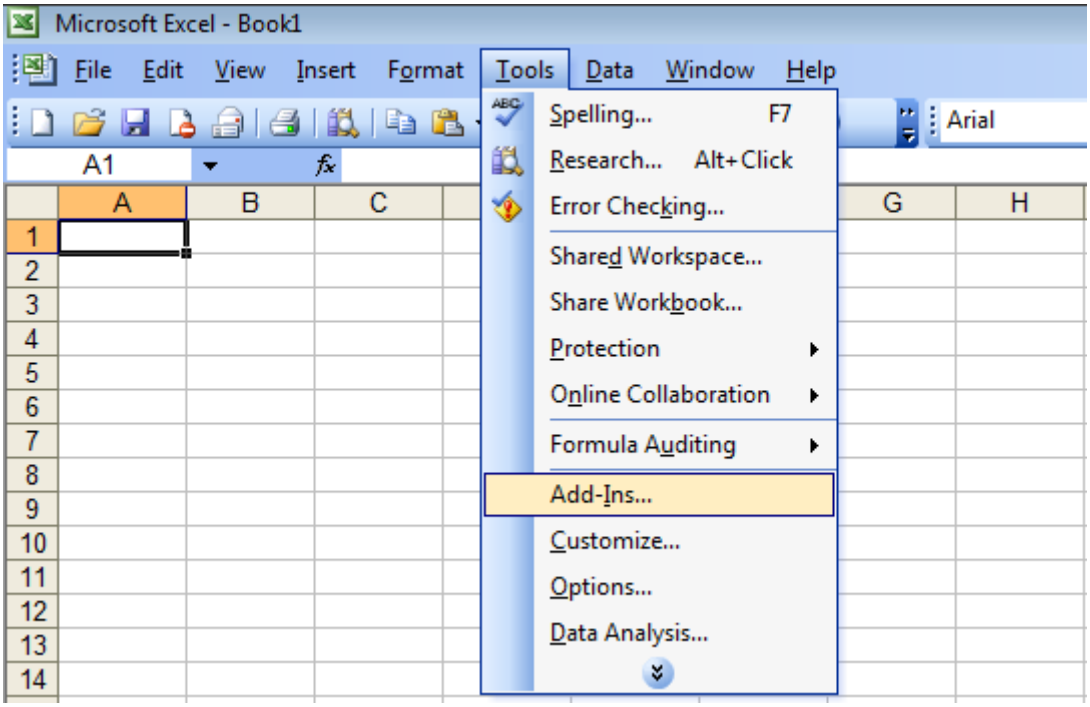


圖 5 微軟 Office Excel Add-In 示意圖

六、情緒分析（Sentiment Analysis）

[1] 隨機森林分類器: 在機器學習中，隨機森林是一個包含多個決策樹的分類器，並且其輸出的類別是由個別樹輸出的類別的眾數而定。這個方法是結合 Breimans 的"Bootstrap aggregating"想法和 Ho 的"random subspace method" 以建造決策樹的集合。

將情緒與意圖進行量化會遭遇到一些困難及障礙，包含:俚語、諷刺、誇張的詞彙可能會阻礙數據的分析。情感分析（Sentiment Analysis）可以被歸類為一個分類問題，其任務是將訊息分為兩類，以它們所傳達的是不同情感來決定。以 Twitter 為例，Twitter 的數據進行情感分析將耗費大量金錢與精力，但是 Twitter 的數據的顯著優點是很多的動態消息有作者提供的情緒指標。當發動態消息的作者使用表情符號，他們將自己的發言標註上情緒狀態。這樣被註解的動態消息可以用來訓練情感分類器（Sentiment Classifier）（Bifet, 2013），然而人類的情緒表達是會隨著時間、地域、國家不同而改變的，因此分析的演算法與判斷依據也會因此而不同，情緒判斷的準確性也會因此降低。

七、文字探勘 (Text Mining)

資料採礦（Data Mining）與文字探勘（Text Mining）關係緊密，相較於前者顯著的結構化，後者長短不一、沒有規律，且尚有現今生活中隨口說出來的字彙，或是從社群網站、BBS 衍生出的用語，都屬於非結構化的資料。又像是 Web 頁面、新聞張貼，以及電子郵件訊息等文件，都包含非結構性、自由型態的文字，或是許多符合特定電腦語言的語法及文法規則，構成文字和語句的字串流。這些非結構化的資料，隨著使用者每天更新及增減而無法事先定義。一般說來，非結構化的文字探勘功能主要是從獲得的文本資料中，分析訊息得到的詞頻基本特質，建立詞頻關聯表與正負向情緒判別，因此非結構化的文字探勘可以進行字詞的重要性分析、關聯性分析、集群分析、文章的正負向情緒分析、語意分析、脈絡分析...等。

面臨事件當中報導陳述的真實意義可能會因為文字的前後聯結不同而有所差異，並且有可能因為不同的生活經驗、受教育情形與社經背景的不同而對同一段文字的意義有不同的判斷，大數據分析如何進行文字語意的判斷，演算法如何分析複雜的語言關係與文字詞義結構，將是大數據分析在運用上將受到的質疑。文字探勘的技術類似社會科學研究方法中的內容

分析法，Berelson（1952）定義內容分析方法是針對傳播的明顯內容，做客觀、系統與定量的研究，內容分析方法是運用量化的分析策略，將所欲分析文本中的文字進行斷字與斷句的拆解，拆解為可供分析的單位（字、句、陳述等），接著針對文本進行編碼表的製作，訓練編碼員將分析文本進行量化的編碼，再進行統計分析。大數據分析則是運用類似的方法進行文字探勘，差別在於大數據分析方法是設計出一套演算法進行斷字與斷句，並且將正負面語意例字與例詞寫入程式中訓練演算法自動進行判斷。然而內容分析方法在方法論上的客觀性也遭到挑戰，包括編碼員的訓練、分析單位、信效度甚至於研究的可重製性都備受質疑（游美惠，2000）。大數據分析所運用的演算法是使用電腦軟體進行分析，政府應用大數據精進公共服務與政策分析之可行性研究可減去因編碼員差異導致的差異，而演算法程序、斷字斷句邏輯與判斷正確性都是可能存在的問題。

八、Fanpage

Fanpage Karma 能分析 Facebook 粉絲團、Twitter、Instagram 等，操作簡單，並提供完善的分析報告，各時段貼文表現、粉絲喜好、廣告價值、競品分析，還有最新功能是結合 Pinterest、Google Analytics。Fanpage Karma 提供使用者建立網絡關係圖，讓使用者可以知道自己的網站與那些相關網站是有關聯性的，如圖 6。

圖 6 關注者流向

九、深度學習（Deep Learning）

20 世紀 80 年代末期，用於人工神經網路反向傳播演算法（Backpropagation，簡稱為 BP）[2]的發明，給機器學習帶來了希望，掀起了基於統計模型的機器學習熱潮。這個熱潮一直持續到今天。利用 BP 演算法可以讓一個人工神經網路模型從大量訓練樣本中學習統計規律，從而對未知事件做預測。這種基於統計的機器學習方法比起過去基於人工規則的系統，在很多方面顯出優越性。這個時候的人工神經網路，雖也被稱作多層感知機（Multi-layer Perceptron）[3]，但實際是種只含有一層隱層節點的淺層模型。20 世紀 90 年代，各種各樣的淺層機器學習模型相繼被提出，例如支撐向量機（SVM，Support Vector Machines）[4]、Boosting[5]。這些模型的結

[2]反向傳播演算法（Backpropagation，簡稱為 BP）：是「誤差反向傳播」的簡稱，是一種與最優化方法（如梯度下降法）結合使用的，用來訓練人工神經網路的常見方法。該方法計算對網路中所有權重計算損失函數的梯度。這個梯度會反饋給最優化方法，用來更新權值以最小化損失函數。

[3]多層感知機（Multi-layer Perceptron）：是一種前向結構的人工神經網路，映射一組輸入向量到一組輸出向量。

[4]支撐向量機（SVM，Support Vector Machines）：是一種監督式學習的方法，可廣泛地應用於統計分類以及回歸分析。

[5]Boosting: 對一份數據，建立 M 個模型（比如分類），一般這種模型比較簡單，稱為弱分類器(weak learner)每次分類都將上一次分錯的數據權重提高一點再進行分類，這樣最終得到的分類器在測試數據與訓練數

構基本上可以看成帶有一層隱層節點（如 SVM、Boosting），或沒有隱層節點（如羅吉斯回歸[6]）。這些模型無論是在理論分析或是應用皆能獲得很好的成果展現。相較之下，由於理論分析的難度較大，訓練方法又需要很多經驗和技巧。

2006 年，加拿大多倫多大學教授開啟了深度學習在學術界和工業界的浪潮。目前多數分類、迴歸等學習方法為淺層結構演算法，其在有限樣本和計算單元情況下，對於複雜函數的表示能力有限，深度學習可通過學習一種深層非線性網路結構，實現複雜函數逼近，表徵輸入資料分散式表示，並展現了強大的從少數樣本集中學習資料集本質特徵的能力。深度學習的實質，是通過構建具有很多隱層的機器學習模型和海量的訓練資料，來學習更有用的特徵，從而最終提升分類或預測的準確性。深度分析推動大數據智慧應用預測決策、精準推薦、語義化，這些涉及人的思維、影響、理解的延展，都成為大數據深度分析的關鍵應用方向。深度學習成為大數據智慧分析的核心技術相比於傳統機器學習演算法，深度學習借助深層次神經網路模型，能夠更加智慧的提取資料不同層次的特徵，對資料進行更加準確有效的表達。而且資料量越大，深度學習演算法越有優勢，可以得到更好的結果。目前，深度學習已經在圖像分類檢索、語音辨識等領域產生重大突破。深度學習的興起凸顯出複雜機器學習模型在利用大數據方面的突出優勢，進一步考慮大數據動態性、分布性、關聯性的新型機器學習技術將很快湧現。

據上都可以得到比較好的成績。

[6] 羅吉斯迴歸(Logistic regression): 是離散選擇法模型之一，屬於多重變量分析範疇，是社會學、生物統計學、臨床、數量心理學、計量經濟學、市場行銷等統計實證分析的常用方法。

第四節、大數據分析技術工具現況與發展趨勢

在目前大數據的分布式數據處理、串流式數據分析、機器學習以及大規模數據分析技術工具可以分為技術平台與常見的處理工具。隨著視覺化分析的成長應用趨勢，視覺化分析技術工具已漸漸成為發展趨勢，以下即針對技術平台與技術工具進一步說明。

一、技術平台

(一)Hadoop

Hadoop 是一個能夠對大量數據進行分布式處理的軟體框架，也是以一種可靠、高效、可伸縮的方式進行處理的工具，因為它假設計算元素和存儲會失敗，因此它維護多個工作數據副本，確保能夠針對失敗的節點重新分布處理。Hadoop 可以並行的方式工作，通過並行處理加快處理速度。Hadoop 也能夠處理 PB 級數據。Hadoop 它的成本比較低，任何人都可以使用。

Hadoop 是一個能夠讓用戶輕鬆架構和使用的分布式計算平台。用戶可以輕鬆地在 Hadoop 上開發和運行處理海量數據的應用程序。它主要有以下幾個優點：

1. 高可靠性:Hadoop 按位元儲存和處理數據的能力值得人們信賴。
2. 高擴展性：Hadoop 是在可用的電腦（或伺服器）集簇間分配數據並完成計算任務，這些集簇可以方便地擴展到數以千計的節點中。
3. 高效性：Hadoop 能夠在節點之間動態地移動數據，並保證各個節點的動態平衡，因此處理速度非常快。
4. 高容錯性：Hadoop 能夠自動保存數據的多個副本，並且能夠自動將失敗的任務重新分配。

Hadoop 帶有用 Java 語言編寫的框架，因此運行在 Linux 生產平台上是非常理想的。Hadoop 上的應用程序也可以使用其他語言編寫，比如 C^{++}。

(二)HPCC（High Performance Computing And Communications）

1993 年，由美國科學、工程、技術聯邦協調理事會向國會提交了「重大挑戰項目：高性能計算與通信」的報告，也就是被稱為 HPCC 計劃的報告，即美國總統科學戰略項目，其目的是通過加強研究與開發解決一批重要的科學與技術挑戰問題。HPCC 是美國實施高性能計算和通信計畫，該計畫的實施將耗資百億美元，其主要目標要達到開發可擴展的計算系統及相關軟體，以支持太位級（TB）網路傳輸性能，擴展研究和教育機構及網路連接能力。

該平台主要由五部分組成：

1. 高性能計算機系統（HPCS），內容包括今後幾代計算機系統的研究、系統設計工具、先進的典型系統及原有系統的評價等。
2. 先進軟體技術與演算法（ASTA），內容有巨大挑戰問題的軟體支撐、新算法設計、軟體分支與工具、計算及高性能計算研究中心等。
3. 國家科研與教育網路（NREN），內容有中接站及 10 億位級傳輸的研究與開發。
4. 基本研究與人類資源（BRHR），內容有基礎研究、培訓、教育及課程教材，通過高性能的計算訓練，來提供必需的基礎架構以支撐這些調查和研究活動。
5. 訊息基礎結構技術和應用（IITA），目的在於保證美國在先進訊息技術開發方面的領先地位。

(三)Storm

Storm 的設計針對的是流式數據，用於大數據的即時分析，亦是很可靠的計算系統。它同樣是一個開源項目而且開發人員可以使用所有的主流高級語言。Storm 主要用於以下應用：在線機器學習、連續計算、即時分析、ETL（Extraction-Transformation-Loading）的縮寫，即數據抽取、轉換及加載)、分布式 RPC（一種通過網路從遠端請求服務的電腦）。Apache Storm 有配置方便、可用性高、容錯性好及擴展性好等諸多優點，處理速度也極快，每個節點每秒可以處理數百萬個 tuple。

(四)Apache Drill

為了幫助企業用戶尋找更為有效、加快 Hadoop 數據查詢的方法，Apache 軟體基金會最近發起了一項名為"Drill"的開源項目。該項目將會創建出開源版本的 Google Dremel Hadoop 工具（Google 使用該工具來為 Hadoop 數據分析工具的互聯網應用提升速率）。而"Drill"將有助於 Hadoop 用戶做到更快查詢海量數據集的目的。

“Drill"項目其實也是從 Google 的 Dremel 項目中獲得靈感，該項目幫助 Google 做到海量數據集的分析處理，包括分析抓取 Web 文檔、跟蹤安裝在 Android Market 上的應用程序數據、分析垃圾郵件、分析 Google 分布式構建系統上的測試結果等等。通過開發"Drill" Apache 開源項目，組織機構將有望建立 Drill 所屬的 API 介接及彈性大的體系架構，從而幫助支持廣泛的數據源、數據格式和查詢語言。

(五)RapidMiner

RapidMiner 是世界領先的資料採礦解決方案，其資料採礦任務涉及範圍廣泛，包括各種數據藝術，能簡化資料採礦過程的設計和評價。功能和特點如下所示：

1. 免費提供資料採礦技術和資料庫
2. 100%用 Java 代碼（可運行在操作系統）
3. 資料採礦過程簡單、強大和直觀
4. 內部 XML 保證標準化的格式來表示交換資料採礦過程
5. 可以用簡單腳本語言自動進行大規模進程
6. 多層次的數據視圖，確保有效和透明的數據
7. 圖形用戶界面的互動原型
8. 命令行（批處理模式）自動大規模應用
9. Java API（應用編程接口）
10. 簡單的插件和推廣機制
11. 強大的視覺化引擎，許多尖端的高維數據的視覺化建模
12. 400 多個資料採礦經營商支持

耶魯大學已成功地應用在許多不同的應用領域，包括文字探勘、多媒體數據挖掘、功能設計、數據流挖掘、整合開發環境方法和分布式資料採礦。

(六)Apex

Apex 是一個企業級的大數據動態處理平台，能夠支持即時的串流式數據處理，也可以支持批次數據處理。它可以是一個 YARN 的原始程序，能夠支持大規模、可擴展、支持容錯方法的串流式數據處理引擎。Apex 支持一般事件處理並保證數據一致性（精確一次處理、最少一次、最多一次）。以前 DataTorrent 公司開發的基於 Apex 的商業處理軟體，其代碼、文檔及架構設計顯示，Apex 在支持 DevOps 方面能夠把應用開發清楚的分離，用戶代碼通常不需要

知道他在一個流媒體處理集群中運行。Malhar 是一個相關項目，提供超過 300 種常用的實現共同的業務邏輯的應用程式模板。Malhar 的連結庫可以顯著的減少開發 Apex 應用程式的時間，並且提供了連接各種存儲、文件系統、消息系統、資料庫的連接器和驅動程序。並且可以進行擴展或定製，以滿足個人業務的要求，所有的 malhar 組件都是 Apache 許可下使用。

二、資料採礦工具

百分之八十的數據是非結構化的，需要一個程序和方法來從中提取有用資訊，並且將其轉換為可理解、可用的結構化形式。資料採礦過程中有大量的工具可供使用，比如採用人工智慧、機器學習，以及其他技術等來提取數據。目前常用的有六種資料採礦工具：

(一) RapidMiner

該工具是用 Java 語言編寫的，通過基於模板的框架，提供先進的分析技術。該款工具最大的好處就是，用戶無需寫任何代碼。它是作為一個服務提供。除了資料採礦，RapidMiner 還提供如數據前處理和視覺化、預測分析和統計建模、評估和部署等功能。它還提供來自 R 的學習方案、模型和算法。RapidMiner 分布在 AGPL 開源許可下，可以從 SourceForge 上下載。

(二) WEKA

WEKA 原生主要是為了分析農業領域數據而開發的。該工具基於 Java 版本，是非常複雜的，並且應用在許多不同的應用中，包括數據分析以及預測建模的視覺化和演算法。與 RapidMiner 相比優勢在於，它在 GNU 通用公共授權下是免費的，因為用戶可以按照自己的喜好選擇自定義。WEKA 支持多種標準資料採礦任務，包括數據前處理、收集、分類、回歸分析、視覺化和特徵選取。

(三)NLTK

NLTK 提供了一個語言處理工具，包括資料採礦、機器學習、數據抓取、情感分析等各種語言處理任務。需要做的只是安裝 NLTK，然後將一個包拖拽到最喜愛的任務中就可以了。因為它是用 Python 語言編寫的，可以在上面建立應用，還可以自定義它的小任務。

(四)Orange

Python 之所以受歡迎，是因為它簡單易學並且功能強大。如果是一個 Python 開發者，當涉及到需要找一個工作用的工具時，那麼沒有比 Orange 更合適的了。它是一個基於 Python 語言，功能強大的開源工具，並且對初學者和專家級使用者均適用。此外，它不僅有機器學習的組件，還附加有生物資訊和文本挖掘，可以說是充滿了數據分析的各種功能。

(五)KNIME

KNIME 提供了一個圖形化的用戶介面，以便對數據節點進行處理。它是一個開源的數據分析、報告和綜合平台，同時還通過其模塊化數據的流水型概念，集成了各種機器學習的組件和資料採礦，並引起了商業智能和財務數據分析的注意。KNIME 是基於 Eclipse，用 Java 編寫的，並且易於擴展和補充插件。

(六)R-Programming

它主要是由 C 語言和 FORTRAN 語言編寫的，並且很多模塊都是由 R 編寫的，這是一款針對程式語言和軟體環境進行統計計算和製圖的免費軟體。R 語言被廣泛應用於資料採礦，以及開發統計軟體和數據分析中。近年來，易用性和可擴展性也大大提高了 R 的知名度。除了數據，它還提供統計和製圖技術，包括線性和非線性建模，經典的統計測試，時間序列分析、分類、收集等等。

以主計總處統計專區薪資及生產力統計、受僱員工薪資調查統計為例，發現民國 90 年~民國 104 年因為資料有時間性可以探討時間序列，分析步驟如下說明（如圖 7 至圖 16 表示）。

1. 在「分析方法」選單中，選擇時間序列，及點選想要的建模方式。

圖 7 R-Web 介面-功能選擇

2. 步驟二進行參數設定。依序輸入依變數及自變數，選擇是否要對變數進行轉換，選擇完畢後，可以選擇由系統自動選取最佳模型預覽。

圖 8 R-Web 介面-參數設定

3. 步驟三可以選擇自己想要做的差分及自我回歸 P、移動平均 Q 之參數選擇。在輸入參數設定按下開始分析，按下開始分析後，即可完成一些數列圖，如下面三張圖，以下圖 9 是未經過差分。

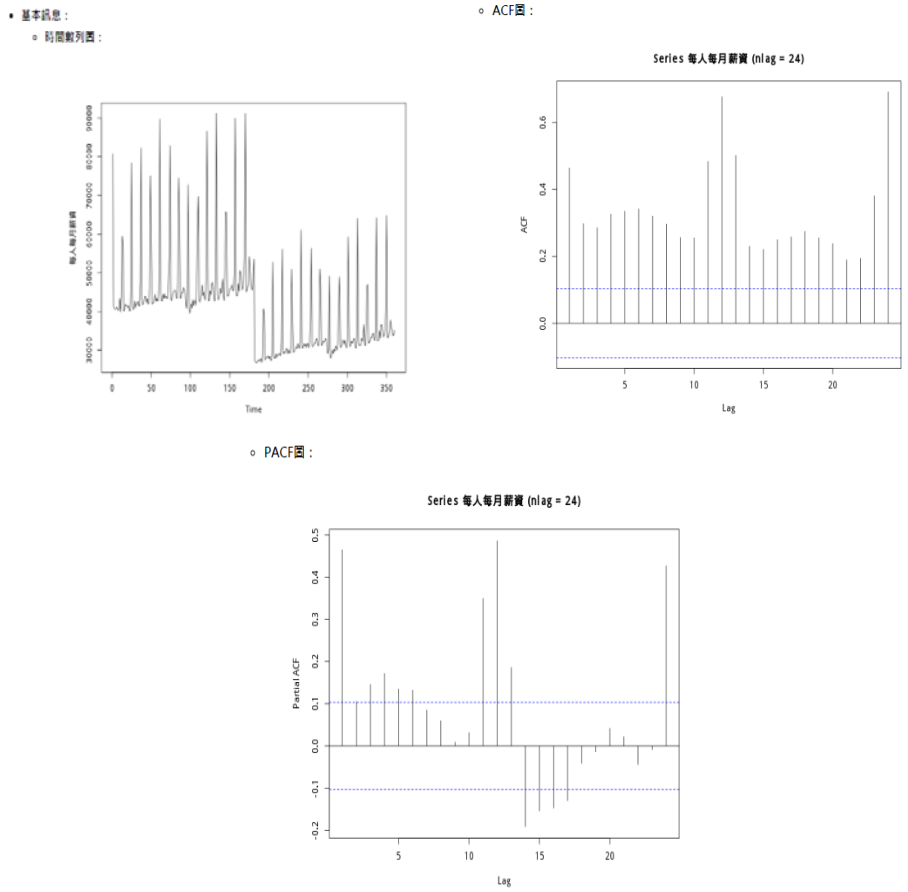


圖 9 時間序列未差分

4. 最後會跑出以下圖 10 參數之估計以及檢定之 P 值，也可以對殘差做檢定。在圖 10 中可以分別看到，AR1 模型及 AR2 模型的 P-value 值，以檢視模型的可用度。圖 11 針對殘差進行獨立性檢定，可以從檢定 p-value 值檢視模型中的殘差是否獨立。

- 模式配適：
  - 模式係數估計：

係數 coefficient	估計值 estimation	標準差 Std. err.	t-統計量 t-value	p-值 p-value
AR1	-0.4338	0.0513	-8.4561	< 1e-04
AR2	-0.297	0.0512	-5.8008	< 1e-04

圖 10 模式係數估計

- 殘差分析：
  - 殘差資料獨立性(Independence)檢定：

虛無假設：資料互相獨立			
檢定方法 method	卡方統計量 Chi-square statistic	自由度 d.f.	p值 p-value
Box-Pierce檢定	1.4821	1	0.2235
Box-Ljung檢定	1.4944	1	0.2215

圖 11 殘差資料獨立性檢定

5. 以預測 30 筆資料為例，從圖 12 及圖 13 可以發現，資料第 337 筆的每人每月消費值突然升高，此時間點為 12 月份，猜測可能有獎金發放，所以造成真實消費值較高。

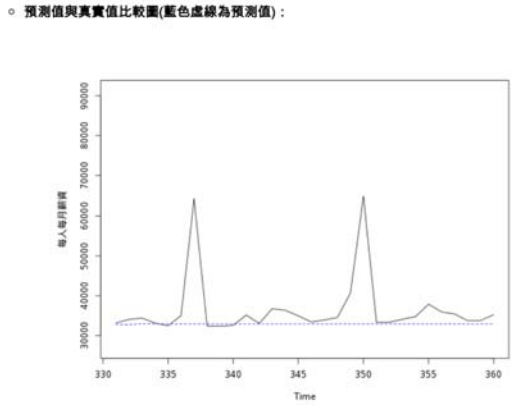


圖 12 預測值與真實值比較圖

- **資料預測：**
  - **預測值資料表：**

資料筆數	真實值	預測值	絕對誤差百分比%
331	33212	32873.9107	1.018
332	34069	32804.8729	3.7105
333	34413	32986.8309	4.1443
334	33171	32929.9407	0.7267
335	32487	32900.212	1.2719
336	35053	32929.7044	6.0574
337	64282	32925.9761	48.7789
338	32459	32918.8228	1.4166
339	32357	32922.9681	1.7491
340	32600	32923.3255	0.9918
341	35176	32921.9454	6.4079
342	33073	32922.4261	0.4553
343	36751	32922.6307	10.417
344	36375	32922.4013	9.4917
345	34985	32922.4382	5.8956
346	33429	32922.4905	1.5152

圖 13 預測值資料表

6. 以歲計為例，輸入主計總處預算執行與政府歲入決算之合計，選擇民國 89 年~民國 103 年時間段開始分析，得到下圖 14 是未經過差分的時間數列圖，隨時間呈現正向成長趨勢。

- 基本訊息：
  - 時間數列圖：

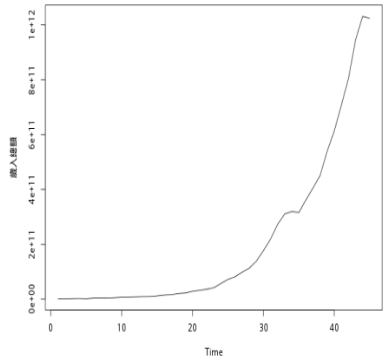


圖 14 時間序列(歲計)

7. 進行 MA1 模型的參數估計檢定及殘差檢定，依 P 值檢視該模型適當性。

- 模式配適：
  - 模式係數估計：

係數 coefficient	估計值 estimation	標準差 Std. err.	t-統計量 t-value	p-值 p-value
MA1	1	0.0879	11.3766	< 1e-04

圖 15 模式係數估計(歲計)

- 殘差分析：
  - 殘差資料獨立性(Independence)檢定：

虛無假設：資料互相獨立			
檢定方法 method	卡方統計量 Chi-square statistic	自由度 d.f.	p值 p-value
Box-Pierce檢定	2.276	1	0.1314
Box-Ljung檢定	2.4311	1	0.1189

圖 16 殘差資料獨立性檢定(歲計)

(七) SAS Enterprise Miner

SAS 完全以統計理論為基礎，功能強大，有完備的數據探索功能。基本內容支持 SAS 統計模塊，使之具有傑出的力量和影響，它還通過大量資料採礦演算法增強了那些模塊。SAS 使用它的 SEMMA 方法學以提供一個能支持包括關聯、聚類、決策樹、神經元網路和統計迴歸在內的廣闊範圍的模型資料採礦工具。

(八) SPSS

「統計產品與服務解決方案」軟體，用於多個領域和行業，是世界上應用最廣泛的專業統計軟體。

第五節、大數據分析應用案例現況與發展趨勢

大數據創新為政府和企業帶來新視界、新機會與新挑戰，更是城市管理、企業管理、國家治理的新變革。本研究收集美國、澳洲、英國與中國有關主計領域大數據分析之研究文獻，從中可以學習到各國在大數據的發展趨勢。

一、美國

美國政府在大數據應用上起跑得快，在大數據時代下遙遙領先，當我國政府仍在起跑點躊躇的同時，美國政府早在 2012 年就砸下 2 億美元，啟動了多項大數據國家級計畫來應戰，成為大數據發展的領頭羊。不僅如此，美國更在今年全面回顧過去大數據政策，尋找新機會和政策方向，準備好為大數據的下一個階段全力以赴。科學與技術顧問委員會提出一份大數據政策研究報告，報告中指出了大數據的 3 大機會點，以及 3 大疑慮，同時也做了一項大數據隱私調查，調查結果發現高達 8 成的受訪民眾非常在意政府如何使用和收集資料，且對於相關資料收集的機構並不信任。

(一)在開發工具技術方面，「計算軟體」及「資料視覺化管理」是下一階段工具技術主流。

1. 美國國防部每年投入 6,000 萬美元開發足以計算大量資料的軟體及工具。
2. XDATA 計畫也力推開源軟體，來提供使用者在不同應用環境下更彈性地處理大量資料。
3. 美國能源部邀請來自六個國家實驗室與七所大學的專家，共同開發新工具，用資料視覺化管理能源部內的超級電腦。

(二)大數據計畫擴展與延燒，從資料到知識

1. 藉由政策、資源和標準化的推動，廣泛使用與共享巨量且複雜的生物醫學數據。
2. 技術方面，開發並傳播新的分析方法與軟體。
3. 教育訓練方面，不僅加強資料科學家、電腦工程師及生物資訊學家的專業培訓。

(三)將大數據知識加以應用，從資料到知識到行動

推動大數據和分析技術與支援，提高經濟成長、創造就業、教育、健康、能源、可持續發展、公共安全、先進的製造、科學工程和全球發展，最後將大數據獲得的新知識見解，發揮作用，並培育區域創新。

(四)大數據朝向多面向且變動速度快的資料

1. 隨著網際網路應用、穿戴技術、先進的感應監測技術的不斷演進，現在的資料來源除了公眾網路、社群媒體、來自州政府的紀錄與數據、來自商業交易的數據、地理空間的數據等，還包括了新的資料收集來源，像是感應器、相機、地理間觀測技術。
2. 人們的生活已經處處皆是資料的收集管道，而這樣的資料量也將是前所未有的龐大，需要更高更複雜的分析技術與能力。

(五)參與式預算增加政府預算透明度

世界各地有上千個都市推行參與式預算制度，期望藉由公民參與預算決策，改善政府的「無感」問題，讓政府施政與公共建設支出計畫更能滿足民眾偏好，同時提高政府預算透明度。

美國第一個參與式預算，於 2009 年，誕生於芝加哥。芝加哥市有「議員補助款」（Aldermanic Menu Program）的預算制度。市長每年編列一筆錢給每位選區的議員作為地方建設之用，在 2010 到

2012 年度，這筆補助款的額度是 132 萬美元。補助款只能用於資本支出來改善基礎設施，除了用途的規定之外，這筆錢要怎麼用，完全由議員自行決定。紐約市和芝加哥一樣，也有「議員補助款」的預算制度；每位市議員，每年可以支配大約一百萬美元的地方建設預算。芝加哥和紐約市進行參與式預算的程序步驟非常相似，芝加哥 49 選區所舉辦的參與式預算的過程，也經歷三個階段。

1. 舉辦鄰里住民大會

實行的第一年（2009-2010），利用兩個月的時間舉辦九場住民大會。主辦單位向參與者介紹參與式預算的基本理念，以及補助款的性質與用途，請參與者提出想法來討論怎麼使用這筆款項。參與者在會中進行腦力激盪，討論他們社區的問題，對如何改善社區的基礎設施，分享他們的看法。

2. 規劃工作

住民大會所選出的 60 位社區代表，分成五個委員會：藝術與創新、公園與環境、交通與公共安全、街道，和運輸。在每個委員會中，社區代表的任務是要把住民大會提出的意見，轉化為具體的計畫方案，評估社區的需求，與專家和市政人員會面，了解計畫的可行性。

3. 公開展覽計畫及投票

向住民說明計畫內容，聽取回饋意見，修正計畫，最後擬定候選方案。經過長達四個月的工作，社區代表總共提出了 36 個預算候選方案，然後進入參與過程的第三階段：投票。議員根據投票結果將預算項目提交給市政府，市府依照一般預算程序通過後，計畫便可執行。

預算的編定與審議是政府施政非常重要的一環。透過市民參與，達

到除弊與興利的雙重目的。未來政府預算的編制過程，匡列一定比例或金額，開放公民團體或個人提案，經過專家評選及網路投票後，交給相關局處編列預算執行，發揮民間創意，推動符合民眾需要的計畫，減少政府不必要的浪費。

(六)未來趨勢

運用不斷進步的物聯網技術，促進產業與訊息化經濟的結合，加速經濟發展。

(七)大數據應用案例

1. 芝加哥大學和芝加哥城市數據計算中心（UCCD）計畫在路燈上安裝感測器，如圖 17 所示，以收集空氣品質、光線強弱、音量、熱度、人潮與風量等資料，進行大數據分析。

圖 17 燈罩感測器 （照片提供／UCCD）

(1) 目的

此專案的目的是更加了解芝加哥，但同時也注意到隱私權問題，因此感測器只會透過路人的行動裝置計算路過的人次，而不會蒐集行動裝置的數位地址。

(2) 分析方式

該裝備可以檢測到已經打開藍牙的移動裝置，以追蹤特定區

域的行人密度，另外計畫幫助對花粉過敏的人避開花粉密集的區域。

(3) 成本

每一個感測器的建置成本為 500 美元，而每年的電力成本則為 15 美元。另外，包括高通、英特爾、思科等企業已經承諾給予每年 100 萬美元以上的研發成本。

2. 美國太空總署地球觀測站利用過去 150 年間全球年平均死亡率與空氣污染數據分析模型，製作出全球空氣污染地圖，如圖 18 所示，顯示中國東部、印度北部及歐洲為空氣汙染嚴重地區。

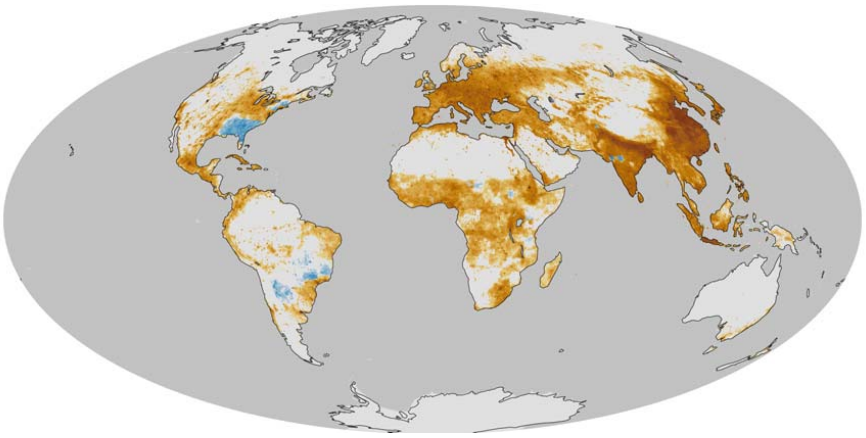


圖 18 全球空氣污染地圖

(1) 目的

利用過去 150 年間全球年平均死亡率與空氣污染數據分析模型，進而製作出全球空氣污染地圖，顯示不同地區間的空氣污染對人體健康造成的直接影響。

(2) 分析方法

NASA 地球觀測站則利用北卡羅來納大學的空氣污染數據分析模型，並結合 1850 年至 2000 年之間每千平方公里的年平均死亡人數，繪製出全球空氣污染地圖。此地圖能以視覺化

方式呈現不同地區間的空氣污染程度對人體健康的直接影響，像是相較於淺棕色區域，地圖上越深褐色區塊則表示有越多人因空氣污染而死亡，至於藍色區塊則代表該地區空氣品質已獲得明顯改善。

3. 紐約大學魯丁交通研究中心透過蒐集紐約 Citi Bike 公共自行車的停靠使用紀錄，完成視覺化的城市自行車動態路線圖，如圖 19 所示。

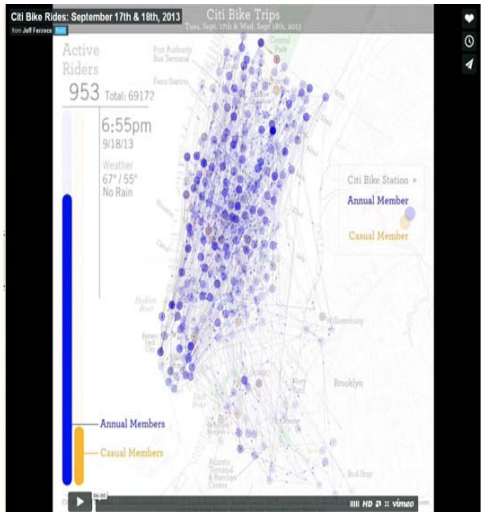


圖 19 視覺化的城市自行車動態路線圖

(1) 目的

紐約則是有人將這些公共自行車站點所紀錄下的時間，經過數據分析打造出一場視覺化的自行車動態秀，快速掌握城市的交通尖峰分布。

(2) 分析方法

透過蒐集連續兩天 48 小時來自 75,000 個紐約 Citi Bike 公共自行車的停靠使用紀錄，再經由數據分析所完成的視覺化的城市自行車動態路線圖。在這個動態自行車路線上，可以看到像是密密麻麻的自行車小點，以點對點的方式往返穿梭在各個 Citi Bike 站點，宛如像是一個散狀分布網路，而每隔一段

時間經過，自行車路線的車潮分布也會發生變化。透過這些往返路線的自行車動態圖，除了能協助政府了解民眾騎車路線的喜好外，也能快速掌握城市的交通尖峰時間分布，針對這些交通繁忙的時段提供更有效的解決方案。

4. Adobe 數位物價指數（Digital Price Index），2015 年 1 月至 2016 年期間，Adobe DPI 顯示 13.07%的累計通貨緊縮。然而，同期的 CPI 報告的佔 7.09%的累計通貨緊縮。仔細一看，DPI 能夠顯示價格跌幅由黑色星期五、網路星期一等節日相關的促銷訊息。如圖 20 中看出 DPI 上升，一些以前未被發現的趨勢。憑藉其動態數據時 Adobe 可以查明季節性的價格波動，以更好地識別宏觀和微觀的趨勢。

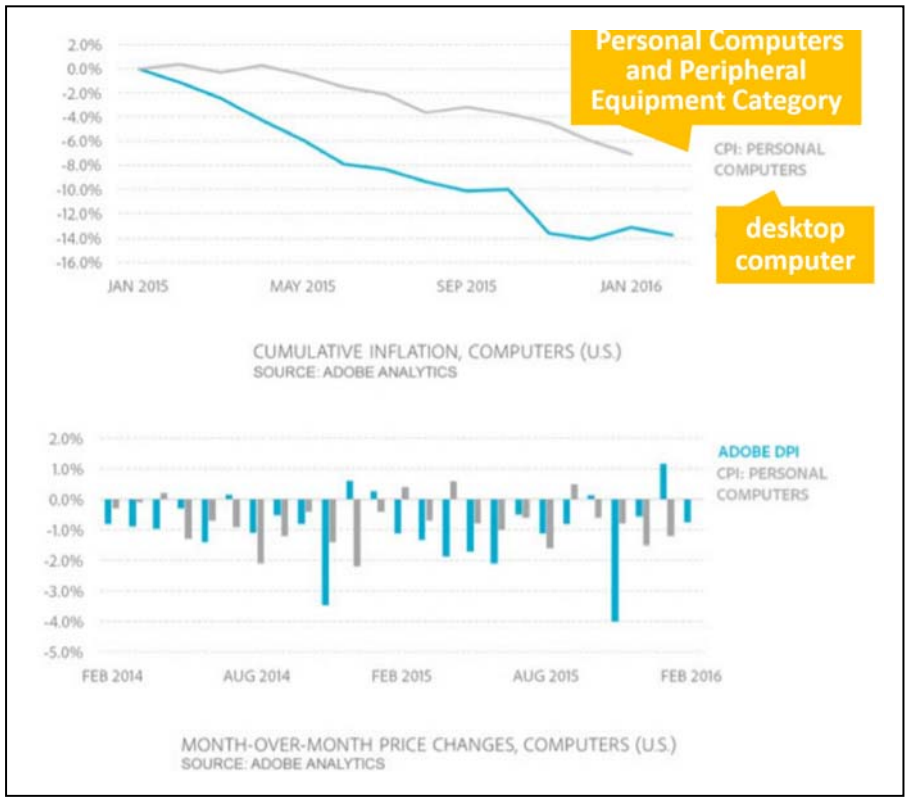


圖 20 Adobe 數位物價指數

(1) 目的

數位物價指數係為提供社會具數位經濟動態且能即時反映之資料分析觀點。

(2) 範圍

電子產品與食品類。

(3) 資料期間

2014/12-2016/12 共 2 年資料。

(4) 資料量

美國 80 億筆網路瀏覽、140 萬筆產品線上交易數據。

(5) 公式

採費氏指數理想公式計算。

二、澳洲

每年入境澳洲的旅客以千萬計，因此批准旅客入境與否，耗費澳洲政府大量資源。其實，在旅客到達澳洲國土，政府已經運用大數據技術，透過邊境風險識別系統審核每個旅客是否能夠入境，目前這套系統已在雪梨、墨爾本和布里斯班上線，為國家安全把關，也替政府省下大量人力及成本。大數據分析技術能找到資料之間意料外的關係，並為存在的問題提供重要的見解，或許能夠找到創新的解決方案，也能在資訊爆炸的年代，快速的對訊息及趨勢產生反應。

(一) 大數據在國家服務及政策發揮空間大

1. 澳洲政府至 2012 年為止，總共擴增了 93,000TB 的儲存容量，藉由分析這些日益增長的資料，澳洲政府能夠提供民眾更多元的服務，邊境風險識別系統便是一例。
2. 在社會福利、緊急事件處理，抑或是防詐騙等方面，大數據技術都能有很大的幫助。

(二)未來趨勢

由先驅計畫的技術需求，反饋回教育課程，強化大數據相關技術的教育訓練。

(三)大數據應用案例

1.澳洲醫院事先預測急診人數有效調度醫療資源

(1)目的

澳洲聯邦科學與工業研究組織（CSIRO）開發患者住院和預測工具來分析歷史資料，縮短急診病患等待時間。

(2)分析方法

澳洲聯邦科學與工業研究組織（CSIRO），開發了患者住院和預測工具（Patient Admission and Prediction Tool，PAPT）。PAPT藉由分析過去醫院急診資料，並考慮人口成長率和季節因素，來預測在接下來的一小時、一天、下個禮拜，特定節日，會有多少患者抵達急診室、這些急診患者的醫療需求，以及會有多少人住院或出院。這能幫助醫院更有效率地分配醫院資源，像是排班的員工人數、床位管理等。目前，這個工具已推廣至31家在澳洲昆士蘭的醫院。

三、英國

英國預計投入 10 億新臺幣在學校基礎教育，包含數據處理以及數學領域上，並在業界推廣先驅者師徒制計畫，促使大數據知識的流動。據估計，到了 2020 年，全球將有 500 億個裝置連上網際網路，智慧城市與物聯網時代來臨，即時分析大量的感知器資料，並適應性服務應用將越來越多。英國政府為滿足商業行為及公民需求，因此不光只有運算資料量的障礙，還有運算即時性的挑戰，由於感知器的資料可能是持續不停地串流傳送，所以可預期的是非結構性的資料比例將大幅成長，因此大數據技術將是目前有前景的解決方案。

大數據的價值在於挖掘資料中隱含的意義，這也許能找出消費者真正的需要，甚至是幫助政府制訂政策，英國政府引用電信比價網 Billmonitor 的分析資料，英國行動裝置用戶有 76%的用戶擁有長期合約，但是每月資料傳輸使用量遠不及應繳月費，導致一年平均浪費將近 200 歐元，所有用戶一年約浪費 50 億英鎊，當消費者更了解自己的使用習慣，將可以選擇更好的消費方案，而國家也能藉由這類的統計資訊，訂定相關消費者保護法，維護消費者權益。

(一)現況

在 2013 年英國政府為推動國家大數據技術，所提出的大數據發展策略報告書中，列出大數據 3 個發展方針，分別是加強處理資料的能力、強化基礎建設和化資料為力量。在 2014 年 9 月，英國政府要求 5 到 16 歲學生所學的計算機課程最終成果，需讓學生會建立 App 和程式開發，而課程會包含這些技能所應具備的細項能力，包含統計、機率、高等運算和建模。

1. 加強處理資料的能力

在現有職場的人力，英國則計畫以師徒制度提升大數據專業技能，推行先驅者計畫制定師徒制度的標準，由 8 家業界廠商率先執行此計畫，而英國政府也投入資金，贊助企業舉辦教育訓練。

2. 強化基礎架構

用 G-Cloud 政府雲計畫推動資訊與通信科技，所有公部門、地方當局的 IT 採購皆需將公有雲優先列入考慮，而因為 G-Cloud 計畫的推動，目前公部門形成跨政府的經濟規模，不只架構具有彈性規模且成本低廉。

3. 開放資料發揮最大的利益

在大數據政策上，資料就是主角，因此政府資料的開放以及保護

需要被重視。在 2010 年，英國政府開放超過 10,000 個資料庫。

(二) 大數據應用案例

1. 不同於一般政府開放的人口、教育等統計資料，倫敦政府利用 32 個地區的選舉資料結果，創造出視覺化的政治色彩地圖，如圖 21 所示。

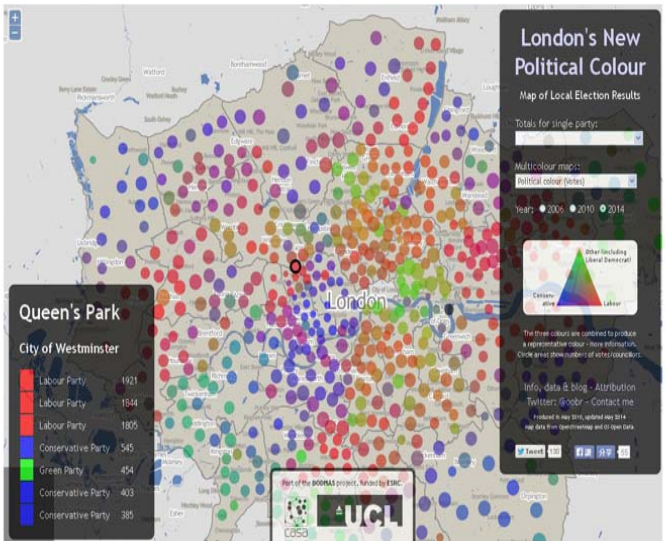


圖 21 倫敦社區政治地圖

(1) 目的

一般對於政府開放的統計類資料，例如人口、勞工、教育、文化及選舉等資料較無法直接拿來應用，但倫敦政府則是找到了新應用方法，將地方選舉資料和社群開放街圖（Open Street Map）結合，創造出視覺化的政治色彩地圖。

(2) 分析方法

由倫敦政府將 32 個地區選舉資料，與開放街圖結合。只要在地圖上縮放或滾動就能看到填滿繽紛色彩的圓圈，而這些顏色所代表是，各地區選民的綜合政黨傾向。此外，該地圖同時也能單獨顯示單一政黨、地區選舉的版圖分布。也讓民眾

更了解每一地區政黨勢力分布，例如：2014 年勞工黨的勢力版圖偏向倫敦中心和東部，而保守黨則是偏向在外圍和西部。

2. 紅綠燈裝感測器，依行人數調節車流。倫敦一天有 1000 萬輛汽車上路，其中包含 9000 輛公車。如何有效率使用道路？政府透過大數據與各式 App 應用，讓行車順暢、路人也省時。

(1)目的

英國最嚴重的城市問題是交通阻塞。即便已有地下鐵、地上鐵、公車、客運、火車和輕軌 DLR 等選擇，但是仍舊應付不了龐大交通量。尤其倫敦，一天有 1000 萬輛汽車在路上跑，其中包含 9000 輛公車。

(2)分析方法

英國逐步發展智慧交通號誌系統 SCOOT，透過道路上的感應器，了解車流量後，計算出最適合的交通號誌變化，讓汽車、公車和單車的道路使用更有效率。2012 年的倫敦奧運，就是利用 SCOOT，消化來自世界各地的千萬名選手和旅客。目前，倫敦境內的 6300 個交通號誌中，已有 3500 個裝設 SCOOT 系統，預計兩年內，將增加近千個。

最近，位於南倫敦的 Tooting Bec 地鐵站外，更引進了新型的智慧交通號誌系統，這款系統是在紅綠燈上裝設感測器和攝影機，可測試行人過馬路的狀況。如果偵測到行人很多，便延長綠燈時間；反之，若是沒有行人，就將號誌快速轉為紅燈。

四、中國

當前中國政府掌握著相對齊全的資料，卻一直存在著縱向和橫向

分割的問題，並沒有真正打通和共用。可以預見，假如政府內部的資料可以打通，並且開放公共存取權限，有所收穫的絕不僅僅是企業，還有媒體和輿論監督的力量。幾乎每一級政府和每一個部委都有自己的資料庫，公安部有一個覆蓋近 14 億人口的人口資料庫，國家工商總局有企業法人資料庫，金融、醫療、稅務、質監、社保、教育、氣象等都有各自的資訊庫，如何服務於公民，而不僅僅是政府或企業。

(一)DATA 國家數據

由中國國家統計局成立的「DATA 國家數據」是政府單位最為齊全的國家級統計數據平台，如圖 22 所示。

圖 22 DATA 國家數據網站

(二)百度數據研究中心

百度作為全球最大的中文搜尋引擎，具有強大的搜尋引擎技術和占絕對優勢的中國市場佔有率。百度數據研究中心成立於 2006 年，是百度公司下屬的網路資料研究機構，依託於百度平台，致力於開展線民搜索行為研究和網路輿情監測，為各行業客戶提供專業的網路資料產品和服務。

中國最大搜尋引擎「百度」將用自己的資料庫測量中國經濟；推出百度版消費者物價指數與就業指數。百度將從超過 6 億的使用者、3000 個購物地點、2000 個工業園區中蒐集資訊，創造量尺與指數。中國有 70～80%以上的網路調查都出自百度，因此資料庫已經商業化與科學化，百度也是中國人工智慧與大數據應用的領頭羊。有關百度數據研究中心平台，如圖 23 所示。

圖 23 百度數據研究中心平台

(三)DATA 數據堂

數據堂是中國首家專注於互聯網綜合資料交易和服務的公司，致力於融合和各類大數據資源，實現資料價值最大化，推動相關技術、應用和產業的創新。數據堂為中國大數據交易和服務領域的領頭羊，已經在全國中小企業股份轉讓系統(新三板)掛牌上市，成功登陸資本市場，資料定制、資料商城、移動應用資料服務是數據堂旗下三大核心業務。有關數據堂平台，如圖 24 所示。

圖 24 數據堂平台

第六節、小結

綜合本章節參考文獻，從大數據分析起源、國外主計領域大數據現況與趨勢、分析方法與技術工具的應用現況與趨勢及國外各國應用案例，本研究將大數據未來發展整理成九大趨勢。

趨勢一：數據資源化，將成為最有價值的資產

隨著大數據應用的發展，大數據價值得以充分的體現，在企業和社會層面成為重要的戰略資源，數據成為新的戰略制高點，是大家搶奪的新焦點。《華爾街日報》在一份題為《大數據，大影響》的報導宣傳，數據已經成為一種新的資產類別，就像貨幣或黃金一樣。Google、Facebook、亞馬遜、騰訊、百度、阿里巴巴和 360 等企業正在運用大數據力量獲得商業上更大的成功，並且金融和電信企業也在運用大數據來提升自己的競爭力，因此，大數據將持續成為機構和企業的資產，並作為提升機構和企業競爭力的有力武器。

趨勢二：大數據在更多的傳統行業的企業管理落地紮根

一種新的技術往往在少數行業應用取得了好的效果，對其他行業就有強烈的示範效應，目前大數據在大型互聯網企業已經得到較好的應用，其他行業的大數據尤其是電信和金融也逐漸在多種應用場景取得效果。因此，未來大數據將作為一種從資料中創造新價值的工具，會於許多行業的企業得到應用，帶來廣泛的社會價值。企業將更了解大數據，更能運用大數據分析並滿足客戶需求及潛在需求，而在業務營運智慧監控、精進企業營運、客戶生命週期管理、精緻化行銷、經營分析和戰略分析等方面將會獲得更多的應用。企業管理既是藝術亦是科學，大數據分析的科學方法運用於企業管理，將會帶來更顯著的進步，讓更多擁抱大數據的企業實現智慧企業管理。

趨勢三：大數據和傳統商業智慧融合，資料整合平台解決方案出現

來自傳統商業智慧化領域者將大數據當成一個新增的資料來源，而大數據從業者則認為傳統商業智慧化只是其領域中，處理少量資料時的一種方法。大數據使用者更希望能獲得一種整體的解決方案，即不僅要能收集、處理和分析企業內部的業務資料，還希望能導入互聯網上的網路瀏覽、Facebook 等社群及圖像、語音等非結構化資料。除此之外，還希望能結合移動設備的位置資訊，如此，企業就可以形成一個全面、完整的數據價值發展平臺。無論是大數據還是商業智慧，目的都是為分析的服務，數據全面整合起來，更有利於發現新的商業機會，此為大數據商業智慧化的展現。同時，由於行業的差異性，很難研發出一套適用於各行業的大數據商業智慧分析系統，相信更多的大數據商業智慧型制定化解決方案將在電信、金融、零售等行業出現。

以預測台北市儲蓄力及消費力的研究為例，透過臺北市 Open Data 與村里消費力資料整合，在 Open Data 中包括有台北市的地理資料、各村里的消費力、生活品質與超商數量等公開資料，如圖 25 及圖 26 所示，可以清楚知道台北市的各村里內超商數量、商圈、公車站等資訊。

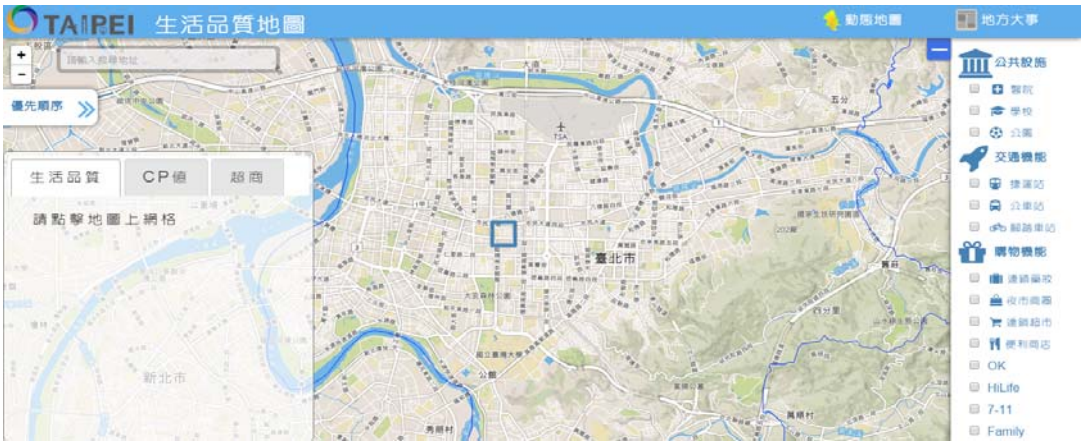


圖 25 台北市生活品質地圖

圖 26 台北市生活品質地圖網格示意圖

再結合地理資料，以資料採礦技術、模型建立及透過地理定位視角，觀察資訊彼此間的地緣關聯性，可以看出各村里土壤液化程度、犯罪率、噪音及空氣品質等資訊，如圖 27 及圖 28 所示。

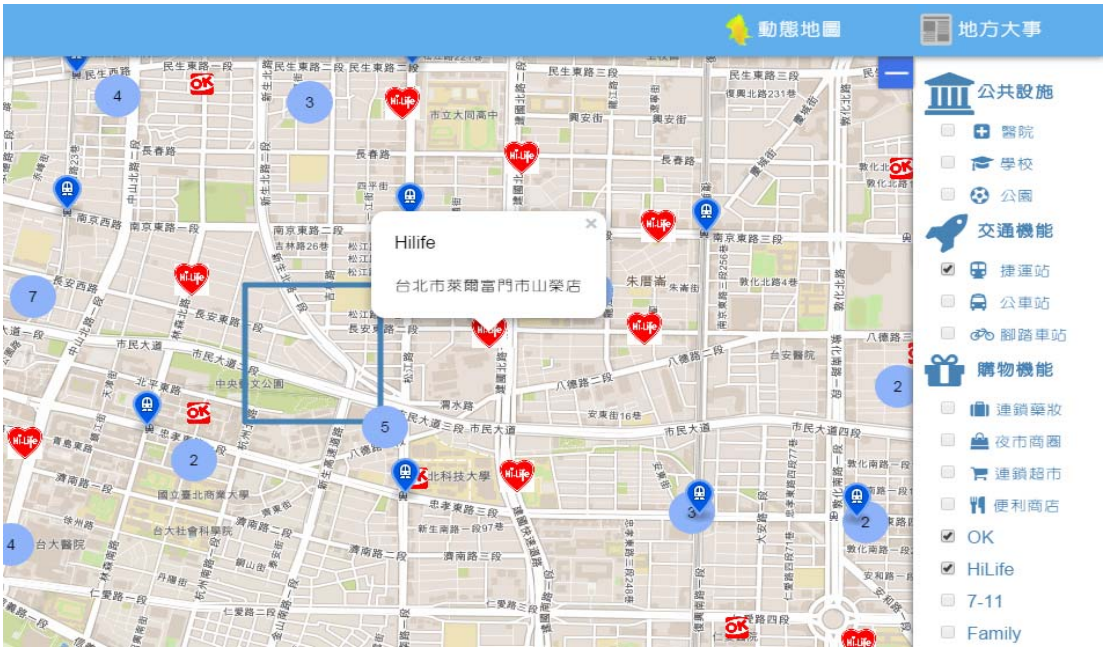


圖 27 地理資訊定位

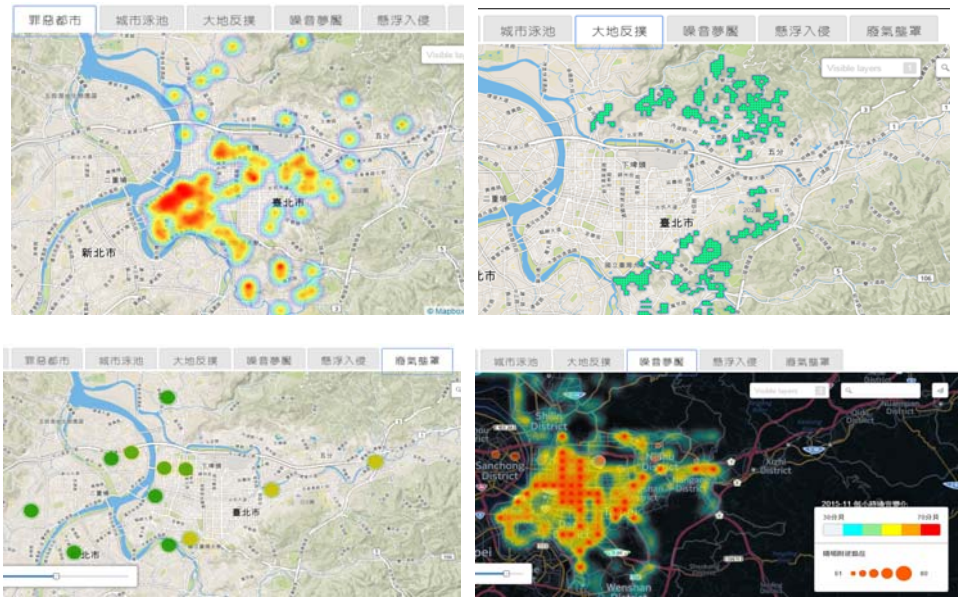


圖 28 焦點資訊聚焦

再透過村里定位與消費支出項目（食品/煙酒/服飾/生活開支）可以了解該區域的消費支出與收入儲蓄分布狀況，紅點愈密集的區域代表消費力愈高，綠點愈密集的區域代表居民的收入水準愈高，物業、服務業、零售業等業者可以透過這樣的模型分析作開店選址的參考，如圖 29 及圖 30 所示。

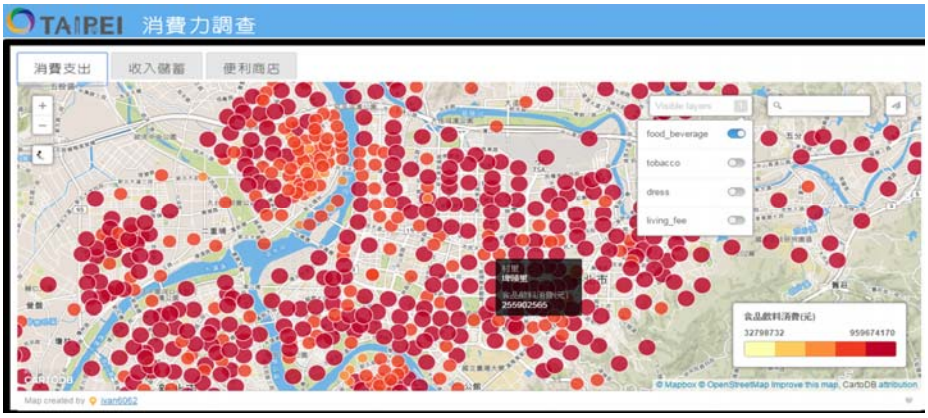


圖 29 地區消費支出項目分布

圖 30 地區儲蓄分布

趨勢四：數據將越來越開放，數據共用聯盟將出現

大數據越關聯越有價值，尤其是公共事業和互聯網企業的資料開放資料將越來越多。依目前現況發展，美國、英國、澳洲等國家的政府都在政府和公共事業上的數據做出努力，而國內的一些城市和部門也在逐漸開展數據開放的工作，對於不同的行業，數據愈共用也愈有價值。例如，每一醫院想獲得更多病情特徵庫以及藥效資訊，那麼就需要全國，甚至全世界的醫療資訊共用，從而可以通過平臺進行分析，獲取更大的價值。數據將會呈現一種共用的趨勢，不同領域的數據聯盟將出現。

趨勢五：大數據安全愈來愈受重視，大數據安全市場將愈發重要

隨著數據價值愈來愈重要，大數據安全穩定也將會逐漸被重視，網路和數位化生活也使得犯罪的分子更容易獲取他人的資訊，也有更多的騙術和犯罪手段出現，所以，在大數據時代，無論對於數據本身的保護，還是對於數據而演變一些資訊的安全議題，企業將對大數據分析有較高的要求，大數據安全是跟大數據業務相對應，與傳統資訊安全相較，大數據安全的最大區別是安全廠商，在思考安全問題的時候，首先要進行業務分析，並且找出針對大數據的業務威脅，然後提出有針對性的解決方案。例如，對於數據儲存量的情境，目前很多企業採用開源軟體如 Hadoop 技術來解決大數據問題，由於其開源性，因此，安全問題也更需注意的。市場需要更多專業的安全廠商針對不同的大數據安全問題來提供專業的服務。

趨勢六：大數據促進智慧城市發展，為智慧城市的引擎

隨著大數據的發展，大數據在智慧型城市將發揮著愈來愈重要的作用，由於人口聚集給城市帶來交通、醫療、建築等各方面壓力，需要城市能夠更合理的進行資源佈局和調配，而智慧型城市正是城市治理轉型的最優良解決方案，智慧型城市是通過物與物、物與人、人與人的互聯與互通能力、全面感知能力和資訊利用能力，通過物聯網、移動互聯網、雲端運算等新一代的資訊技術，實現城市高效率的政府管理、便捷的民生服務、可持續的產業發展。城市數位化到城市智慧化，關鍵是要實現對數位資訊的智慧處理，其核心是引入了大數據處理技術，大數據是智慧城市的核心智慧引擎：智慧型安防、智慧型交通、智慧型醫療、智慧型管區等，都是以大數據為基礎的智慧型城市應用領域。

趨勢七：大數據將催生一批新的工作職位和相應專業

一個新行業的出現，必將在工作職位方面有新需求，大數據的出現也將推出一批新的就業職缺，例如，大數據分析師、數據管理專家、大數據演算法工程師、數據產品經理等等。具有豐富經驗的數據分析人才將成為稀有的資源，數據驅動型工作將呈現爆炸式的增長。而由於有強烈的市場需求，大學也將逐步開設大數據相關的科系，以培養相應的專業人才。企業也將和大學緊密合作，協助大學聯合培養大數據人才。如 2014 年，IBM 全面推進與大學在大數據領域的合作，引入強大的研發團隊和業務夥伴，推動「大數據平台」和「大數據分析」面向，進行行業產學研究創新合作，以及系統化知識體系建設和高價值人才培養。

趨勢八：大數據多方位改善我們的生活

大數據不僅運用於企業和政府，也應用於我們的生活。在健康方面，我們可以利用智慧型手環監測，對我們的睡眠模式來進行追蹤，瞭解睡眠品質；我們可以利用智慧型血壓計、智慧型心率偵測儀遠端便可監控身在異地的家裡老人的健康情況，讓遠在他方的外出工作者更加放心在出門。在交通方面，我們可以利用智慧型導航出行 GPS 數據瞭解交通狀況，並根

據堵塞情況進行路線及時調整。在居家生活方面，大數據將成為智慧型家居核心，智慧型家電實現擬人化智慧，產品通過感測器和控制晶片來捕捉和處理資訊，再根據住宅空間環境和使用者需求自動設置控制，甚至提出優化生活品質的建議，如我們的冰箱可能會在每天一大早建議我們當天的食譜。

趨勢九：政府統計的應用

政府發布的統計依其來源至少有四類，公務統計、調查統計、普查統計及編算統計，每月發布的進出口、稅收數字屬公務統計，就業、物價、外銷訂單屬調查統計，工商普查，人口普查則為普查統計，至於經濟成長率、每人 GNP，則是依聯合國國民經濟會計制度（SNA）所編算的統計。隨著網路日趨普及，公務統計取得的資料愈來愈豐富。政府提供開放資料（Open Data）格式供民間下載加值分析，有助於這些公務統計取得過去難以獲致的資訊。

第三章、研究設計

第一節、研究方法與架構

本研究計畫係從大數據（Big Data）思維與角度出發，分析技術則以機器學習、資料採礦 CRISP-DM 及統計等演算法為主，運用資料映射（Data Mapping）技術將政府其他資料庫如：政府歲計會計資訊管理系統（GBA）、臺閩地區工商及服務業普查資料、全民健保資料檔資料、財政部營利事業資料、科技部產業創新調查資料等，以及外部資料如：全國達康資料、全國意向資料等，進行跨機關及跨領域資料串連，以利後續資料採礦分析之用。有關本計畫研究概念，如圖 31 所示。

本計畫研究重點項目包括：蒐整國外先進國家有關主計領域大數據分析之研究文獻（含企業及政府面向）、規劃與分析方法、技術工具及應用案例等發展趨勢。透過專家座談及文字探勘技術等方法，蒐整國內政府部門、民眾、產業界與學術界對於主計領域之數據資料需求。

一、根據發展趨勢及數據資料需求，研提有關主計領域大數據分析之關鍵核心議題。

二、選擇至少一個以上關鍵核心議題，整合運用歲計、會計、統計資料，結合大數據分析技術，完成以下三項研究案例，萃取資料特徵與樣態，建構適當資料模型，並提出具體可行性之應用架構、平台環境與作法建議。

（一）建立以欄位及模型方法連結二個以上來源資料庫，並產生至少一個應用目的資料庫之研究案例。

（二）從普抽查或統計調查原始資料，抽取部份資料，進行去識別化及函數映射等方法，建立仿真統計調查應用資料庫之研究案例。

（三）建立以資料採礦（Data Mining）及資料驅動（Data Driven）大數

據分析方法之研究案例。

三、針對本研究之期中、期末報告，邀集至少五位專家學者（其中二位需為業務領域專家)，辦理共二場專家學者審查會，以及辦理至少二場研究成果分享發表會，並提出對主計業務相關施政決策。

四、透過本研究在技術領域所累積的知識、各項技術與經驗，移轉給主計總處。

為完整進行，本研究計畫將以三階段方式進行，研究流程見圖 33。分述如下：

第一階段：資料準備期

包含蒐集國內外相關文獻，探討核心議題、可用資料庫檢視。接著，舉辦一場期中專家審查座談，藉由專家的審定與座談確認分析方向的正確行及可行性。將提出對主計業務相關施政決策、後續研究建議。

第二階段：資料建模期

針對第一階段的議題訂定、規劃與討論後，本階段主要即是建立資料模型的階段。故本階段係以資料採礦 CRISP-DM 的程序作為本研究執行的方法與步驟。

第三階段：資訊分享期

本最後階段，將先撰寫初版的期末報告，並舉辦一場期末專家審查座談，對於研究結果的討論，及對主計業務相關施政決策、後續研究建議。最後，將召開二場本研究成果分享會，並透過本研究在技術領域所累積的知識、各項技術與經驗，移轉給主計總處。

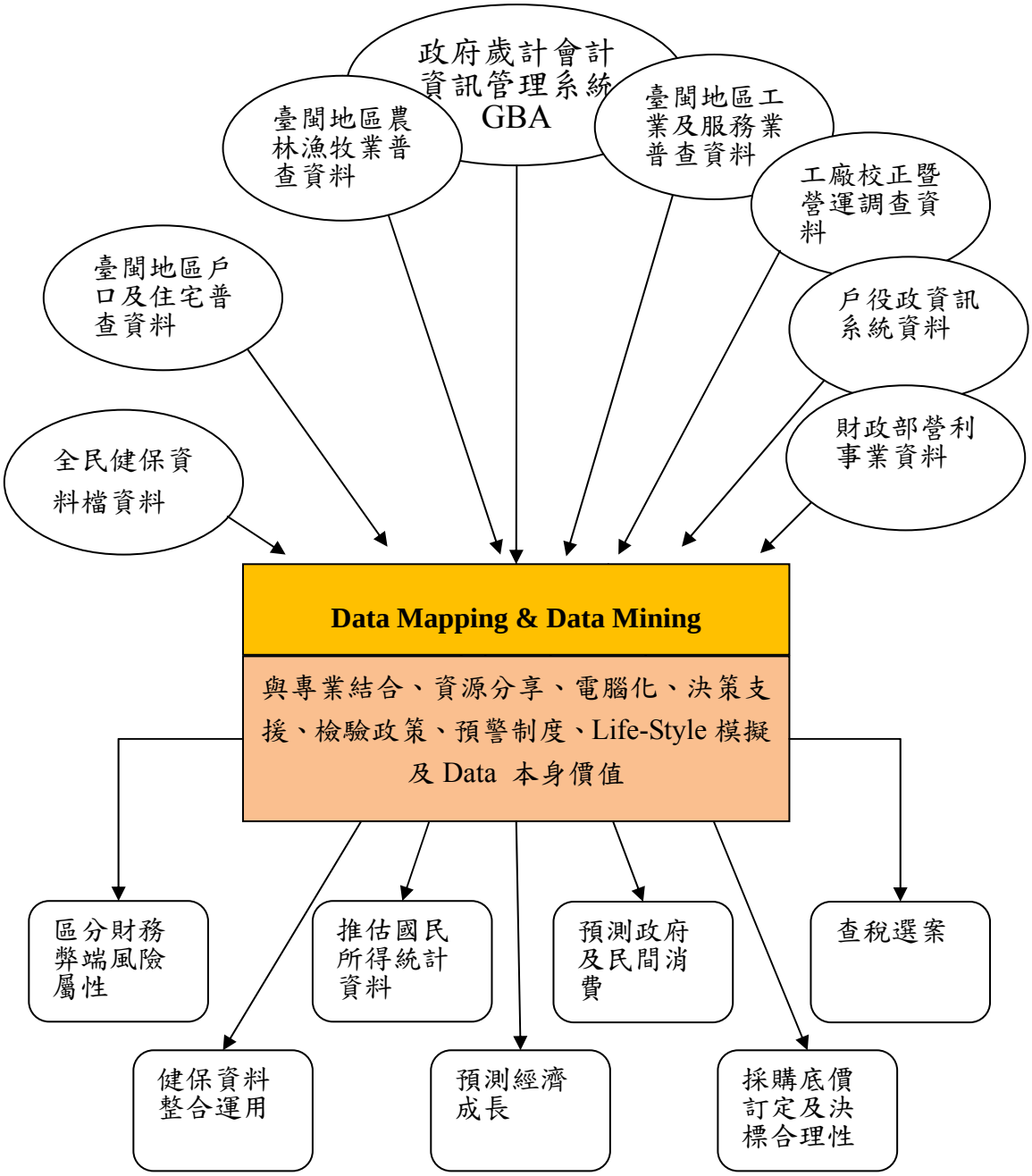

政府歲計會計資訊管理系統 GBA
臺閩地區農林漁牧業普查資料
臺閩地區工業及服務業普查資料
工廠校正暨營運調查資料
臺閩地區戶口及住宅普查資料
戶役政資訊系統資料
全民健保資料檔資料
財政部營利事業資料
Data Mapping & Data Mining
與專業結合、資源分享、電腦化、決策支援、檢驗政策、預警制度、Life-Style 模擬及 Data 本身價值
區分財務弊端風險屬性
推估國民所得統計資料
預測政府及民間消費
查稅選案
健保資料整合運用
預測經濟成長
採購底價訂定及決標合理性

圖 31 研究概念圖

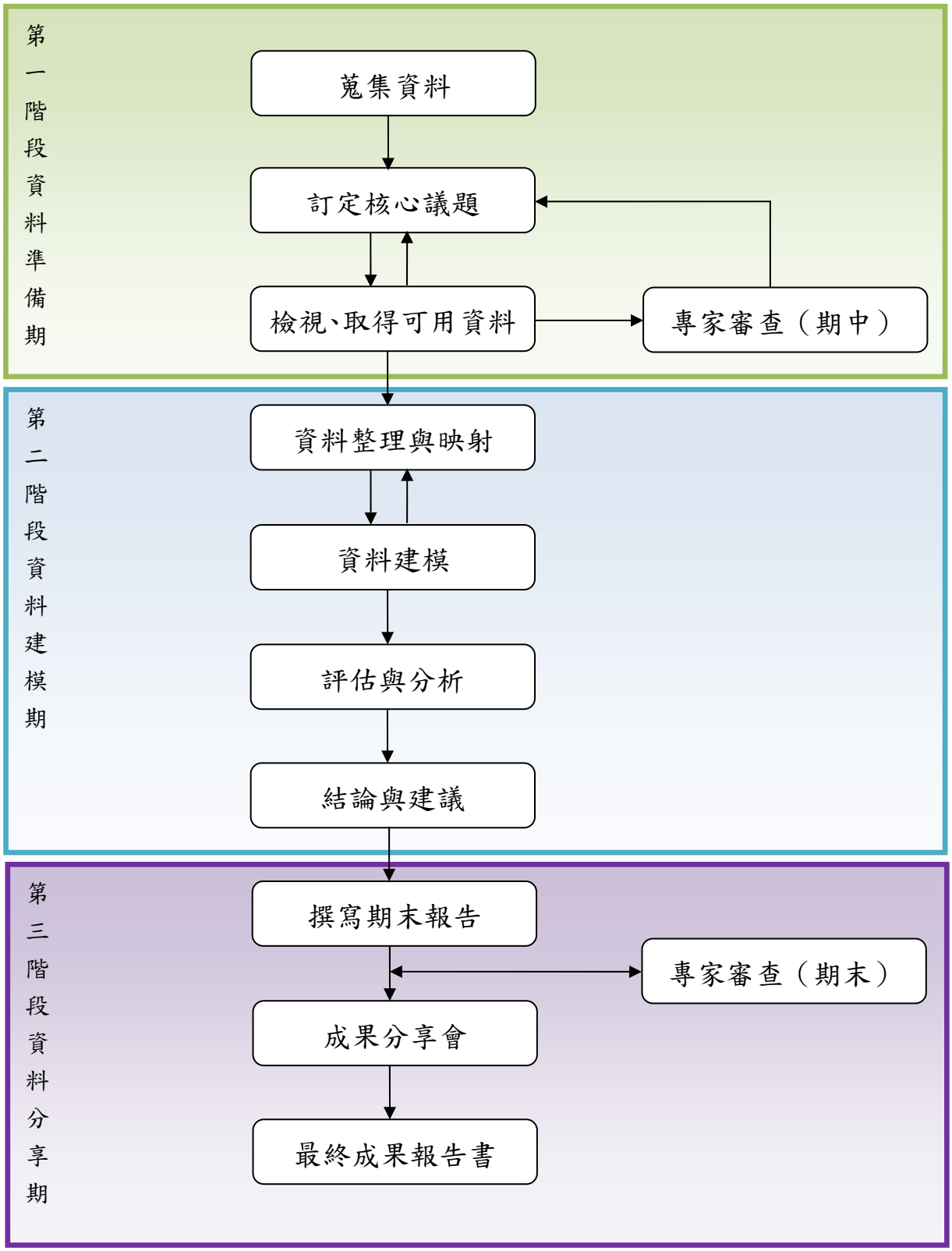

第一階段資料準備期
蒐集資料
訂定核心議題
檢視、取得可用資料
專家審查（期中）
第二階段資料建模期
資料整理與映射
資料建模
評估與分析
結論與建議
第三階段資料分享期
撰寫期末報告
專家審查（期末）
成果分享會
最終成果報告書

圖 32 研究流程圖

第二節、 需求蒐整

一、蒐整主計領域之需求

為能達到需求蒐整的完整性，本研究分別從四個面向蒐整需求，包括有國內政府部門的需求、民眾輿情的需求、產業界的需求以及學術界的需求，以能更瞭解國內各領域對於主計大數據發展的想法。

(一)國內政府部門需求

以主計總處來說，本研究分別透過與處本部、綜合規劃處、公務預算處、基金預算處、會計決算處、綜合統計處、國勢普查處、地方統計推展中心、內部審核小組及主計資訊處等十個單位專家需求訪談，完整蒐整主計領域之數據資料需求。各單位需求說明如下：

1.處本部

(1)強化本總處統計資料連結、整合及加值，將統計處國民所得資料、地方統計發展中心之家庭收支資料、國勢普查處之薪資、就業人口資料連結，加值並提供國家決策者進行決策參考。

(2)如何引用他機關或非本總處所擁有的數據，解決歷年調查瓶頸，發揮主計的效能。

(3)蒐整有關預算、會計、統計之國外大數據分析案例，建立一可行之主計三連環案例，將預算、統計、會計串聯。

2.綜合規劃處

(1)建立專案效益評估模型，以檢視每項專案是否有達到成效。

(2)主計三連環內控機制建立。

3.公務預算處

目前歲計部份估算不夠精細，希望未來 4 年中程額度推估會相對較準，尤其針對法律義務的支出分析約佔 7 成，需依法律義務支出估準。

4.基金預算處

(1)業權基金部分，本處訂定預算盈餘目標，促使各基金勉力達成，因此，希以大數據分析為基礎協助訂定具客觀性之預算盈餘目標。

(2)另為配合政府各項政策推動，政府現設非營業特種基金計 105 單位，因基金種類眾多，建議可先挑選某類基金，透過本研究案建立基金財務預警機制模型，以利後續推廣應用。

5.會計決算處

(1) 在計畫效能面加入追踪、管理效能等管考機制。

(2) 藉由大數據資料之收集，評價（evaluate）計畫的效能、價值性，建立計畫或預算模型，將計畫->預算->執行->考核->評核->KPI，建立推展 SOP。

(3) 在歲計會計資訊管理系統用脈絡關連找出關係 pattern，從既有報帳中間找出 pattern 或 fraud detect。

6.綜合統計處

物價指數 CPI 可否從交易檔、消費、電子發票抓資料，或從 channel 價格、比價網等無個資資料之網站，透過篩選查價時間點進行爬文，取代某些問項不用調查。

7.國勢普查處

跨資料的整合與加值，以公務檔案串聯(ID 串聯)來簡化調查問題，目前遇到挑戰是資料去識別化導致檔案串聯或整合不易。

8.地方統計推展中心

(1)地方政府公務統計報送及資料彙集制度建立。
(2)家庭收支調查，資料連結，以減少問項為目標。
(3)綠色國民所得，細分程度不夠細、強勢作為行政力，完整建構。
(4)推動公務統計管理資訊系統：地方基本資料建構，資料整編。

9.內部審核小組

(1)對於機關每日發生經費（記帳憑證金額與傳票文字說明），進行支付體系與內部審核關連分析，並回饋至管理面參考。
(2)透過資料採礦建立並回饋內審的 Knowhow。
(3)有關外界關切議題，如：節能減碳，運用文字探勘找出相關預決算書資料與金額，將議題與經費金額建立連結，並作為預算執行績效管控參考。
(4)強化會計與統計資料連結。

10.主計資訊處

(1)思考國家需要哪些資料、我們目前缺什麼資料，進行主計機構資料流整合評估（Dataflow 先建立）

(2)建置具智能能力之資料框架與機制、資訊環境與平台，統一資料格式便於收集，盤點本總處資料，並且建立資料屬性、定義，以供資料分析連結，例如以生產為例，哪些需要生產，從農業、工業資料大結合。

(3)需考量哪些數據可以取得、有限時間可以做完、規則及樣態（pattern）。

(4)分析做有趣、讓別人有感，並輔以視覺化動態。追本溯源，大方向、小實證、資料整合，哪些不足再做調查。

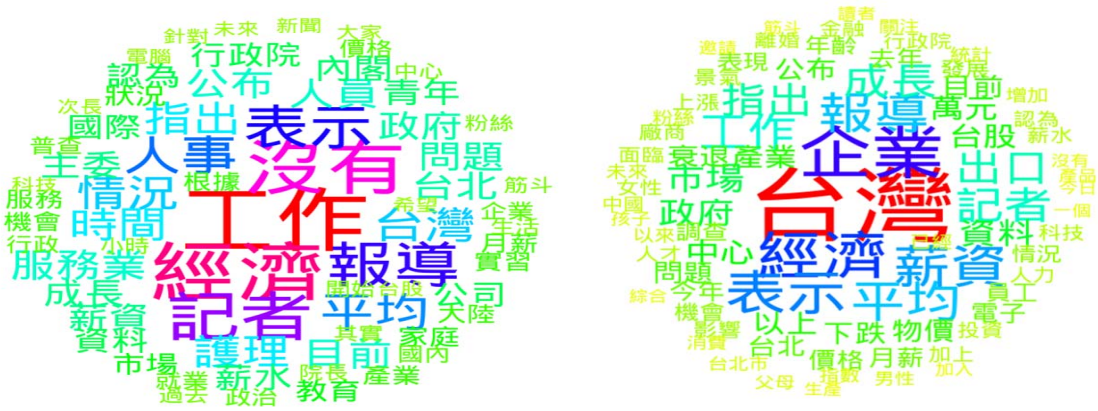
(5)可以使用文字探勘內部數據及外部數據，在預算計畫前後 100 天爬文進行輿情分析比較。

二、民眾輿情分析

本研究為能進一步了解民眾與媒體對於主計相關新聞的反應，針對 Ettoday 新聞中有關主計總處的新聞進行輿情分析，其分析結果如下說明。

（一）文字雲

針對主計總處去（民國 104）年年底全年經濟成長率（GDP）預測，將民國 104 年經濟成長率預測，由原本的 1.56%，下修到 1.06%，勉強保 1 的情況，以新聞分為去年、今（民國 105）年兩個部分，來看文字探勘是否有何差異。圖 33 左邊為今年年初到現在新聞，圖 33 右邊為去年一整年新聞。可以發現，圖 33 左邊出現了「工作」、「沒有」、「經濟」等文字，可見民眾與媒體最為關注的是在工作與經濟。

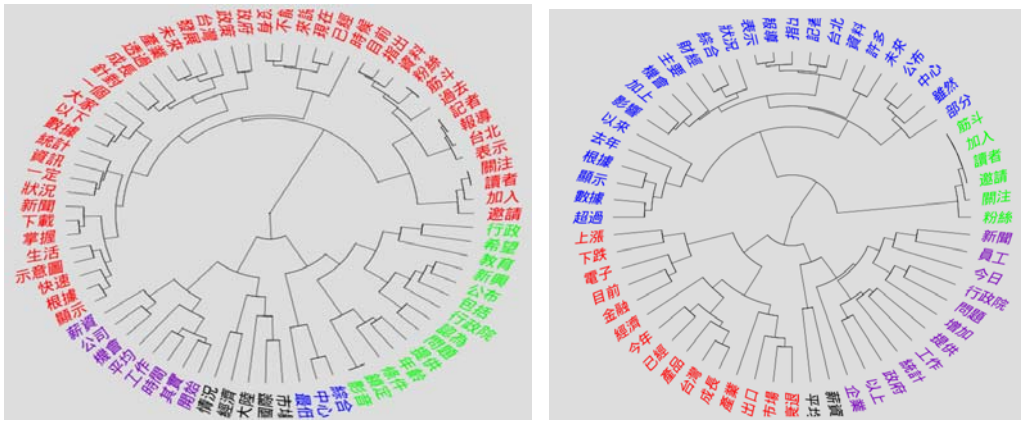


民國 105 年　　民國 104 年

圖 33 民眾輿情文字雲分析

(二)集群分析

圖 34 左邊為民國 105 年年初到 8 月新聞，圖 34 右邊為民國 104 年一整年新聞，可以發覺兩圖皆分為五群，其中圖 34 紅色字及綠色字部分為雜訊。其餘顏色為民眾討論關心的議題在生活、未來發展、成長等大幅成長。

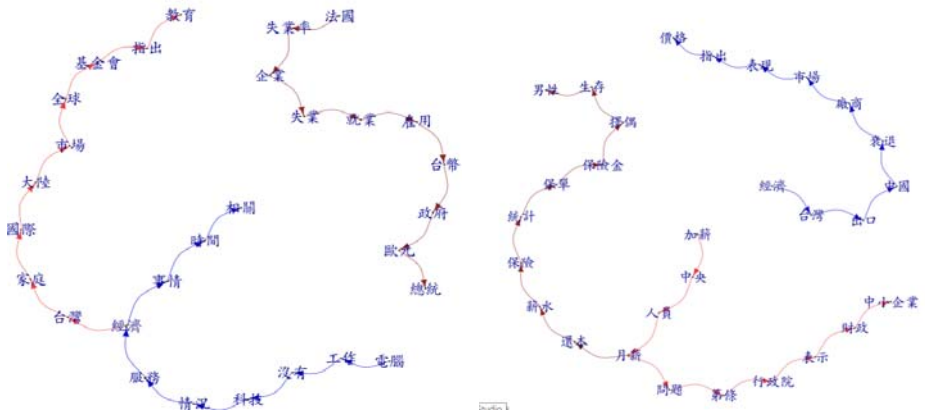


民國 105 年　　民國 104 年

圖 34 文字探勘集群分析

(三)脈絡分析

圖 35 左邊為民國 105 年年初到 8 月新聞，圖 35 右邊為民國 104 年一整年新聞。從脈胳分析可以看到失業→企業→政府的討論形成一條脈絡，教育→全球→市場→經濟→服務→沒有工作另成一條民眾關注的脈絡。

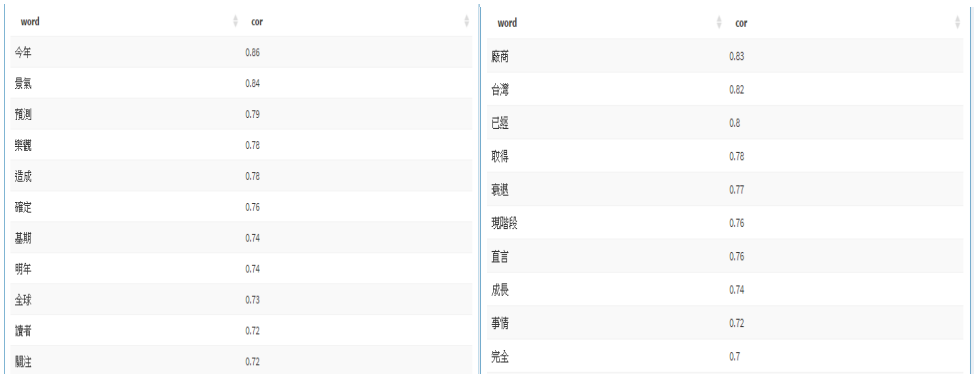


民國 105 年　　民國 104 年

圖 35 脈絡分析

(四)關聯分析

圖 36 左邊為民國 105 年年初到 8 月新聞，圖 36 右邊為民國 104 年一整年新聞。輸入關鍵字為經濟，圖 36 左邊景氣跟樂觀與經濟的相關性較高，圖 36 右邊為廠商及衰退較高。可見今年民眾對於經濟的討論更較去年熱絡。



word	cor
今年	0.86
景氣	0.84
預測	0.79
樂觀	0.78
造成	0.78
確定	0.76
基期	0.74
明年	0.74
全球	0.73
讓著	0.72
關注	0.72

word	cor
廠商	0.83
台灣	0.82
已經	0.8
取得	0.78
覺醒	0.77
現階段	0.76
直言	0.76
成長	0.74
事情	0.72
完全	0.7

民國 105 年　　民國 104 年

圖 36 關聯分析

綜合上述，從輿情分析可以看到，國內民眾對於主計大數據渴望能有更多與「經濟發展面」與「就業市場面」等兩面向的資訊提供。

三、產業界需求挖掘

國內會計審核技術發展在過去數十年經歷若干重大改變，從企業帳務以人工記帳，發展到企業以電腦記帳，採用會計套裝軟體的 2.0 時代，對傳統會計審核或將形成一股變革的浪潮。而現今國內外大型會計師事務所都看到 Data &Analysis（簡稱 D&A）會計審核技術發展趨勢，不約而同研發有助於數據分析的會計審核工具，試著採用最新數據分析的會計審核工具，試著將數據分析會計審核概念嵌入查核方法中，因此面對會計審核工作的新世代專業人才的訓練，也是未來政府可以取經的地方。

資誠所長張明輝在 2016 資誠玉山論壇提到根據《資誠台灣大數據調查分析報告》，將近 90%的企業認為大數據會影響未來企業營運，但只有 8%的受訪企業現在正運用大數據。也就是說，大家都知道大數據很重要，卻不知道如何有效運用大數據。可以預見，「數據分析能力」及「數據決策能力」將會是台灣企業未來面對的重要挑戰。隨著「互聯網+」、「工業 4.0」是我國當前和今後推動經濟轉型發展的兩個主動力，二者的主要結合點為

智能製造，而工業大數據是實現智能製造的驅動力之一。國內產業大數據應用中主要存在以下幾點問題：

(一)挖掘產業大數據價值的技術體系尚未建立

當前，我國還處於推進智能製造的探索階段，對多數企業而言，能夠自我感知、自我記憶的數據採集感應系統尚未建立，處理複雜數據結構的數據處理技術仍需提高，高效的資料庫維護和管理機制還需完善。

(二)行業內外數據整合應用不足

目前，我國大數據整體應用仍處於初級階段，行業內部數據和外部數據整合應用不足，跨行業的互動聚合效應尚未顯現，無論是工業或商業大數據亦是如此。

(三)企業各部門數據集成應用難度較大

企業內部數據的儲存應用是實現生產、業務協同的首要環節，但目前眾多企業內部部門之間訊息孤立情況比較嚴重，基本數據都是由系統採集和統計，不同部門之間的數據尚未打通和整合，致使數據利用率極低，為產業大數據的應用增加了門檻。

(四)產業大數據加工服務能力有待強化

由於客戶需求、生產環境等不同，不同行業、不同企業對數據的採集、處理過程和挖掘方向也各不相同，這就要求產業大數據加工服務企業需兼備產業行業專業知識與大數據處理能力。

目前，我國國內企業的數據向前預測能力有待提升及強化，多數只是將數據用於後向披露與原因分析。因此，產業未來透過主計數據的整合朝向加強產業大數據應用，加強明確產業大數據應用的技術、標準、產業，制定發展路徑，規劃並推動建立挖掘大數據價值的核心智能技術體系；健全推廣應用機制，實施一批具有特色的大數據應用示範項目，探索大數據產業的新模式、新業態。並構建有效人才培育引進和激勵機制，積極營造

有利於產業大數據人才培養和發展的外部環境，這也是可以向業界取經之處。

四、學術界的觀點

政府財務必須透過行政、主計、公庫及會計審核等組織，以達成相互配合及分立制衡之功能，而主計功能的發揮，更是促使國家有限資源合理分配及有效運用的重要關鍵。主計工作是從數據資料中來發掘有用的資訊，對政府來說這是機遇也是挑戰，資料量大並非大數據主要挑戰，大數據的主要挑戰是即時性資料變化快，品種多及非結構化。大數據要求重視資料的挖掘和利用，對各種複雜資訊和資料要進行可信性分析，政府在加強資訊公開的同時也要重視資料的安全和隱私保護，自主掌握大數據開發是對主計大數據創新能力的考驗。

就學術界的觀點分析看來，未來政府在主計工作面對大數據的發展，應用心思考的是：

（一）大數據發揮協同效應需要政府主計及跨界達成競爭與合作的平衡

基於從民間或政府其他單位提出了更多的數據資訊的要求，如果沒有對整體產業鏈的總體把握，單從自己掌握的獨立資料是無法瞭解產業鏈各個環節資料之間的關係，因此對民眾或產業做出的判斷和影響十分有限。大數據最具有想像力的發展方向是將不同的行業及屬性資料整合起來，提供全方位立體的資料繪圖。透過資源分享，也可以降低主計工作成本，減少不經濟支出，且有助組織學習與知識累積。

(二)大數據結論的解讀和應用

主計機關不只是製造資料庫的單位，如何由資料庫中找到寶藏，做為支援決策及檢驗政策的方針，為其應努力的方向。大數據可以從資料分析的層面上揭示各個變數之間可能的關聯，但是資料層面上的關聯如何能具體的應用到各行各業中?如何制定可執行方案應用大數據的結論?這些問題不但能夠解讀大數據，同時還需深諳行業發展各個要素之間的關聯。

(三)大數據下的政府統計

我國將政府統計資料應用在公共行政上，但各部會對應用資料的概念尚不普及。科技部兩年前推動大數據擴大運用計劃，由各部會以民眾的角色提出需求及運用計劃，最後一共產生16個計畫，包括「教育適性學習的應用」、「空氣汙染及健康關聯性大數據研究」等成果的報告，針對主計資料大數據的需求創新研究仍有待擴展。但是已有一實際應用案例產生，即是國發會建立的「政府物價資訊看板平台」，這一平台整合了財政部、行政院消保處及主計總處等單位的資訊，從這個網路平台可以查到賣場、便利商店近期的食品、飲料、日常用品售價。在資訊透明化之下，物價難以再哄抬，非理性的通膨預期可望大大降低，這是一個以民眾使用角度思考所產生的資料應用服務，可作為未來推展更多主計大數據應用之參考。

可以善加運用的政府資料非常多，除了電子發票提供零售價格、銷售品項資訊之外，政府勞保統計、出入境統計也有助於了解國內就業、海外就業的情況。此外，經常引起各界關注的貧富差距問題，財政部總會將五百多萬戶家庭的稅檔資訊彙整成十等分位、

二十等分位的所得分配統計，這份資料近年已成為外界研判台灣貧富差距的重要參考。

不論是電子發票、勞保統計、入出境資料、綜所稅檔皆屬公務統計，無需抽樣推估即可取得，非僅有助於我們了解台灣經濟社會，其對政府抽樣調查所取得的相關統計也有相輔相成之效。例如國發會本次建置的物價看板平台，可以驗證消費者物價指數（CPI）的變化，勞保投保人數可以佐證主計總處的失業調查，而綜所稅檔二十等分位資訊同樣可用以研判家庭收支調查統計。此外，政府還擁有公保、全民健保、高速公路交通量、海關出進口及不動產實價登錄等公務統計，這些數據若能善加運用，也有助於政府擬定策略以提升施政效能。

第三節、 關鍵核心議題綜整與研析

一、關鍵核心議題綜整

本研究透過第一節專家需求訪談結果，完整蒐整主計領域之數據資料需求及關鍵核心議題。本節將需求綜整進行分析歸類為七大需求，詳細說明如表 5（需求單位序依訪談時間序排列）。

表 5 關鍵核心議題綜整表

關鍵核心議題	需求單位	需求內容
人口與統計資料加值整合應用需求	處本部	強化本總處統計資料連結、整合及加值，例如將統計處國民所得資料、地方統計發展中心之家庭收支資料、國勢普查處之薪資、就業人口資料連結，加值並提供國家決策者進行決策參考。
	綜合統計處	物價指數 CPI：可否從交易檔、消費、電子發票抓資料，或從 channel 價格、比價網等無個資資料之網站，透過篩選查價時間點進行爬文，取代某些問項不用調查。但會遇到品項花色是否一致性的挑戰。
公務檔案串連跨資料整合	處本部	蒐整有關預算、會計、統計之國外大數據分析案例，建立一可行之主計三連環案例，將預算、統計、會計串聯。
	國勢普查處	跨資料的整合與加值，以公務檔案串聯(ID 串聯)來簡化調查問題，目前遇到挑戰是資料去識別化導致檔案串聯或整合不易。
	主計資訊處	需思考的是國家需要哪些資料、我們目前缺什麼資料，進行主計機構資料流整合評估(Dataflow 及資料收集流程建立)。

關鍵核心議題	需求單位	需求內容
公務檔案串連跨資料整合	地方統計推展中心	地方政府公務統計報送及資料彙集制度建立；家庭收支調查，資料連結，以減少問項為目標。
	內部審核研究小組	強化會計與統計資料連結
	綜合規劃處	建立主計三連環 SOP
歲計資料整合應用	公務預算處	目前歲計部份估算不夠精細，希望未來 4 年中程額度推估會相對較準，尤其針對法律義務的支出分析約佔 7 成，需依法律義務支出估準。
預算及計畫部分文字、數字應用	基金預算處	業權基金部分，本處訂定預算盈餘目標，促使各基金勉力達成，因此，希以大數據分析為基礎協助訂定具客觀性之預算盈餘目標。
	綜合規劃處	建立專案效益評估模型。
計畫效能追蹤及控管	會計決算處	在計畫效能面加入追踪、管理效能等管考機制。
	內部審核研究小組	對於機關每日發生經費（記帳憑證金額與傳票文字說明），進行支付體系與內部審核關連分析，並回饋至管理面參考。
客製化 Text Mining	主計資訊處	可以使用文字探勘內部數據及外部數據，在預算計畫前後 100 天爬文進行輿情分析比較。
	內部審核研究小組	有關外界關切議題，如：節能減碳，運用文字探勘找出相關預決算書資料與金額，將議題與經費金額建立連結，並作為預算執行績效管控參考。
	民眾輿情	聽取更多民眾對「經濟發展面」與「就業市場面」的需求。

關鍵核心議題	需求單位	需求內容
資料平台 跨域整合	處本部	如何引用他機關或非本總處所擁有的數據，解決歷年調查瓶頸，發揮主計的效能。
	主計資訊處	建置具智能能力之資料框架與機制、資訊環境與平台，統一資料格式便於收集，盤點本總處資料，並且建立資料屬性、定義，以供資料分析連結，例如以生產為例，哪些需要生產，從農業、工業資料大結合。
	基金預算處	為配合政府各項政策推動，政府現設非營業特種基金計 105 單位，因基金種類眾多，建議可先挑選某類基金，透過本研究案建立基金財務預警機制模型，以利後續推廣應用。
	產業界	需要產業大數據應用，規劃並推動建立挖掘大數據價值的核心智能技術體系。

資料來源：本計畫整理

二、關鍵核心議題研析

（一）議題研析

為達到可以整合運用歲計、會計、統計資料，結合大數據分析技術，建構適當資料模型之計畫目的，從七大議題中分別以「業務領域問題大小」、「解決方法技術實作難易程度」、「資料取得與成熟度」三方面來評析議題重要性，以挑選數議題作為本年度計畫進行實作以及於期中報告中展現。

評估標準說明：

1. 業務領域問題大小：問題及單位在 1~2 項者給予 1 分
   3~4 項給予 2 分
   4 項以上給予 3 分
2. 解決方法技術難度：解決方法技術難度困難，給予 1 分
   解決方法技術難度普通，給予 2 分
   解決方法技術難度容易，給予 3 分
3. 資料取得與成熟度：目前已有資料但使用程度尚不成熟，給予 1 分
   目前已有資料使用程度尚可，給予 2 分
   目前已有資料且使用程度高，給予 3 分

表 6 關鍵核心議題研析表

關鍵核心議題	業務領域問題大小	解決方法技術實作難易程度	資料取得與成熟度	總分
人口與統計資料加值整合應用需求	**1**	**2**	**3**	**6**
公務檔案串連跨資料整合	**3**	**2**	**3**	**8**
歲計資料整合應用	**1**	**2**	**3**	**6**
預算及計畫部分文字、數字應用	**1**	**2**	**1**	**4**
計畫效能追蹤及控管	**1**	**2**	**1**	**4**
客製化 Text Mining	**2**	**3**	**3**	**8**
資料平台跨域整合	**2**	**2**	**3**	**7**

從表 6 分析結果可得，前 4 個重要的評析議題分別是「公務檔案串連跨資料整合」、「歲計資料整合應用」、「資料平台跨域整合」及「客製化 Text Mining」等 4 項關鍵核心議題，本研究先就此 4 項核心議題作為本年度計畫進行實作以及於期中報告中展現為優先，其餘議題將視本研究團隊資源分配亦會安排進行。

(二)核心議題初步構想

1. 公務檔案串連跨資料整合及資料平台跨域之整合議題

為能整合各單位之間的公務檔案，需先進行檔案串連整合，甚至進行跨域資料整合。

(1)目的

希望藉由具有所需欄位的另一個相關資料庫，依其關聯推估出目標資料庫所需欄位之值，以增進後續分析之可行性與價值。

(2)概念

在資料平台跨域整合議題的具體作法，是透過資料映射方法，工商普查公司資料（約 127 萬筆）與技術創新公司資料（約 1 萬 3 千筆）之間，可依其員工人數、行業別、資本額、所在區域……等欄位建立其關聯性，進而將後者的某些欄位 mapping 到前者。圖 37 說明 資料映射概念。

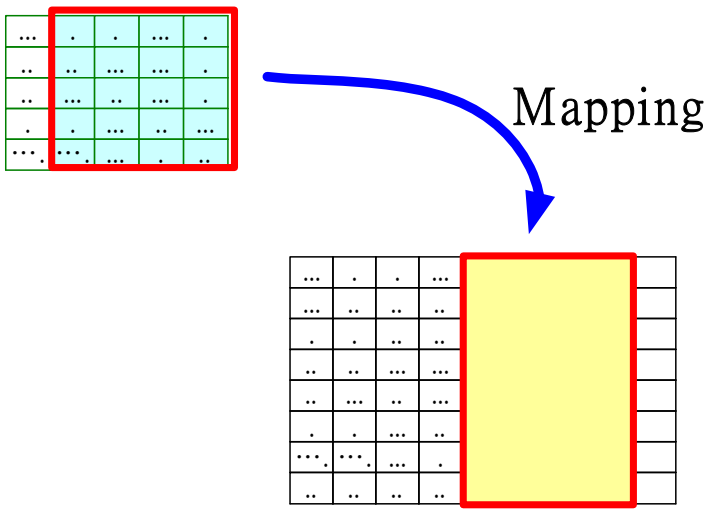


圖 37 資料映射概念圖

在公務檔案串連跨資料整合議題，會進行地方統計推展中心的「地方政府公務統計報送及資料彙集制度建立；家庭收支調查，資料連結，以減少問項為目標。」及國勢普查處的「以公務檔案串聯（ID 串聯）來簡化調查問題」為進行方向。

(3)主架構及實驗資料

實作資料庫將採用民國 100 年工商及服務業普查資料庫（Industry，Commerce and Service Census DataBase）約 127 萬筆資料量為研究中的目標資料庫，以民國 100 年第 3 次技術創新調查資料庫（Technology Innovation Survey DataBase）資料量約 1 萬 3 千筆為研究中的輔助資料庫。有關核心議題主架構及實驗資料如圖 38 所示。

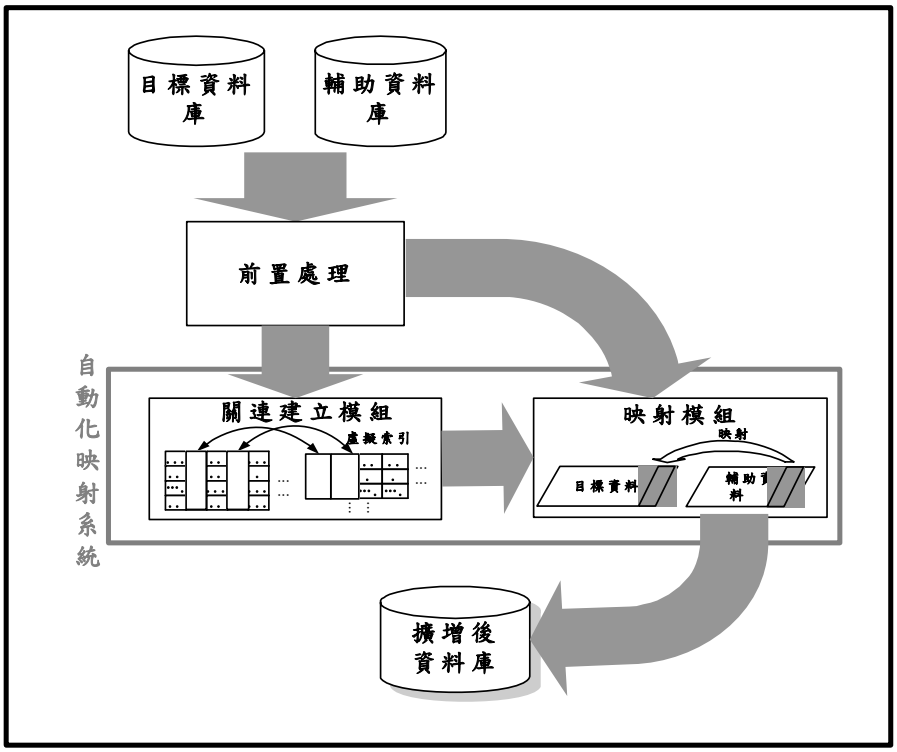


圖 38 核心議題主架構

2. 客製化 Text Mining

為主計總處開發一套客製化的 Text Mining 平台（如圖 39），讓同仁可以簡易的使用文字探勘與輿情分析，透過業務及政策需求即時修正符合之功能，隨時應變需求，針對施政方針進行分析。

圖 39 主計總處客製化 Text Mining 平台

其功能面包括有詞雲-生成詞雲、集群展示、脈絡分析、關聯分析、情感指數及 Facebook 粉絲團動態圖等，分析介面示意圖如圖 40-圖 45 所示。

圖 40 詞雲-生成詞雲

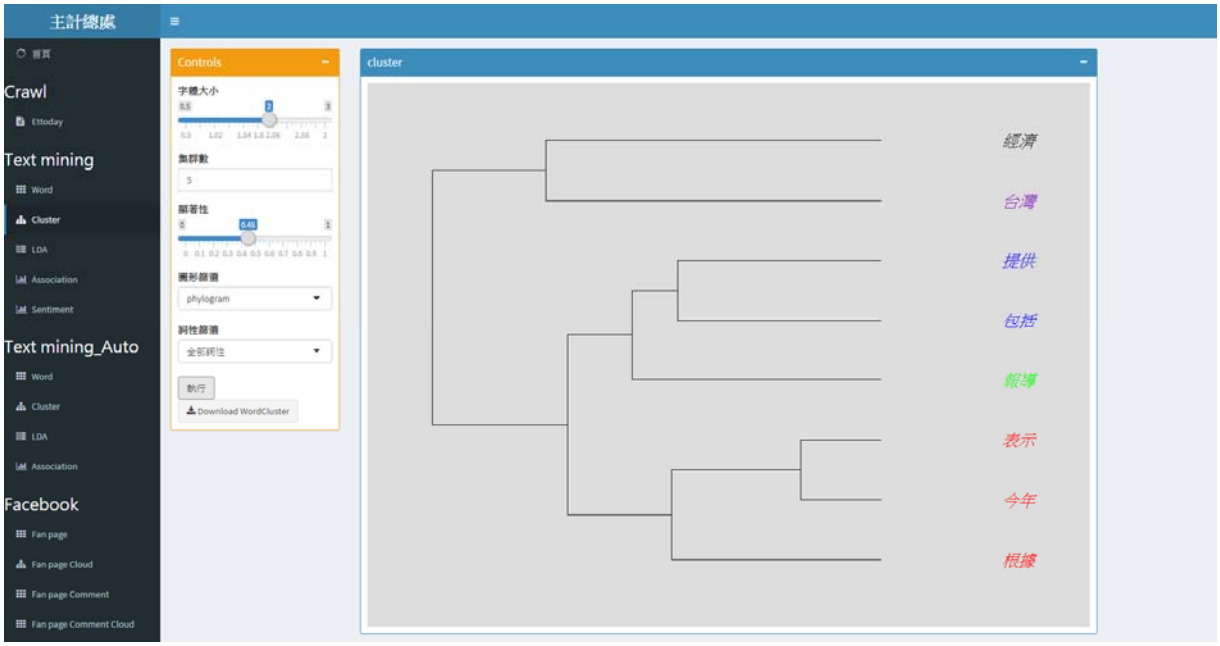


圖 41 集群展示

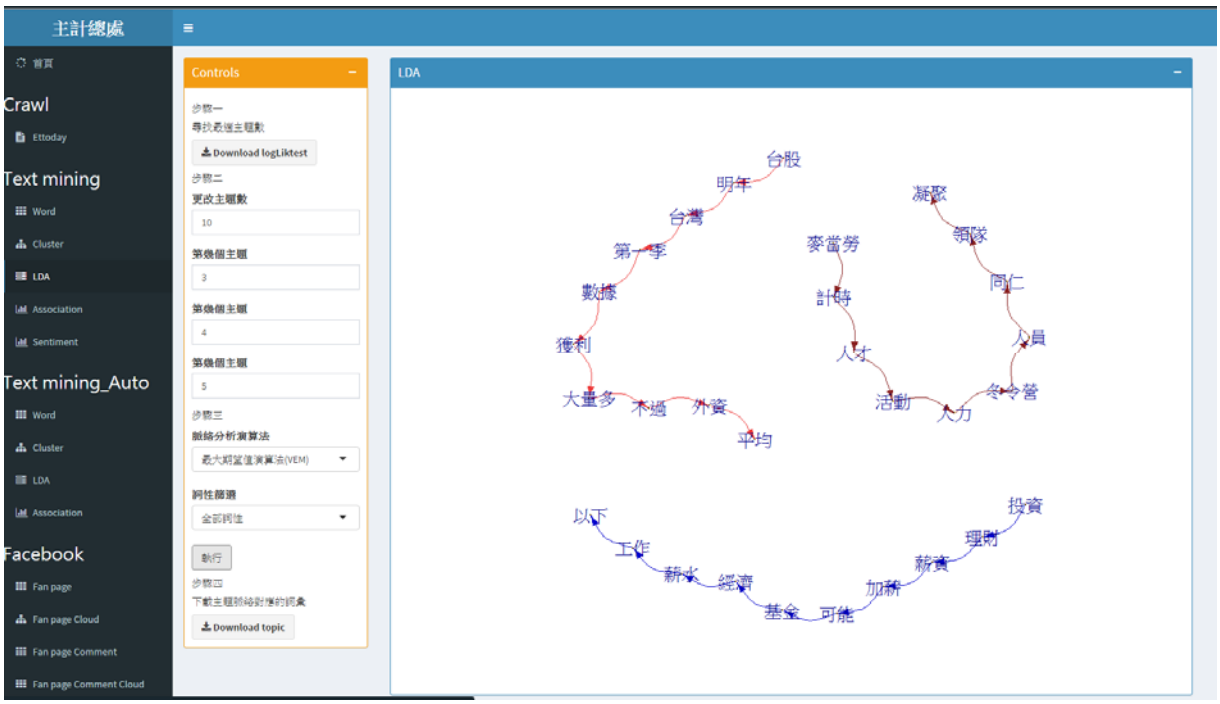


圖 42 脈絡分析

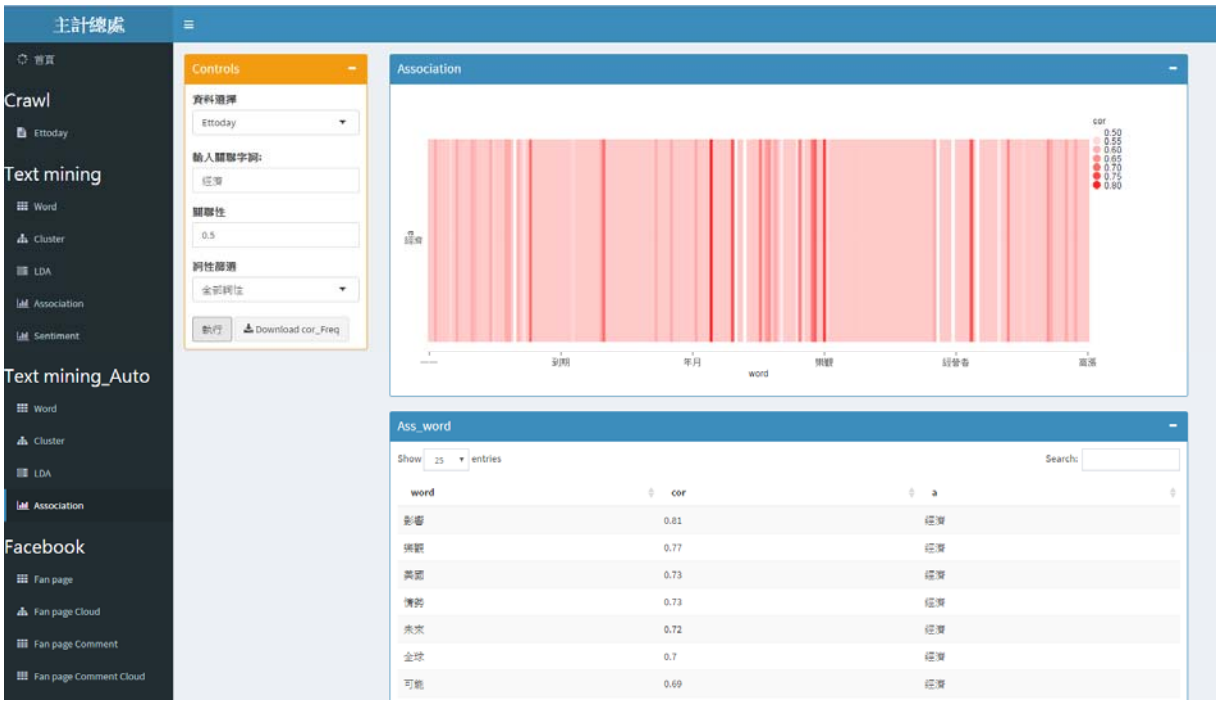


圖 43 關聯分析

圖 44 情感指數

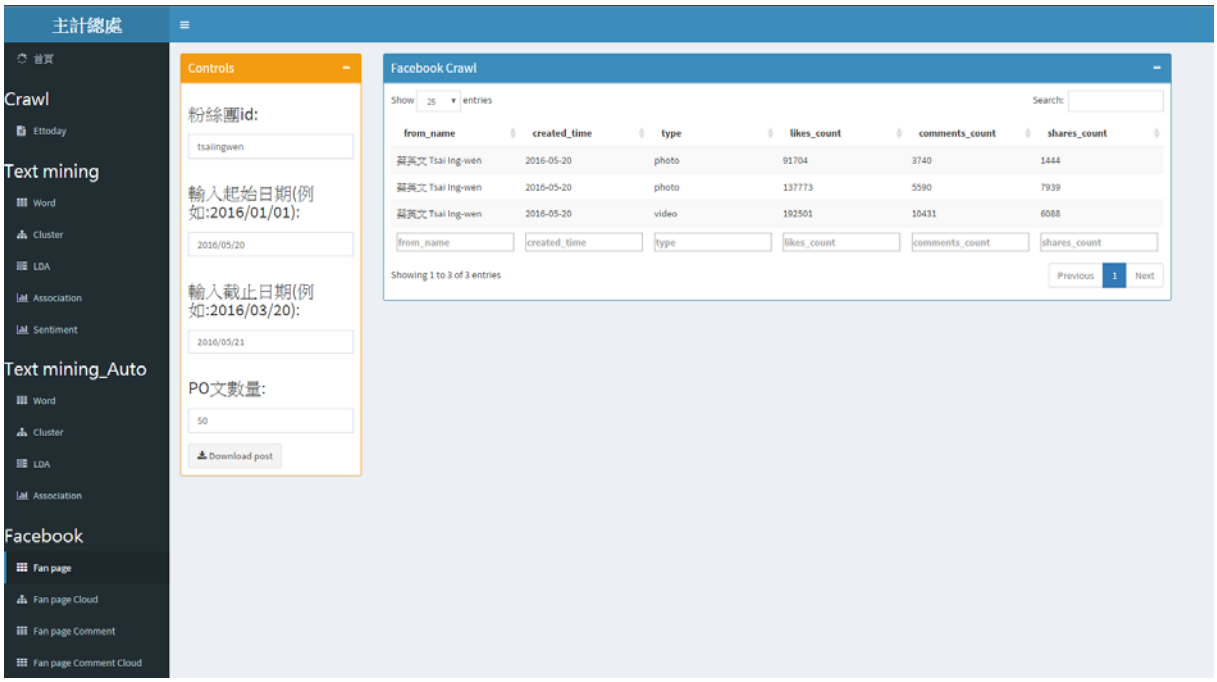


圖 45 Facebook 粉絲團爬文

第四章、村里常住人口推估模型

第一節、研究方法與資料來源

人口及住宅普查（以下簡稱人口普查）目的，為了解全國人口之質量、家庭結構、就學、就業及住宅使用狀況，供國家研訂施政計畫、規劃國家建設發展之主要參據，截至目前為止，我國分別於 1956、1966、1980、1990、2000、2010 年實行 6 次人口普查。

但隨著現今民眾逐漸重視隱私權保護議題，訪問調查日益困難，世界各國已開始運用公務登記資料，代替過去傳統普抽查或調查方式。因此，本計畫順應趨勢，並建置人口普查更細緻推估模型，採用創新大數據分析技術（如：隨機森林演算法、類神經網路等），應用函數映射方法，結合普抽查、統計調查去識別化資料、內政部、其他開放及跨私領域資料，建立仿真統計調查應用資料庫，作為未來主計總處普抽查調查方法參考。

更進一步說明，本計畫採用內政部民國 105 年 7 月人口資料、主計總處民國 99 年人口普查資料與全國達康民國 105 年 6 月常住人口資料[7]建立村里常住人口推估模型。傳統的常住人口方式調查方式乃透過數據網格化，藉以此將數據微型化，一般網格單位從村里至最小範圍網格皆可採用 GIS 軟體處理，但是技術上僅僅強調視覺上巨微型網格，並非專業的智能空間分析與地理智慧。全國達康採用精確的數據研究方法，在智能地理分析中，建構應用面上的「地理差異化」，例如人口數據分別在白天/夜晚/居住/工作的數量，因此得到常住人口數字。本次計畫所使用的民國 105 年 6 月常住人口資料是全國達康例年都會進行調查的資料，因此在使用上不需另外清整。

[7] 全國達康是台灣最早踏入電子地圖網際網路服務，並成功的將累積 20 餘年的 GIS 地理資料轉換為有價值的資料，包括有 200M*200M 網格社經數據(圖層/數據)、行政區界（圖層）、人口統計網格數據(圖層/數據)、全台知名商業地標資訊等，是擁有商業數據的民間單位。

透過與普查年資料銜接，以函數映射方法估計到一定期間各縣市村里的人口常住資料，掌握調查時點最新人口數量及分布，藉由具有所需欄位的另一個相關資料庫，依其關連推估出目標資料庫所需欄位之值，以增進後續分析之可行性與價值，本次以民國 99 年人口普查資料及民國 105 年全國達康常住人口資料，依其縣市村里欄位建立其關連性，進而將後者（全國達康常住人口資料）mapping 到人口普查資料。其研究方法說明如下：

一、建立 Generating Classification Model

以 assisted DB 中，與 target DB 共同(相似)的欄位建立 classification model。

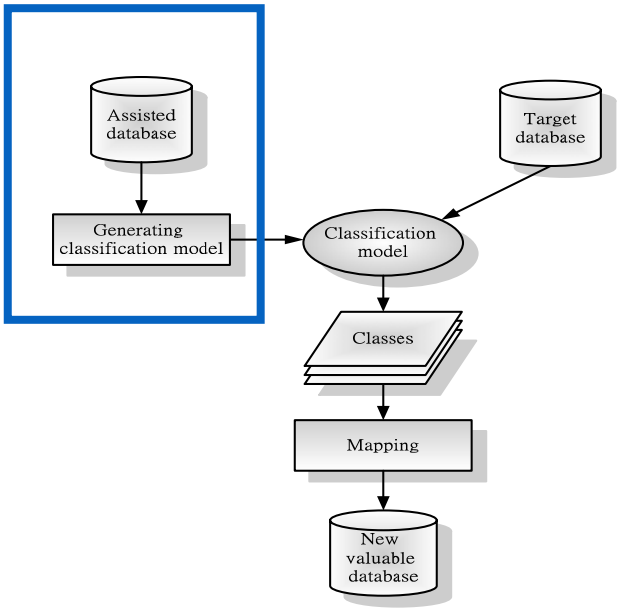


圖 46 Generating classification model 建立步驟

二、Classify Target Database

將 target DB 代入所建立的 classification model 中，使 target DB 拆成與 assisted DB 相對應的各個資料群。其目的是將 assisted DB 與 target DB 拆成相似的資料群， 以降低後序步驟（mapping）之誤差。

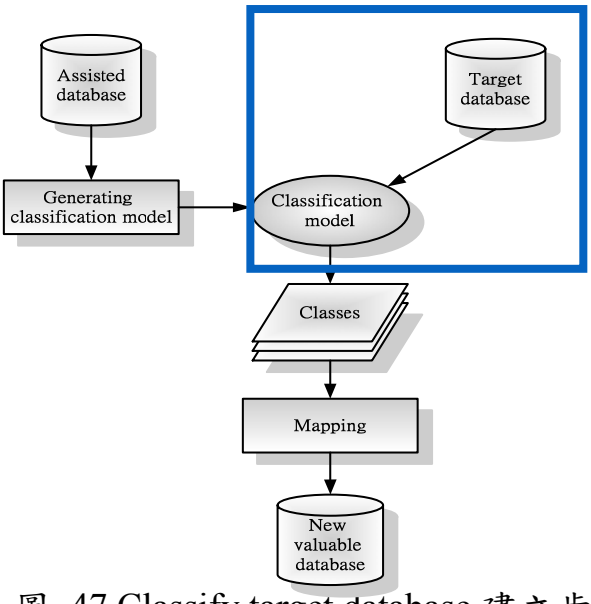


圖 47 Classify target database 建立步驟

三、建立 Mapping Index

運用迴歸方法找出具有解釋力之欄位，此欄位可以當為 mapping index。

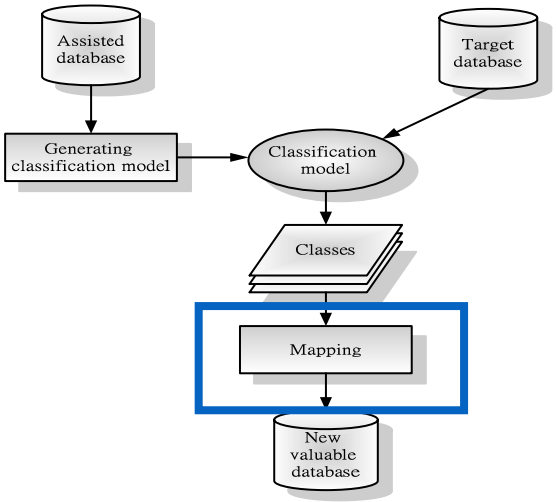


圖 48 Mapping index 建立步驟

$$\overline{Y'_{ij}} = \frac{\sum_{i=1}^{n} Y'_{ij}}{N} \qquad \text{For j=1,2,3,….k}$$

$$Y'_{ij} = a\overline{Y'}_{1j} + b\overline{Y'}_{2j} + c\overline{Y'}_{3j} \ldots\ldots n\overline{Y'}_{nj}$$

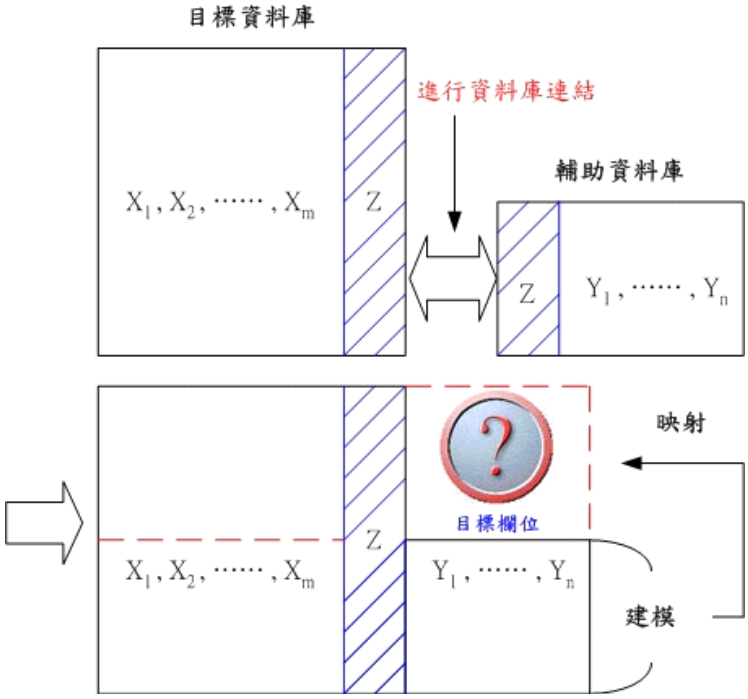


圖 49 函數映射方式

四、映射效果評估

每種控制變因組合下重複進行 10 次，以測試演算法採用簡單隨機抽樣，讓樣本分佈型態接近母體分佈，並且在穩定性。在本研究中，採用全國普查人口數作為 Target DB，將全國達康常住人口數作為 Assisted DB 進行 mapping，並採用線性迴歸方式推估。

第二節、建立模型歷程

一、人口資料匯整連結與整合

全國達康已成功的將累積20餘年的GIS(Geographic Information System 地理資訊系統）專業技術與經驗，以台灣電子地圖服務網為基礎，提供精確有效的生活資訊，例如人口統計（白天/夜晚/工作/居住）（圖層/數據）、人口統計網格數據（圖層/數據）、全台知名商業地標資訊等等，以減少民眾對地域陌生因而發生的成本。從圖 50 全國達康民國 105 年 6 月常住人口資料中可以應用的資料包括各縣市村里的公司數、工廠數、商店數、戶數等。

縣市別	區域別	ALLNAME	公司數	工廠數	商店數	VNAME_M	戶數	人口數	人口數-男	人口數-女
新北市	新北市板橋區	新北市板橋區留侯里	46	2	100	留侯里	691	1673	786	887
新北市	新北市板橋區	新北市板橋區流芳里	49	1	115	流芳里	644	1573	732	841
新北市	新北市板橋區	新北市板橋區赤松里	97	1	198	赤松里	326	847	414	433
新北市	新北市板橋區	新北市板橋區黃石里	37	1	130	黃石里	441	1182	578	604
新北市	新北市板橋區	新北市板橋區挹秀里	235	2	167	挹秀里	736	1800	881	919
新北市	新北市板橋區	新北市板橋區湳興里	152	4	365	湳興里	2319	5683	2741	2942
新北市	新北市板橋區	新北市板橋區新興里	98	0	403	新興里	1478	3278	1579	1699
新北市	新北市板橋區	新北市板橋區社後里	195	1	340	社後里	4801	9209	5118	4091
新北市	新北市板橋區	新北市板橋區香社里	104	0	163	香社里	1576	4508	2249	2259
新北市	新北市板橋區	新北市板橋區中正里	167	7	206	中正里	1613	4566	2232	2334
新北市	新北市板橋區	新北市板橋區自強里	110	0	89	自強里	2051	5426	2649	2777
新北市	新北市板橋區	新北市板橋區自立里	88	1	65	自立里	1092	2948	1427	1521
新北市	新北市板橋區	新北市板橋區光華里	114	0	125	光華里	1097	2816	1390	1426
新北市	新北市板橋區	新北市板橋區國光里	57	0	62	國光里	1565	3883	1903	1980
新北市	新北市板橋區	新北市板橋區港尾里	58	0	60	港尾里	1105	3055	1440	1615
新北市	新北市板橋區	新北市板橋區金華里	91	0	182	金華里	1462	4174	2050	2124
新北市	新北市板橋區	新北市板橋區港德里	107	0	204	港德里	1938	5115	2532	2583
新北市	新北市板橋區	新北市板橋區民權里	58	1	64	民權里	1069	2762	1362	1400
新北市	新北市板橋區	新北市板橋區建國里	135	2	146	建國里	1418	3736	1835	1901
新北市	新北市板橋區	新北市板橋區漢生里	191	3	225	漢生里	1680	4509	2178	2331
新北市	新北市板橋區	新北市板橋區公?里	144	5	206	公館里	1468	4126	2022	2104

圖 50 全國達康常住人口資料

首先我們利用主計總處民國 99 年人口普查資料與全國達康的民國 105 年 6 月常住人口資料，得到民國 99 年與民國 105 年普查與常住人口之比較如圖 51。

地區	性別	人口數_99普查	人口數_達康10506	比值
台北市士林區	M	150047	140414	1.068604
台北市大同區	M	55838	63787	0.875382
台北市大安區	M	141946	146296	0.970266
台北市中山區	M	101980	107873	0.945371
台北市中正區	M	73258	77370	0.946853
台北市內湖區	M	132351	137751	0.960799
台北市文山區	M	133511	132445	1.008049
台北市北投區	M	127319	124422	1.023284
台北市松山區	M	87330	98621	0.885511
台北市信義區	M	107404	109328	0.982402
台北市南港區	M	59353	59790	0.992691
台北市萬華區	M	90725	95542	0.949582

圖 51 民國 99 年人口普查與民國 105 年全國達康常住人口之比較表

此圖 51 中的變數有地區、性別、人口數_99 普查、人口數_達康 10506 以及比值，其中地區變數為行政區，可以細到鄉鎮區，而比值變數則是以人口數_99 普查資料除上人口數_達康 10506 資料所得到，比值大於 1 表示此地區在民國 99 年的普查人口數大於民國 105 年 6 月時全國達康的常住人口數，小於 1 則反之。圖 51 主要是顯示民國 99 年人口普查與民國 105 年全國達康常住人口間之差異的比值，該比值是用來進行映射資料時調整常住人口之加權指數，在進行後續推估時讓降低普查與全國達康常住人口之間的差異。

二、函數映射建立村里常住人口仿真資料庫

利用全國達康的民國 105 年 6 月的常住人口資料算出每個里在其所在區中所佔的比例，再用此比例去調整內政部的民國 105 年 7 月的鄉鎮區人口資料，即可以推出民國 105 年 7 月各區的村里常住人口資料，如圖 52。

縣市別	區域別	名稱	公司數	工廠數	商店數	戶數	人口數	人口數-男	人口數-女
新北市	板橋區	留侯里	46	2	100	691	1672	786	887
新北市	板橋區	流芳里	49	1	115	644	1572	732	841
新北市	板橋區	赤松里	97	1	198	326	847	414	433
新北市	板橋區	黃石里	37	1	130	441	1181	578	604
新北市	板橋區	挹秀里	235	2	167	736	1799	881	919
新北市	板橋區	漳興里	152	4	365	2319	5680	2740	2940
新北市	板橋區	新興里	98	0	403	1478	3276	1578	1698
新北市	板橋區	社後里	195	1	340	4801	9204	5115	4089
新北市	板橋區	香社里	104	0	163	1576	4506	2248	2258
新北市	板橋區	中正里	167	7	206	1613	4564	2231	2333
新北市	板橋區	自強里	110	0	89	2051	5423	2648	2775
新北市	板橋區	自立里	88	1	65	1092	2946	1426	1520
新北市	板橋區	光華里	114	0	125	1097	2814	1389	1425
新北市	板橋區	國光里	57	0	62	1565	3881	1902	1979
新北市	板橋區	港尾里	58	0	60	1105	3053	1439	1614
新北市	板橋區	金華里	91	0	182	1462	4172	2049	2123
新北市	板橋區	港德里	107	0	204	1938	5112	2531	2582
新北市	板橋區	民權里	58	1	64	1069	2761	1361	1399
新北市	板橋區	建國里	135	2	146	1418	3734	1834	1900
新北市	板橋區	漢生里	191	3	225	1680	4507	2177	2330

圖 52　民國 105 年 7 月各區的村里常住人口推估資料表

圖 52 的變數有縣市別、區域別、名稱、公司數、工廠數、商店數、戶數、人口數、並且再從人口數中細分出人口數-男及人口數-女兩項變數。行政區的部分細到村里，除此之外還有戶數、公司數、工廠數、商店數等變數可供未來分析之使用。

第三節、視覺化呈現發展

本計畫採用 Microsoft Power BI 的互動式資料視覺效果 BI 工具，以全新的方式檢視本計畫資料。Power BI for Office 365 的元件包括 Power Query、Power Map、Power Pivot 與 Power View。其中，Power Query 允許使用者於 Excel 中搜尋及存取公開資料或企業中的資料。 Power Map 則是一個 3D 的視覺化資料工具，能夠繪製地圖或探索地域及臨時資料，也能與這些資料進行互動；Power View 能在 Excel 中建立互動式的圖表；運用名為 Q&A 的自然語言搜尋引擎，使用者可進行自然語言查詢。微軟表示，企業用戶並可建立 BI Sites 共享該工具產生的分析結果與報表，而藉由

HTML5 的支援，也能利用行動裝置存取資料。

利用 Power BI 中的 Power Pivot 樞紐分析圖做資料視覺化的呈現如圖 53。此為縣市下各鄉鎮區的人口數直條圖，可以從左上的年度按鈕做年度篩選，右方的縣市按鈕則可以做縣市的篩選，以此圖為例，即為新北市民國 105 年 7 月各區人口數。

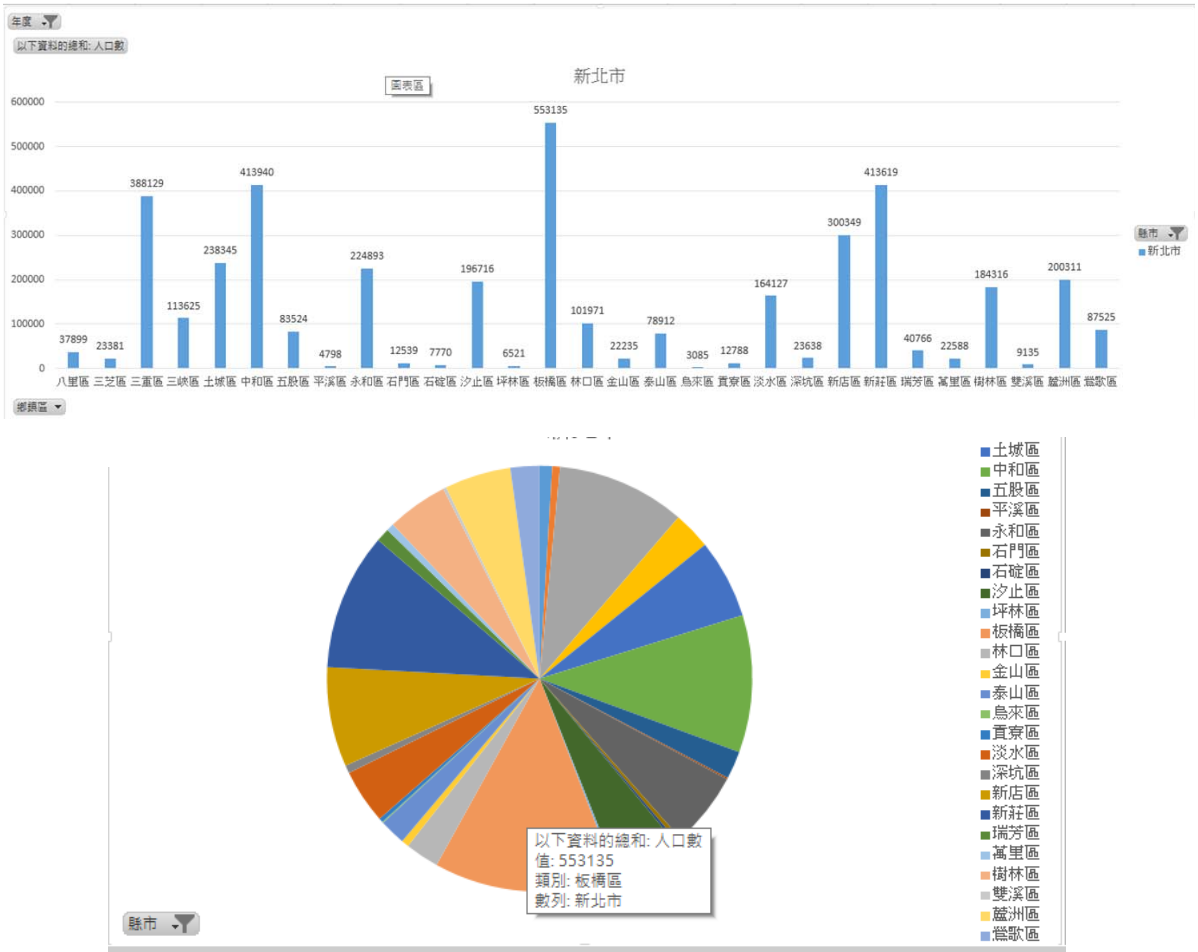


圖 54　民國 105 年 7 月各縣市鄉鎮區常住人口圓餅圖 （以新北市為例）

此為縣市下各鄉鎮區的人口數圓餅圖，一樣可以從左上的年度按鈕做年度篩選，左下方的縣市按鈕則可以做縣市的篩選，圓餅中的顏色代表此縣市的鄉鎮區，圓餅的大小則代表人口數的大小，以圖 54 為例，即為新北市民國 105 年 7 月各區人口數，而從圖形中可以明顯看出佔了新北市人口數最大比例的區為板橋區。

再利用 Power BI 中的 Power Map 功能做出的常住人口地圖立體直條圖，結合了地圖與資料視覺化，可以輕易地看出人口集中的趨勢，以南北為兩大高峰。(如圖 55)

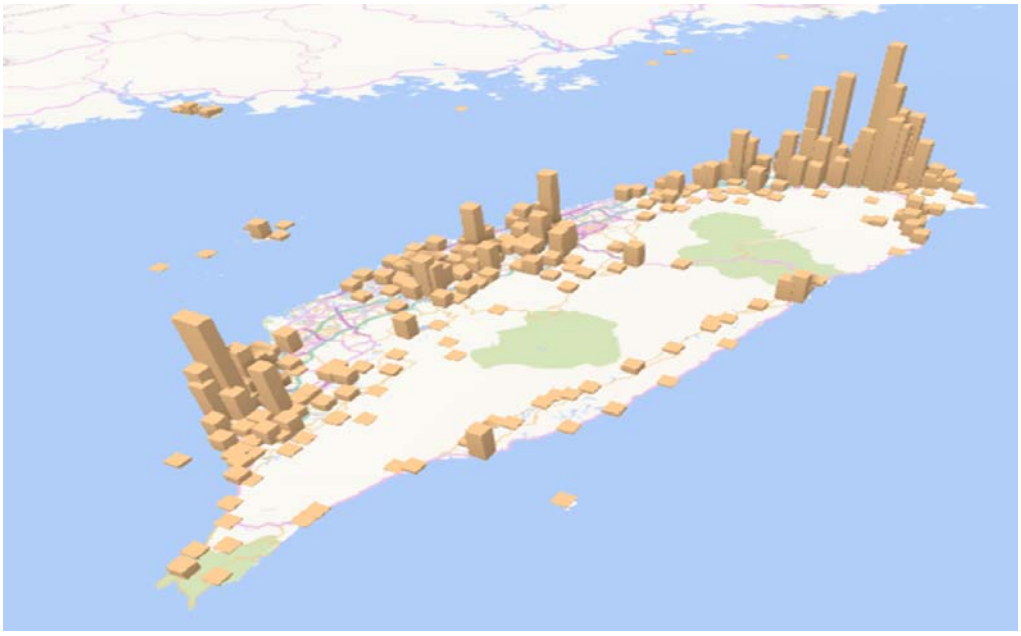


圖 55　民國 105 年 7 月各縣市鄉鎮區常住人口立體直條圖

此為另一種地圖資料視覺化的呈現，為民國 105 年 7 月各縣市鄉鎮區常住人口熱力圖（如圖 56)，顏色越紅的部分人口越密集，可看出人口大多集中在南北兩方。

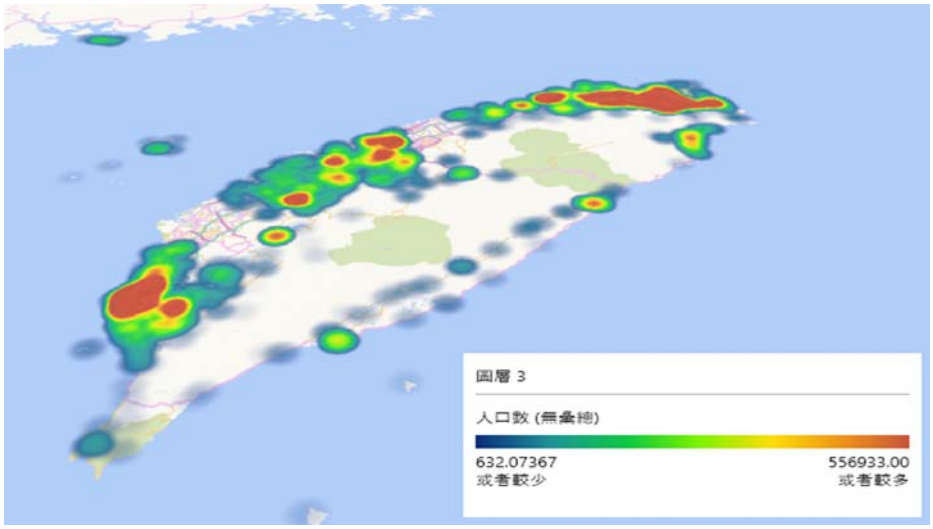


圖 56　民國 105 年 7 月各縣市鄉鎮區常住人口熱力圖

第四節、村里常住人口推估應用效益

本次建立的研究案例，不僅成功從普抽查或統計調查去識別化資料，結合其他跨公私領域資料，進行函數映射等方法，建立仿真村里常住人口統計調查應用資料庫，在未來更可以達到下列三方面應用效益：

一、就政策效益而言：呈現完整人口質量與家庭組成，提供小地區人口及分布狀況，以為政府規劃人力運用，制定住宅、都市計畫、產業、交通運輸建設及均衡城鄉發展政策之參考。

二、就企業經濟效益而言：結合空間地圖（GIS）與普查資料，展示小地區統計，可供廠商經營選址應用，提高企業經營成效。

三、就統計效益而言：運用本仿真村里常住人口資料檔，以供各有關調查抽樣設計與推估之參據；並建立時間序列，供學術團隊研究人口、社會變遷及城鄉發展所需。

四、延續運用村里常住人口模型，連結地方統計推展中心之家庭收支資料，以及其他跨領域相關資料，建立並推估臺澎金馬地區不同城市(村里、鄉鎮、縣市)間之消費水準及消費力，進而建立村里消費力指標（CCP）與村里人均消費力指標（PCP）。

第五章、歲計估算模型

第一節、 研究方法與資料來源

「歲計估算」模型建立目的是以資料探勘（Data Mining）及資料驅動（Data Driven）大數據分析方法進行案例研究。政府盱衡國內外經濟情勢，估測全國總資源供需狀況，編列歲入、歲出預算，配合財政健全興革措施及年度施政方針等重大政策與規劃目標執行。為期國家財政健全與經濟永續發展，政府亟應正視國家整體財政收入不足支應公共支出之嚴峻課題；而在精進財政策略，並兼顧經濟發展與財政健全平衡考量前提下，適度調整租稅工具與稅制結構，建構稅收成長與經濟發展之正向互動關係，以充裕國家財源，提升整體財政效能，達成健全財政目標。

但是我國目前一直存在法律義務支出逐年攀升，歲出結構僵化的問題，近 10 年來，民國 91 至 96 年度歲出規模在受到歲入財源不足及舉債額度限制下，難以大幅成長，歲出總額平均約控制在 1.6 兆元；惟民國 97 年度起受金融風暴及莫拉克風災，政府採行擴張性財政政策致歲出增加，而景氣將復甦之際，財政刺激方案不宜急速退場，民國 97 至 100 年度歲出規模控制在 1.7 兆元;民國 101 年度更因振興經濟擴大公共建設特別預算業於民國 100 年度全數編竣，相關延續性計畫須回歸總預算編列及新增如國防部推動募兵制、重大軍事採購、全民健保、勞保費率調整之中央增加負擔數等大額經費需求，歲出遽增至 1.9 兆元。就各項歲出內容來看，屬強制性或義務性必須編列之經費，包括人事費、債務利息、社會保險、福利補助、教育支出及對地方政府之補助支出等，其占整體歲出之比率，由民國 98 年度之 56%，攀升至 102 年度之 69%，約增 13 個百分點。在新興政事支出歲出規模縮減與依法律義務支出逐年攀升的雙重壓力下，歲出結構逐漸僵化，可運用之資源相對不足。如前所述，法定義務支出占歲出比例約達 7 成，嚴重排擠公共建設及其他重要施政計畫之資源配置，連帶影響經濟之

成長。

目前世界大多數國家的政府預算皆採取「年度預算」制，僅對於未來一個年度的收支作估測，依 Boex, Martinez-Vazquez, & Mcnab（2000）的看法，多年度預算制度讓將預算過程置於多年度的架構上，政府必須制定一致的政策目標及順序，並且檢查當前政策是否與財政策略不符，讓預算過程具有持續性，並且提升預算過程中的透明度與課責，但是若是以此種方法會有缺點，可能會過度依賴往前的推估，造成財政政策的無彈性，訂定周全的多年度預算可能會非常消耗行政成本。在 OECD 國家已全面實施多年度預算制度，其中部份 OECD 國家採行的多年度預算制度其共通作法包括：多年度預算收支估測、中程財政策略、依推估數調整各機關基線預算、調整計畫優先順序、定期更新收支估測等。但是從發展中國家實施的經驗中發現，失敗的原因有很多，可能是中程經濟成長與收入預估的不確實，也可能是經濟發展計畫缺乏特別重視的財政政策，因此無法將經濟發展計畫與預算過程結合，所以在推動多年度預算制度時可注意一些地方，例如：訂定目標、非所有機關一體適用、採漸進式推動、鼓勵參與、發展可信賴的預算估測。多年度預算制度有利於預算收支的控制，同時亦可提高資源的配置效率，但必須依賴更精確的預算估測模型。

近年來越來越多的國家預算開始呈現赤字的發展，因此，預測財政赤字在未來擬定政策時是有用的，但是，現在越來越多會影響財政的因子漸漸浮上檯面，基於理論，公共部門的政府預算赤字可能由許多不同原因所致。所以必須利用多種方法加以比較，找出一個最好的預測模式，預測往後幾年的財政赤字的趨勢後，可以讓政府更早擬定完善的政策，加以因應，若能找出一個較好的預測預算模型，即可讓政府減少不確定性的存在而產生的政策，也可以使國家的預算減少浪費，將這些原本可能被浪費的預算投資到其他的政策，讓政府資源更有效的被利用。

目前我國在歲計部分之估算仍不夠精確，針對法律義務的支出分析約占七成，此部分可依法律義務支出部分作為估算之主軸，因在政府歲計中

此支出為佔最大宗之額度，此部分已有看過預算部分的資料型態，但因資料種類太多，須加以整合及整理，以及瞭解資料欄位之定義後方可執行預估。建立預算模型主要是可以即時並且充分反映政府整體的財政狀況，以及減少不必要的資源浪費之情形，除了預測之外，也可以用來監測政府預算，通過建立基本建模的數據庫，利用這些資料來分析預算偏離的情況，並且可以將這些偏離的項目做成一份預算差異的報告，分析以及發現潛在的問題，實現自動預警的機制，為管理層及時採取有效的管理措施並且提供有效的依據，另一方面，充分運用預算分析和審查，從不同角度進行預算的實時分析，為全面實現內部控制管理提供具有深度的數據模型。

第二節、建立模型歷程

本次歲計估算模型建立以法務部所屬歲計會計資料為分析資料來源，並以法務部人事費用途別資料為主要分析項目，由於，法務部人事任用型態較為複雜多樣，以法務部作為預測建模資料較具代表性。透過建立歲計預算模型可以即時並且充分反映政府整體的財政狀況，以及減少不必要的資源浪費之情形，除了預測之外，也可以用來監測政府預算，通過建立基本建模的資料庫來分析預算偏離的情況。本研究先以法務部資料為主，法務部內因有許多機關，先以機關編號 23001000（法務部本部）當範例進行歲計估算模型的建立，分析主軸為以下兩點：一是找出預算-決算（借-貸）差異最大的用途別，並且找出其原因；二是結合決算資料及指數平滑法建立預測模型，並進行未來兩年的預算推估，以作為明後年預算編列時之參考依據。

首先，採分類最細的三級用途別資料先做探討。本研究統計自民國 96 年至民國 103 年度的預算與決算差異以比較其差異程度。從圖 57 可以看到，深綠色顏色區塊，即民國 100 年至民國 103 年度在 010301（職員待遇）之預算與決算差異甚大。在 014201（退休退職給付）其預算-決算之差異動盪幅度較大，差異值不是很大就是很小。

列標籤	96	97	98	99	100	101	102	103
⊞23001	232146151	543531992	322046548	147990222	-40688106	15452740	-12416122	-49041050
⊟23001000	18662538	26367886	28041147	5917168	60537638	92071370	52475939	16267475
010201	150120	790797	820	-769295	-1084880	-3731340	-3156316	-2631360
010301	16020587	13771326	5431866	-735795	45856846	63295849	38705765	42496447
010401	76287	-1141327	375966	-29979	-906280	697027	-342649	-1052984
010402	650416	976896	1117108	-21820	-486281	-1993251	-1413554	-742411
010501	1248764	716499	223920	592075	1515177	2865104	3234797	3757640
011101	-1884270	-1308001	-311649	-1099887	1020623	3338159	2991329	2020822
011102	19800	-5400	0	36000	-6000	494900	-767200	-406800
011110	911051	657540	-51018	-351116	-1055933	6676056	10348463	9755797
011199	320500	8015000	16075000	23250	242000	2000000	4000000	
012131	171170	310209	261984	379536	1247726	1384461	1269380	941178
012141	1443351	1096008	547152	497447	1596769	1823673		
013101	-4110541	602958	26442	53445	1883159	3004621	3483045	4092805
013102	2390808	1967724	2223760	1845600	-1896616	-2827724	-2838128	-2677104
013103	632378	795149	1422747	1709200	1771600			
014201	-3272033	-3445864	-288231	6006473	4200638	8305368	-6776817	-42731073
014301	126876	174131	203628	72863	30504	-71028	-16540	107280

圖 57 法務部本部三級用途別預算決算差異比較

接著再以二級用途別來作探討，從圖 58 可以看到，在民國 100 年度至民國 103 年度中的 0103（法定編制人員待遇）預算-決算差異值較龐大，而 0142（退休退職給付）在大部分的年份預算都是不足的。若是以法務部本部（23001000）來做探討時，趨勢是差不多的，皆是以 0103、0142 這兩個用途別為差異較大。此部分原因可能為職員調派異動或年度終了之預計員額與實際員額有落差，導致 0103（法定編制人員待遇）預算編製過多，而 0142（退休退職給付）因預算編製不足導致是常需要從其他用途別流用預算來彌補不足之缺額。

列標籤	96	97	98	99	100	101	102	103
⊞23001	145863090	445393386	208850222	-6076780	-40488106	18476740	-8844122	-48921050
⊟23001000	16048595	26459886	28141147	6017168	60637638	94210370	60014939	28806475
0102	150120	790797	820	-769295	-1084880	-3731340	-3156316	-2631360
0103	16020587	13771326	5431866	-735795	45856846	63295849	38705765	42496447
0104	726703	-164431	1493074	-51799	-1392561	-1296224	-1756203	-1795395
0105	1248764	716499	223920	592075	1515177	2865104	3234797	3757640
0111	-632919	7359139	15712333	-1391753	200690	12509115	16572592	16369819
0121	-999422	1498217	909136	976983	2944495	3308134	6769380	6441178
0131	-1087355	3365831	3672949	3608245	1758143	2215897	2683917	3454701
0142	-3272033	-3445864	-288231	6006473	4200638	8305368	-6776817	-42731073
0143	897562	407862	476963	-246853	3693555	4105425	3609217	3177110
0151	2996588	2160510	508317	-1971113	2945535	2633042	128607	267408

圖 58 法務部本部二級用途別預算決算差異比較

在找出預算-決算（借-貸）差異最大的用途別後，接著是結合決算資料及指數平滑法建立預測模型，並進行未來兩年的預算推估，以作為明後年

預算編列時之參考依據。指數平滑法是生產預測中常用的一種方法，也適用於中短期經濟發展趨勢預測，所有預測方法中，指數平滑是用得最多的一種。簡單的全期平均法是對時間數列的過去數據一個不漏地全部加以同等利用；移動平均法則不考慮較遠期的數據，併在加權移動平均法中給予近期資料更大的權重；而指數平滑法則兼容了全期平均和移動平均所長，不捨棄過去的數據，但是僅給予逐漸減弱的影響程度，即隨著數據的遠離，賦予逐漸收斂為零的權數。

利用民國 96 年度至民國 103 年度預算決算資料為基礎進行指數平滑法未來兩年預算之推估。指數平滑法是生產預測中常用的一種方法，也用於中短期經濟發展趨勢預測，所有預測方法中，指數平滑是用得最多的一種。本案例採用簡單的全期平均法，是對時間數列的過去數據一個不漏地全部加以同等利用；雖是利用指數平滑法作推估，但為了與實際值符合，所以先以民國 96~ 103 推估民國 104 年，利用民國 104 年的真實資料與推估出的值算出權重(如圖 59)，並且對後來推估出來的民國 105 及 106 兩年度之值乘上此權數做修正（如圖 60）。接著再推估出一倍標準差之信賴區間如圖 61，以作為未來預算編列區間之參考。

	96	97	98	99	100	101	102	103	104(推估)	104(原始資料)	權重	105(推估)	105(修正後)	106(推估)	106(修正後)
010101	0											0		0	
010201	11699760	18992829	17657760	19344710	21273180	29220000	27494928	25761421	30942572.36	25920060	0.837683	30824122.17	25820836.32	32603065.53	27311026.54
010301	7.11E+09	1.92E+10	1.95E+10	1.98E+10	2.53E+10	1.34E+10	1.36E+10	1.38E+10	16773099632	13718769494	0.817903	15486226996	12666232429	15353212412	12557439393
010302		0				0								0	
010303		0		0	0	0								0	
010305	1.44E+08	2.23E+08	2.3E+08	2.34E+08	2.36E+08	82363958	84753792	89073673	84179094.18	90393286	1.073821	68911320.89	73998429.17	51295964.02	55082687.59
010306								0							
010401	15111426	32248272	31826831	36425082	39369903	40222984	43274760	45518794	51363183.07	43468187	0.846291	51379501.83	43481997.42	54378374.67	46019915.77
010402	1.23E+08	1.82E+08	2E+08	2.75E+08	3.98E+08	1.88E+08	2.17E+08	2.11E+08	271329013.6	249041176	0.917857	271906816.9	249571516.7	280904470	257830073.6
010403	33574158	46080906	45102981	44944335	43342248	13915033	13356733	12392698	9782010.643	10906832	1.114989	5436014.389	6061094.989	665085.6222	741562.9989
010501	2.48E+08	7.09E+08	7.02E+08	7.08E+08	9.37E+08	4.85E+08	4.85E+08	4.79E+08	598141922.5	475004393	0.794133	544288289.9	432237432.3	536953279.5	426412456.6
011101	6.51E+08	1.84E+09	1.92E+09	2E+09	2.05E+09	1.47E+09	1.19E+09	1.39E+09	1596404533	1409795989	0.883107	1520841423	1343065679	1515774874	1338591374
011102	1350000	3587400	3902370	5723100	8604900	3988800	7167409	6305100	8064001.036	7452600	0.924181	8455681.944	7814584.222	9078336.578	8390030.071
011110	9.53E+08	2.6E+09	2.68E+09	2.74E+09	2.8E+09	1.9E+09	1.76E+09	1.78E+09	2116023468	1807964722	0.854416	1971176023	1684204719	1942706327	1659879750
011199	49359000	1.32E+08	1.12E+08	1.1E+08	1.4E+08	1.19E+08	1.1E+08	1.08E+08	128971535.5	111042690	0.860986	125215250.3	107808581	128232084.6	110406033.1
012111	76536475	1.16E+09	3.37E+09	3.33E+09	3.88E+09	3.86E+09	3.63E+09	3.47E+09	4890324437	3264551785	0.667553	4621632213	3085185428	4967120770	3315817424
01211101					68720		31710	131050	230390	56870	0.246842	98370	24281.87812	110950	27387.15439
01211102			64850		217990			86580	86580	78180	0.90298	69780	63009.93763	61380	55424.90644
01211103			112050	166950	279050		390900	262050	133200	133200	1	4350	4350	0	0
012121	1.56E+08	2.34E+09	6.9E+09	6.87E+09	6.1E+09	6.11E+09	5.99E+09	5.72E+09	7921254728	5399224113	0.681612	7444312980	5074134784	7920138354	5398463179

圖 59 法務部本部指數平滑未來二年預估表

105(修正後)	信賴區間下界	信賴區間上界	106(修正後)	信賴區間	信賴區間
25820836.32	21440776.32	30200896.32	27311026.54	22930967	31691087
12666232429	8111211634	17221253224	12557439393	8E+09	1.71E+10
	0	0		0	0
	0	0		0	0
73998429.17	28843014.06	119153844.3	55082687.59	14656208	95509167
	0	0		0	0
43481997.42	35754874.97	51209119.86	46019915.77	38418074	53621758
249571516.7	180630528.9	318512504.4	257830073.6	1.89E+08	3.27E+08
6061094.989	-5273740.761	17395930.74	741562.9989	-9263390	10746516
432237432.3	259914861.1	604560003.6	426412456.6	2.54E+08	5.99E+08
1343065679	992425293.1	1693706065	1338591374	9.88E+08	1.69E+09
7814584.222	6000859.222	9628309.222	8390030.071	6576305	10203755
1684204719	1221572675	2146836764	1659879750	1.2E+09	2.12E+09
107808581	85250953.24	130366208.7	110406033.1	87848405	1.33E+08
3085185428	2134882033	4035488823	3315817424	2.37E+09	4.27E+09
24281.87812	-2410.15235	50973.90859	27387.15439	695.1239	54079.18
63009.93763	22368.66424	103651.211	55424.90644	16679.89	94169.92
4350	-93375	102075	0	-96637.5	96637.5
5074134784	3388849148	6759420420	5398463179	3.71E+09	7.08E+09

圖 60 法務部本部信賴區間估計

	96	97	98	99	100	101	102	103	104	105(修正後)	標準差(Max-min)/4)	信賴區間下界	信賴區間上界	106(修正後)	標準差(Max-min)/4)	信賴區間下界	信賴區間上界
29001	6306653497	13739997718	14080371648	14389609129	14941316667	2570582880	2689616771	2843035339	2946322593	1596196171	3092683447	0	4688879617	2157295832	3336280124	0	3552009707
29001000	506257649	495288174	498186083	500566373	447510621	438976184	475190525	505594854	517311698	538877159.2	19583876.5	519289280.7	558461087.7	538386665.2	24975243.79	513411421.4	563361909
29010		1544657336	1625271194	1758127648	2046627379	1955959091	1974971836	2049833882	2114728407	2159621398	162628906.4	1996992486	2322250299	2256811549	153741014.2	2083070534	2390552563
29012		1365732488	1438950860	1544151164	1482836854					1617697532	44604669	1573092863	1662302201	1663348872	44604669	1618744203	1707935341
29015		345831488	1146183583	1239558495	1556462168	1513150980	1559980907	1668740234	1769258678	1820702484	41710171.8	1408596312	2237809656	1961813419	368717749	1598095670	2330531168
29016					207276080	1997506180	2079574053	2190299212	2278198335	2215405017	729799029.6	1485611987	2945198046	2529843582	516480563.8	2013363018	3046324146
29017						9432214766	9433224501	9626357336	9806959958	10022118229	93686298	9928431991	10115804527	10155425784	147475865.6	10007949919	10302901650
29018		2109581195	2208504411	2321584271	2654399684	2612144399	2697904971	2766159070	2866543980	2947114504	204729446.8	2742375058	3151853951	3052969451	209883327.3	2843586124	3262352778
29020		1484592065	1548904727	1655175940	1983848055	1907697157	1951121217	2054852751	2130168665	2195453108	177206558.3	2018246550	2372659666	2286331632	177715260.7	2108616371	2464046892
29030		8982266310	3534194335	3700029761	4059681991	2350874149	2408168522	2506152483	2574790671	3144594840	2084430191	1060164649	5229028030	5015557654	1657848040	0	2159403806
29031			1543333158	1648840164	1964692596	1912301987	1845528662	2068316004	2146980803	2237475278	150911911.8	2066563366	2388387189	2327867783	173535529.9	2154332254	2501403313
29032			1524179455	1624797265	1946401314	1891462066	1959166699	2052463964	2100400880	2159985158	159493858	2000554800	2319416016	2247212836	158951425.6	2088369011	2406368762
29033			1538385057	1640974104	1961164792	1910130408	1959815322	2055190628	2125486631	2184596195	163709176.5	2020887018	2348305371	2274182963	161552784.5	2112680179	2435735748
29034			1552450899	1655849056	1742605857	1808768706	1732822002	1834696635	2025551962	2145249504	168275265.8	1976974238	2313524769	2229424226	198199651.2	2041224575	2437622878
29035			308776157	209282627	474200080	185258218	302950090	177859254	168606381	134377188.7	76398412.25	57978776.48	210775601	115618741.8	84955710.32	30663081.5	200574452.1
29041			2108152468	2285610653	2557128892	2532877080	2594322097	2697818442	2795515744	2874531286	186546760.8	2687984525	3061078046	2980295021	191594704.4	2788700817	3171889726
29042	275743694		1698453504	1814652580	2139246878	2088520559	2144705185	2252042557	2334727380	2405189309	173805262.3	2231384047	2578994572	2502345590	176683951.3	2325661639	2679029541
29043			1916118397	2041066038	2389891882	2253159812	2404219698	2519774430	2602601092	2673881695	190421821.1	2483459875	2864303516	2779098107	189440824.4	2589657282	2968538981
29047			1827384545	1944162164	2266000602	2220369877	2274890404	2372447391	2463762280	2536050725	172669527	2363381196	2708720252	2633006474	177166545.1	2455839929	2810179019
29048			1606874851	1712783081	2029925378	1976138868	2033450049	2127105185	2212718798	2281340811	164600832.7	2116740478	2445941143	2373798323	168616489.9	2205181833	2542414813
29049			1547411665	1649408877	1966412219	1912369866	1963440408	2055542949	2136536980	2201148807	161139082.9	2040009725	2362287890	2290845650	163434285.6	2127411365	2454279986
29050			1974365829	2103066872	2438930084	2404574872	2469784770	2578778899	2679414869	2761524603	190700859.6	2570828743	2952225463	2869883091	196789698.5	2673099398	3066672785
29051			1544681302	1646713253	1962975171	1908628171	1959839642	2053132486	2139646380	2207957779	161010627.7	2046947152	2368968407	2298544203	165819119.3	2132725084	2464363928
29052			1661025700	1766077747	2082529750	2032316307	2083901154	2181602178	2262895380	2329656316	164370922.4	2165285394	2494027238	2421618680	167157654	2254461026	2588776334
29053			1560616140	1661627697	1977283565	1921527022	1972206231	2064622276	2147656959	2213620558	159657997.3	2053962560	2373278555	2302975296	163251104.4	2139724192	2466226401
29054			1610847674	1710370265	2027314060	1972371483	2023705974	2122944348	2209712685	2280340267	161241521.3	2119098746	2441581788	2371994205	167373148.3	2204561057	2539907553
29056			1787416946	1895159904	2217737798	2171356335	2217668828	2319450468	2406068232	2476242787	168044848	2308197599	2644267835	2570728816	172206460.2	2598522856	2742695227
29058			2056910306	2169179713	2499028259	2455980004	2519600780	2634775451	2725588169	2804368799	180782378.3	2623586420	2985151177	2907995379	186864622.2	2721070756	3094800002
29059			1595016243	1698635846	2020658344	1970070066	2027013044	2125051798	2207897980	2276195390	167382720.7	2108813669	2443578110	2370136242	170294786.7	2199841455	2540431029
29060			1509827513	1608640058	1924735202	1865306055	1909160755	2001549923	2085005452	2150218743	155862992.7	1994356550	2506081136	2357336780	160097807.5	2077258922	2597434537
29061			1527464974	1630550273	1953097503	1899042997	1950872547	2043247727	2121096967	2183718564	163789009.4	2019928655	2347508474	2274101288	164083997.6	2110037891	2438164686
29062			1525199743	1627111078	1948718043	1891788748	1944884313	2036783628	2117362489	2181707706	162801374.8	2019403831	2344009081	2271842281	164136990.7	2107715290	2435969271
29063			1555385448	1655077878	1975241799	1915079131	1964606536	2058275374	2140445380	2205948748	159061153.2	2046887594	2365009901	2294860665	162528536.2	2152832839	2457388991
29064			1467726700	1565483240	1878310484	1822965485	1873074955	1963124777	2040296842	2102225953	157103997.7	1945132015	2259339890	2189412597	158627513.2	2090785284	2348089910
29090		1347624258	1316828138	1411638720	1707685348	1779106994	1702699332	1801985154	1992783833	2113294721	168988923.8	1944905797	2282283645	2209820080	199116645.7	2010703434	2408986725
29091		1370108914	1443025249	1483376321	1857645105	1799955085	1727469944	1997622284	2017407702	2105234057	164780056	1940504001	2269964113	2197874943	183781285.9	2014099658	2381656229
29092		2091029666	3120080442	2077610855	3518908822	1853809501	1821722210	1745912572	1889060025	1089168541	428617192	1046306821	1132002603	1226947676	2679275538	9590206138	14948753215
29095		5776523539	5862608081	6003080449	6317143727	6340276321	6172968989	6407595462	6529969676	6630209894	189130901.9	6441078992	6819940796	6731751787	213421588.8	6518310199	6945153376
29096000					67778010												

圖 61 依法務部主管、彙編及單位機關別呈現信賴區間估計

第三節、視覺化呈現發展

透過預算、決算、預算-決算三種資料，可利用 Power BI 做視覺化之呈現如圖 62，以 X 軸為每個機關之預算總合，Y 軸為每個機關之決算總額，圓圈大小為預算-決算之差額，選擇年度時間軸可做動態之呈現。

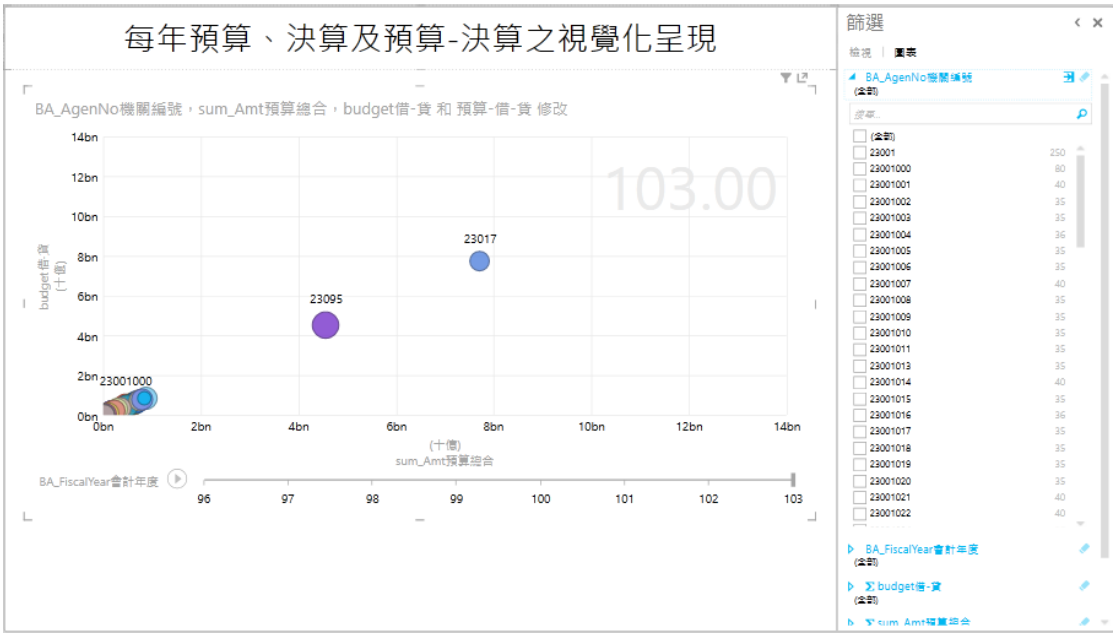


圖 62 法務部本部每年預算、決算及其差額圖

亦可改以長條圖做呈現，可看出每年度之預算-決算之差額，並且以各二級用途別做分類，可得知那些用途別之差額是較多如圖 63，每個長條圖之 X 軸為二級用途別，Y 軸為預算-決算之差額，每個長條圖代表各個年度。

圖 63 年度預算決算分佈圖

第四節、歲計估算模型應用效益

在導入預測模型前，我們觀察到因 2008 年由美國次級房貸所引發的全球性金融海嘯，導致全球經濟陷入嚴重的衰退。全球幾大經濟體紛紛陷入債務危機，像一些歐洲的國家債務規模皆已逼近或超過其 GDP 總值的 100%，伴隨而來的是削赤、削開支、減薪、裁員等等，而美國長期的財政赤字和龐大的政府債務，造成全球金融不安，若無法有效控制不斷擴大的赤字問題，長期下來更可能形成另一場全球經濟危機。然而我國為擴大內需，振興經濟，以大幅赤字預算挹注擴大公共建設支出經費，雖然此方法成功縮小景氣循環，保住經濟成長，但也讓收支短絀再度擴大，政府債務急遽增加。我國自民國 89 年下半年起全球性景氣衰退呈現入不敷出的情形，迄今仍未能收支平衡，如何積極開闢財源，並且控制政府支出規模及檢討支出結構，以縮短收支赤字達成財政收支平衡目標是目前政府刻不容緩需要解決的課題。

在導入預測模型後，善用並且發揮預算預測之功能，而預測功能通常包含了兩個方面，第一個是預測未來期間的預算，第二個是預測單位預算可能執行的結果。在預測未來期間的預算時，要根據財政的發展策略，對未來各個財政目標進行分析和預測，實現中長期預測目標的長遠規劃，為預算規模和預算規劃提供決策的支持，也為預算編製提供參考之依據。在預測預算單位的可能執行結果方面，須要根據各預算資金單位的執行情況及發展趨勢進行年度的比較，為預算調整和適時的糾正偏誤措施提供可行的依據。

本預測模型係以 VBA 開發設計，介面呈現分以下四個步驟進行：

第一步，先按開發人員的 Visual Basic（如圖 64）。

圖 64 歲計估算模型 Visual Basic 畫面

第二步，點選左欄的表單中的 UseForm2（如圖 65）。

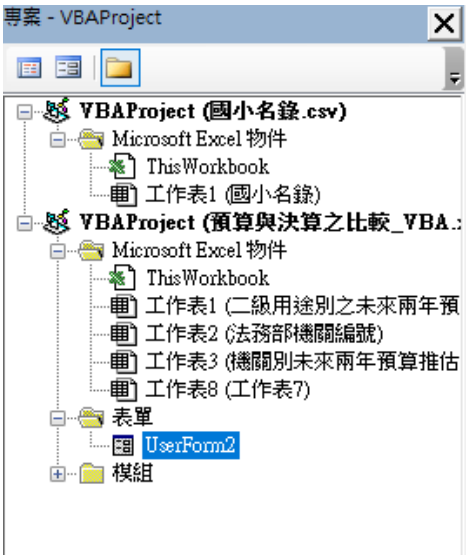


圖 65 歲計估算模型首頁（UseForm2)畫面

第三步，點選表單後會跑出一個視窗，接著按上面的執行巨集按鈕，或是按 F5（如圖 66）。

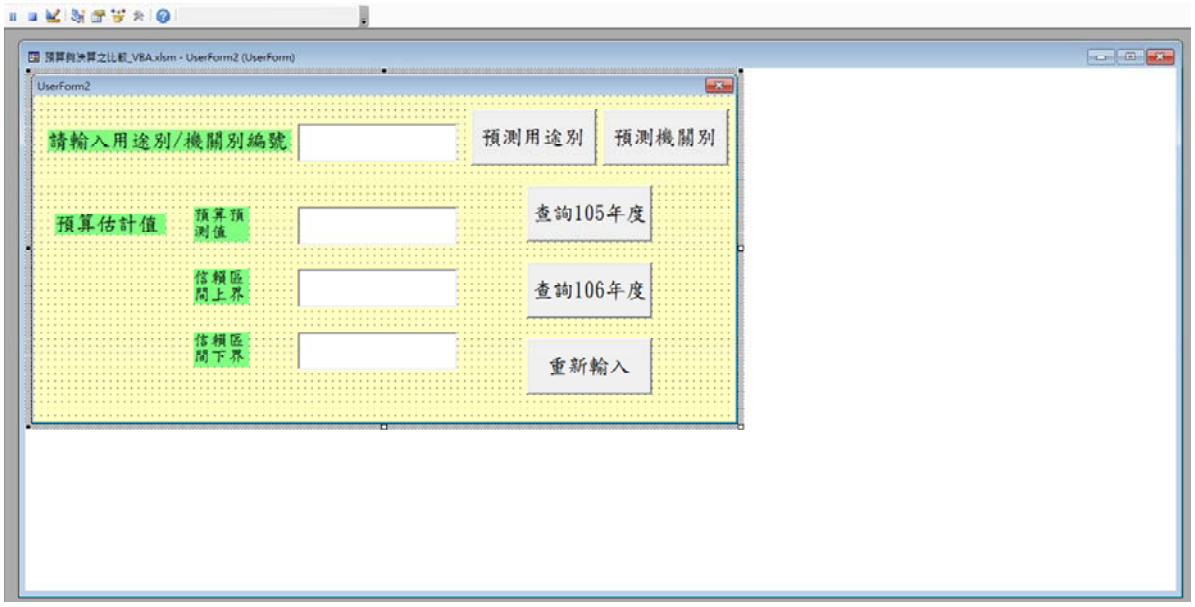


圖 66 歲計估算模型介面-輸入畫面

第四步，可輸入編號的視窗，將想要查詢的用途別編號輸入後，點選預測用途別。或是輸入預測機關別編號，點選預測機關別。

接著再點選要查詢民國 105 年度或是 106 年度，即可得到預算預測值及信賴區間上下界（如圖 67）。

圖 67 歲計估算模型介面-查詢畫面

若是要重新輸入編號，可點選重新輸入按鈕清除視窗上的值，再進行查詢（如圖 68）。

圖 68 歲計估算模型介面-重新輸入畫面

第六章、科技跨域整合資料模型

第一節、研究方法與資料來源

我國政府長期以來相當關心國內產業的科技創新與研發議題，積極推動並執行相關的工商及服務業普查、科技動態調查、產業創新調查等。其中工商及服務業普查係依我國統計法規定，為政府應定期舉辦之基本國勢調查，舉辦普查之目的係為蒐集工商及服務業營運狀況、資源分布、資本運用、生產結構及其他相關產業經濟活動狀況，俾掌握工商及服務業之經營現況與發展趨勢，提供政府研訂產業政策、工商業者發展業務及學術界研究之重要參據。本普查自民國 43 年首次創辦後，已建立每隔 5 年舉辦 1 次之規制，於民國 100 年辦理第 12 次普查，歷次普查結果對於政府重要經建計畫釐訂、工業區開發、產業輔導政策制定、地方產業發展策略研擬等，皆扮演重要角色，以發揮普查支援政策之效益。

科技部自民國 90 年起推動國內產業創新調查，目前已完成三次的調查。另一方面，科技部亦自民國 70 年起每年進行全國科技動態調查，以問卷調查統計全國企業、政府、高等教育及私人非營利部門研發經費與研發人力資料，彙編出版科學技術統計要覽，供產、官、學、研各界參考運用。然這些調查係由不同部會執行及不同調查組織進行，對於想利用這些資料進行大數據研究的使用者而言，存在著最根本且難以解決的問題：調查結果各自表述與使用，無法進行完整的對應關連，故欠缺整合性的分析功能。

為解決上述問題本計劃依循政府進行開放資料（Open Data）的政策，從大數據思維與角度出發，分析技術以機器學習、資料採礦及統計等演算法為本。本研究之目的即是嘗試將過往執行的企業面的科技議題相關的調查，如工商服務業普查及產業創新調查，以資料映射（Data Mapping）方式進行資料庫的串聯進行「科技跨域整合資料」。

由上所述可知，本研究之主要目標：針對民國 100 年度工商普查與民國 100 年第三次產業創新調查 TIS3 進行函數映射（Functional Mapping），有關科技跨域資料庫函數映射研究概念，詳如圖 69。

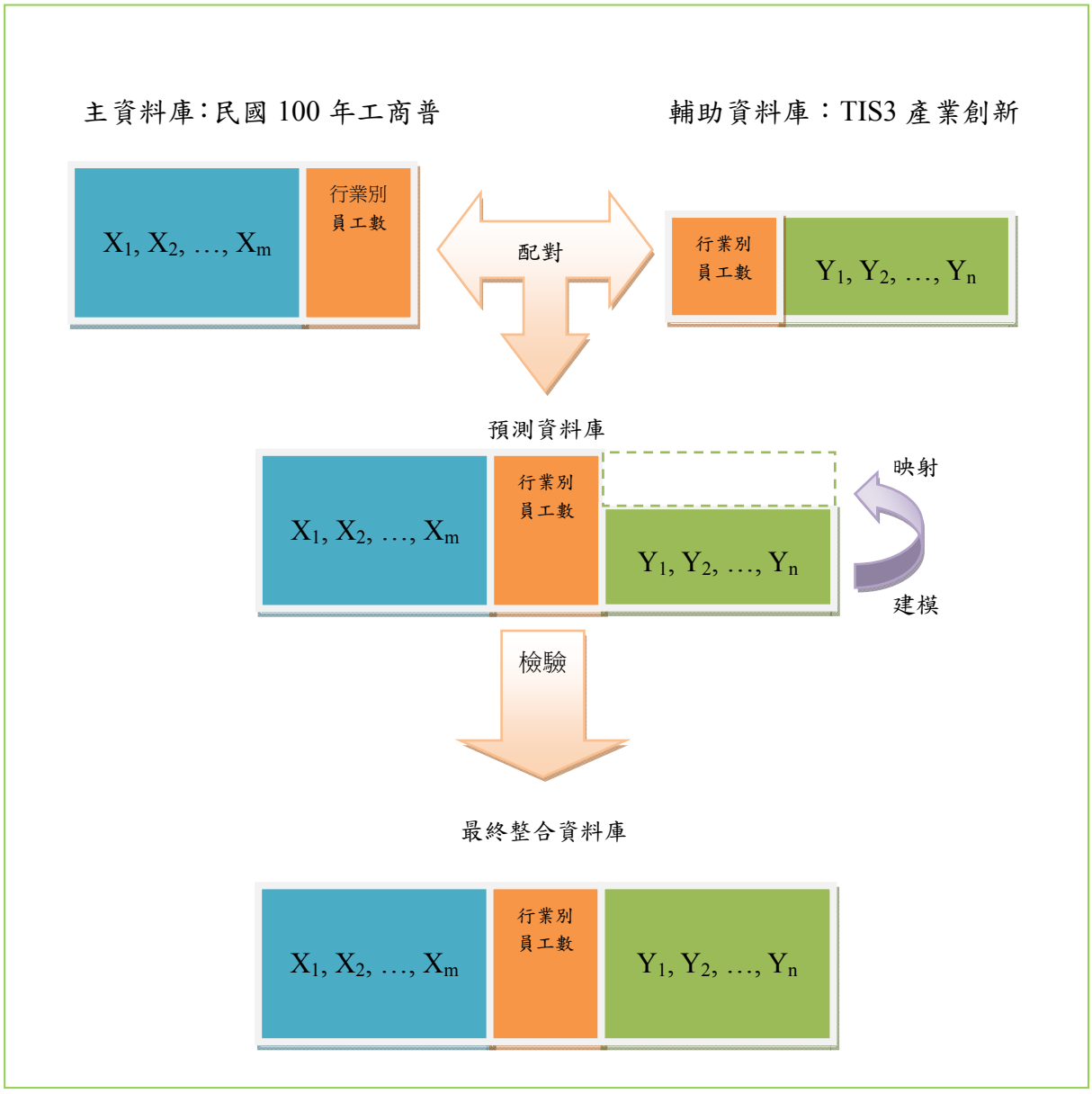


圖 69 科技跨域資料庫函數映射研究示意圖

第二節、建立模型歷程

一、專業認知與定義問題

國內有許多政府或學術機構會針對企業進行不同的調查，然而對於想利用這些資料進行大數據研究的使用者而言，存在著最根本且難以解決的問題：調查結果各自表述與使用，無法進行完整的對應關連，故欠缺整合性的分析功能。前述研究目的提及本研究即是嘗試將過往執行的企業面的科技議題相關的調查,如工商服務業普查及產業創新調查,以資料映射(Data Mapping）方式進行資料庫的串聯進行「科技跨域整合資料」。本研究所使用的資料庫有兩個，分別為民國 100 年工商與服務業普查(簡稱工商普查)、科技部於民國 100 年進行的台灣地區第三次產業創新調查(簡稱創新調查)。以工商普查為主資料庫、創新調查為輔助資料庫，用函數映射（Functional Mapping）方式將創新調查資料映射至工商普查的資料庫中，以擴增工商普查資料庫之應用價值及提升科技跨域資料之整合分析能力。由於創新調查資料內容繁多，本研究將著重在創新活動及研發經費之映射與推估。

二、資料來源之檢視與取得

(一)工商普查

工商及服務業普查係依我國統計法規定，為政府應定期舉辦之基本國勢調查，舉辦普查之目的係為蒐集工商及服務業營運狀況、資源分布、資本運用、生產結構及其他相關產業經濟活動狀況，俾掌握工商及服務業之經營現況與發展趨勢，提供政府研訂產業政策、工商業者發展業務及學術界研究之重要參據。工商普查自民國 43 年首次創辦後，已建立每隔 5 年舉辦 1 次之規制。歷次普查結果對於政府重要經建計畫釐訂、工業區開發、產業輔導政策制定、地方產業發展策略研擬等，皆扮演重要角色，以發揮普查支援政策之效益。工商普查目前已完成第 12 次（民國 100 年)，正進行第 13 次（民國 105 年)。

(二)台灣地區產業創新調查

自 2000 年的初期規劃開始，迄今由科技部支持完成的台灣地區產業創新調查已進行三次，目標是以實證資料來具體描繪台灣的產業創新系統。調查執行係依循歐盟統計局發表的奧斯陸手冊（OSLO Manual）及每一調查年度的問卷，以符合國際調查標準。（OECD & EUROStat，2005）產業創新調查目的為欲瞭解產業界中普遍存在創新活動，包含技術及非技術創新的投入及成果現況。調查對象係以企業為調查單位，調查母體主要為工商及服務業普查，採取分層比例隨機抽樣法。調查項目包括受調查單位現況、創新活動類型、投入創新活動之人力及經費、創新成效、成功及不成功創新、創新活動訊息來源及合作單位、創新活動受阻原因等資料。

三、 資料使用

本研究將採用第 12 次（民國 100 年）工商及服務業普查做為主資料庫（以下簡稱工商普查），以及民國 100 年的第三次產業創新調查作為輔助資料庫（以下簡稱創新調查）。

(一)主資料庫

工商普查資料庫涵蓋 1,272,023 筆企業資料（第一類），變數包含：

表 7 工商普查資料庫之變數

起點	長度	欄位內容
00001	01	表別代號
00002	01	單位級別代號
00003	04	主要業別代號
00007	04	次要業別代號
00011	01	組織別
00012	03	開業年
00015	02	開業月
00017	01	製造業主要經營方式
00018	07	僱用員工－男
00025	07	僱用員工－女

起點	長度	欄位內容
00032	15	僱用員工－全年薪資
00047	07	自營及無酬家屬－在職人數
00054	15	自營及無酬家屬－全年薪資
00069	07	合計－在職人數
00076	15	合計－全年薪資
00091	15	各項支出合計
00106	15	營業收入
00121	15	各項收入合計
00136	15	資產總計（淨額）
00151	01	有無使用電腦或網路設備
00152	01	有無利用電腦資訊系統協助內部管理作業
00153	01	有無透過網路提供營業資訊
00154	01	有無自有品牌
00155	01	有無研究發展支出
00156	15	生產總額
00171	15	生產毛額（附加價值）
00186	15	生產淨額（市價）
00201	15	生產淨額（成本）
00216	15	年底實際運用資產淨額
00231	15	年底實際運用固定資產淨額
00246	01	超商攤計產值檔註記
00247	01	個別資料隱藏註記
00248	10	流水號

(二)輔助資料庫

創新調查資料庫涵蓋 13,841 筆企業資料，變數包含：創新技術與產品相關變數、創新活動相關變數及佔營業額比例等相關變數。

由於本次嘗試推估各縣市的研發經費，本研究欲從輔助資料庫映射（mapping）及推估的創新資訊有：

1. 技術創新活動與花費

(1)民國 96~ 99 年間，是否曾進行各項創新活動

●公司內 R&D

●委外 R&D

●取得機器設備軟體

●取得外部知識

●研發人員培訓

●創新產品行銷活動

●創新產品設計活動

●其他預備創新活動

(2)民國 99 年各項技術創新活動的總花費佔營業額之百分比

(3)上述各項創新活動的花費佔創新活動總花費之百分比

●公司內的研發活動（R&D）（包含 R&D 設備與建築的資本支出）

●委託其他公司或機構研發

●取得機器、設備與軟體的技術（不包含 R&D 設備的資本支出）

●取得外部知識

2. 其他技術商品化的支出費用

3. 映射及推估國內各縣市研發經費

四、 資料整理與資料預處理

(一)資料整理

本階段主要是進行資料格式轉換與一致化、資料標準化、分層、抽樣配對、資料函數映射準備工作...等。使得後續映射與推估作業順利進行。在了解兩個資料庫與欲完成目標後，接著挑選能將兩資料庫配對之連結變數。而在工商普查中主要業別代號與創新調查中大業別相似、工商普查中合計_在職人數與創新調查中民國 100 年員工人數相似。因此本研究選擇此二變數為工商普查以及創新調查的分層依據變數，並做為兩資料庫間的連結變數。

(二)資料分層

為了能將兩資料庫進行分層抽樣，以配對出預測資料庫。本研究將挑選出的分層依據變數轉換成相同格式與名稱的變數。將工商普查中主要業別代號轉換成大業別的代號（標準行業分類：B~S），並命名為 mainind100a（大業別代號）；將工商普查中合計_在職人數與創新調查中民國 99 年員工人數轉換成四個區間（0：0 人、1：1~5 人、2：6~20 人、3：20 人以上），並命名為 emp100（員工合計）。

在抽樣配對成預測資料庫前，本研究應先對資料庫進行預處理。下表為本研究中所產生之衍生變數，如利用開業年轉換成 opy100a(公司年齡)。另外，針對後續的研究所使用的資料，其中「製造業主要經營方式 mbtype100」（90%遺漏值）、「有無利用電腦資訊系統協助內部管理作業 pcmmanag100」(47%遺漏值)及「有無透過網路提供營業資訊 pcinfo」（47%遺漏值），這三個變數遺漏值過多將不予以插補與後續納入建模的討論；「組織別 organization100」(0.1%為遺漏值)、「有無使用電腦或網路設備 pcuse100」（6%遺漏值）、「有無自有品牌 ownbrand100」（6%遺漏值）、「有無研究發展支出 rd100」（6%遺漏值）這四個變數將予以插補與後續納入建模討論。

表 8 本研究所產生的衍生變數

變數名稱		變數型態	變數說明
大業別代號	mainind100a	類別	參照標準行業分類之大行業別
公司年齡	opy100a	連續	100-開業年
員工_合計	emp100	類別	0：0 人 1：1-5 人 2：6-20 人 3：20 人以上

五、建立預測資料庫

資料清理與整理完畢後，接著本研究依分層變數「大業別代號：C~S；員工_合計：1~3」對主資料庫與輔助資料庫各別分層，並在兩資料庫中的相對應層數選取相同數量之樣本來配對，最後合併成預測資料庫含 7,778 筆企業資料，成為預測資料庫。

(一)資料建模與映射

由於建模的應變數（dependent variable）包含類別變數與連續變數，如是否有各項創新活動、創新活動佔營業額百分比、各項創新活動佔創新活動費用之百分比、營業額等，因此本研究嘗試使用的模型為多元迴歸模式、C5.0 分類樹、RandomForest 隨機森林演算法。其中以 RandomForest 隨機森林演算法的正確預測率最高，且此演算法能夠針對類別變數及連續變數進行建模與預測。故本研究之映射模型採用 RandomForest 隨機森林演算法。自變數（independent variables）的選擇來自於主資料庫中的變數，但不含遺漏值過多的變數，如「製造業主要經營方式 mbtype100」（90%遺漏值）、「有無利用電腦資訊系統協助內部管理作業 pcmmanag100」（47%遺漏值）及「有無透過網路提供營業資訊 pcinfo」（47%遺漏值），這三個變數遺漏值過多將不納入後續建模的程序。

(二)建立模型與統計分析

經過不同模型的測試，如羅吉斯回歸、C5.0 演算法及 RandomForest 隨機森林演算法等，其中以 RandomForest 隨機森林能夠獲得最佳的預測正確率，至少達八～九成以上。故本研究後續的模型將透過預測資料庫並以 RandomForest 隨機森林演算法來建構 Data Mapping 用的映射函數及模型，模型欲映射變數預測正確率如表 9 所示。

表 9 技術創新活動預測正確率

項目編號	應變數 （欲映射之變數）	正確率
1	有沒有技術創新活動	99.1%
2	公司內的研發活動	91.8%
3	委託其他公司與研發機構	95.3%
4	取得機器、設備與軟體的技術	82.3%
5	取得外部知識	82.3%
6	研發人員培訓	92.17%
7	為推出創新產品的行銷活動	84.83%
8	為推出創新產品的設計活動	92.68%
9	其他預備創新的相關活動	90.27%

由於表 9 中各項欲映射的變數過於分散，為了提高預測效果，故本研究採取二階段的預測建模。第一階段依據樣本比例合併各項百分比組別後建模，並將合併後的組別進行第二階段的預測建模。

表 10 各項創新活動的花費正確率百分比

項目編號	各項創新活動花費	正確率	
		第一階段	第二階段
1	技術創新活動的總花費佔營業額之百分比	85.39%	98.43%、99.78%、94.74%
2	公司內的研發活動	85.06%	97.5%、96.97%
3	委託其他公司或機構研發	84.55%	97.78%
4	取得機器、設備與軟體的技術(不包含 R&D 設備的資本支出)	85.82%	99.39%、85.06%
5	取得外部知識	85.02%	89.89%
6	其他技術商品化的支出費用(包含研發人員培訓、行銷、設計及可行性測試等活動)	83.54%	97.04%、96.67%、97.06%

(三)全國各縣市研發經費推估

以兩部分來進行全國各縣市研發經費的預估，第一部份為映射工商普查之營業額；第二部份為利用第一部份的營業額、民國 100 年的各項技術創新活動的總花費佔營業額之百分比、公司內部的研發活動（R&D，包含 R&D 設備與建築的資本支出）花費百分比與委託其他公司或機構研發花費百分比來找出全國各縣市研發經費。

1.工商普查之營業額

找出預測資料庫中來自創新調查有營業額的樣本（7,778 筆），再抽樣 8 成的樣本來建立測試集，利用隨機森林挑選出建模自變數作為預測營業額的映射函數，並以所有樣本為測試集，並計算其 MSE(Mean Square Error，均方誤差)，最後映射回工商普查主資料庫。其最適映射函數模型與結果如下表 11。

表 11 工商普查之映射營業額_最適模型

自變數	ttemp100(合計_在職人數) tts100(合計_全年薪資) expen100(各項支出合計) income100(營業收入) tp100(生產總額)	npcost100(生產淨額(成本)) nfaye100(年底實際運用固定資產淨額) mainind100a(大業別代號) rd100(有無研究發展支出)
%Var explained	88.61%	
MSE	12.48	

將工商普查之營業收入（income100）與映射至工商普查之營業額依大業別代號（mainind100a）加總，並對其進行 Kolmogorov-Smirnov 檢定來檢查工商普查之營業收入與映射至工商普查之營業額在大業別代號加總的整體趨勢一致，透過 Kolmogorov-Smirnov 檢定，此檢查是用來檢測樣本趨勢是否一致，檢定後其結果見表 12，P 值大於

0.05 為顯著，即表示兩樣本的整體趨勢為一致的。

表 12 兩樣本之 Kolmogorov-Smirnov 檢定

兩樣本	營業收入(income100)	映射之營業額
最大垂直差距 D	0.375	
P 值	0.2145	

表 13 營業收入與映射之營業額_整體趨勢(千元/%)

	營業收入(income100)	映射之營業額
C_製造業	26,529,390,212(48.48)	12,767,803,144(22.77)
D_電力及燃氣供應業	7,33,867,700(1.34)	363,338,030(0.65)
E_用水供應及污染整治業	1,22,405,561(0.22)	187,232,988(0.33)
F_營造業	1,893,418,520(3.46)	2,810,179,974(5.01)
G_批發及零售業	13,176,465,780(24.08)	11,251,909,212(20.07)
H_運輸及倉儲業	1,207,646,664(2.21)	10,305,660,783(18.38)
I_住宿及餐飲業	552,476,578(1.01)	1,355,055,604(2.42)
J_資訊及通訊傳播業	876,409,716(1.6)	1,128,607,833(2.01)
K_金融及保險業	7,004,267,987(12.8)	3,661,725,695(6.53)
L_不動產業	798,696,929(1.46)	1,313,978,999(2.34)
M_專業、科學及技術服務業	502,169,659(0.92)	1,850,747,104(3.3)
N_支援服務業	294,600,378(0.54)	834,436,743(1.49)
P_教育服務業	81,729,983(0.15)	127,134,195(0.23)
Q_醫療保健及社會工作服務業	652,829,010(1.19)	6,610,481,502(11.79)
R_藝術、娛樂及休閒服務業	92,532,969(0.17)	1,097,840,548(1.96)
S_其他服務業	203,860,932(0.37)	406,955,366(0.73)

2.全國各縣市研發經費

利用第一部份的工商普查之營業額、民國 100 年的各項技術創新活動的總花費佔營業額之百分比、公司內部的研發活動（R&D，包含 R&D 設備與建築的資本支出）花費百分比與委託其他公司或機構研發花費百分比來找出全國各公司的研發經費。並將全國研發經費依「民國 100 年全國各縣市之企業家數比例為權重」推估至各縣市研

發經費（如圖 70）。

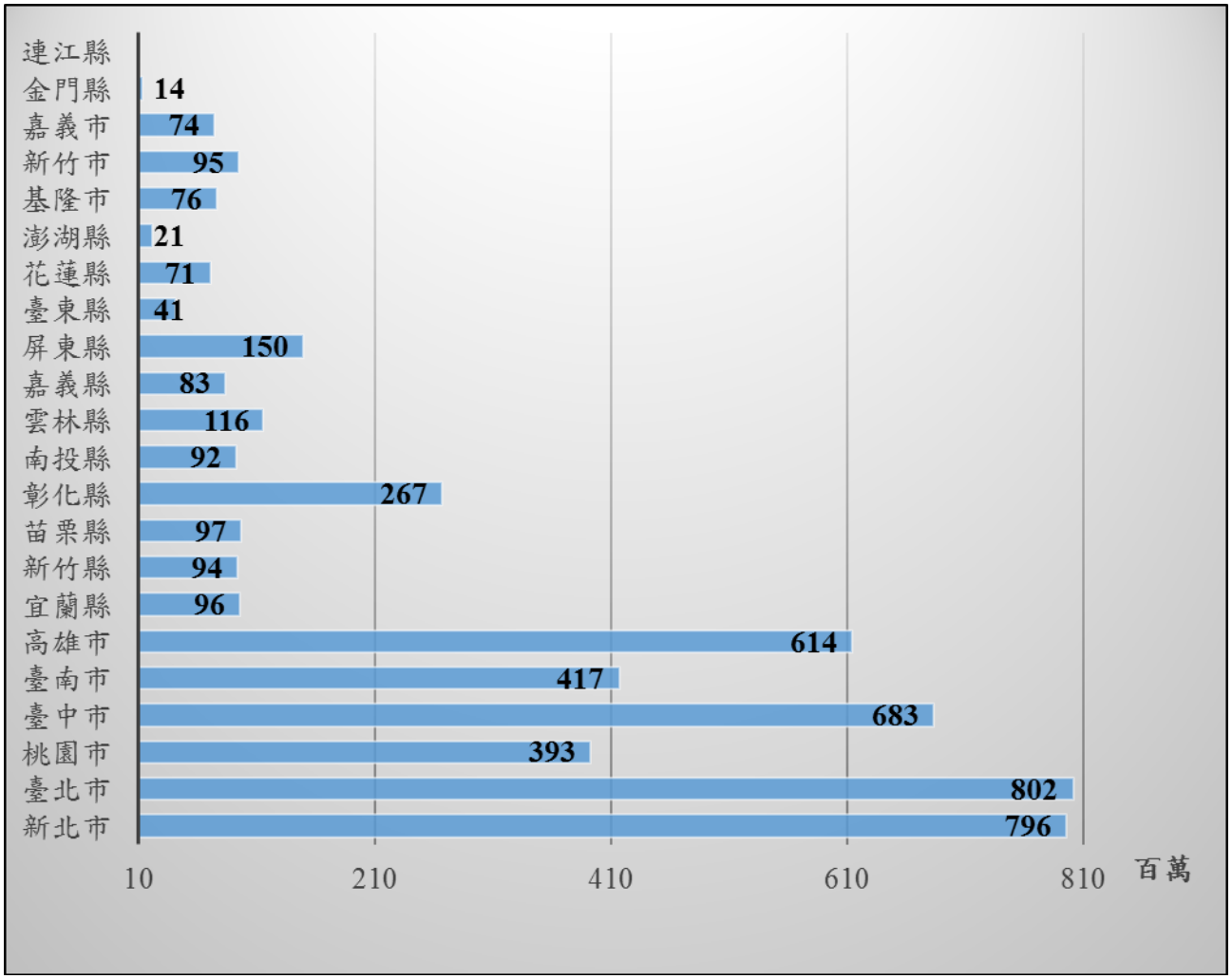


圖 70 全國各縣市之研發經費

第三節、科技跨域整合資料模型應用效益

科技跨域整合資料庫繼續收集、儲存與管理國內主計總處與科技部調查所收集之企業研發與創新資料，未來提供多元資料整合供應與加值應用服務；另外，透過整合模型建立，未來可提供資料使用者整合數個大小不同資料集，將可創造更多資料價值效益。

第七章、客製化 Text Mining

第一節、平台設計背景

資料採礦（Data Mining）與文字探勘（Text Mining）關係緊密，相較於前者顯著的結構化，後者長短不一、沒有規律，且尚有現今生活中隨口說出來的字彙，或是從社群網站、BBS 衍生出的用語，都屬於非結構化的資料。又像是 Web 頁面、新聞張貼，以及電子郵件訊息等文件，都包含了非結構性、自由型態的文字，或是許多符合特定電腦語言的語法及文法規則，構成文字和語句的字串流。這些非結構化的資料，隨著使用者每天更新及增減而無法事先定義。一般說來，非結構化的文字探勘功能主要是從獲得的文本資料中，分析訊息得到的詞頻基本特質，建立詞頻關聯表與正負向情緒判別，因此非結構化的文字探勘可以進行字詞的重要性分析、關聯性分析、集群分析、文章的正負向情緒分析、語意分析、脈絡分析...等。

面臨事件當中報導陳述的真實意義可能會因為文字的前後不同而有所差異，並且有可能因為不同的生活經驗、受教育情形與社經背景的不同而對同一段文字的意義有不同的判斷，大數據分析如何進行文字語意的判斷，演算法如何分析複雜的語言關係與文字詞義結構，將是大數據分析在運用上將受到的質疑。文字探勘的技術類似社會科學研究方法中的內容分析法，Berelson（1952）定義內容分析方法是針對傳播的明顯內容，做客觀、系統與定量的研究，內容分析方法是運用量化的分析策略，將所欲分析文本中的文字進行斷字與斷句的拆解，拆解為可供分析的單位（字、句、陳述等），接著針對文本進行編碼表的製作，訓練編碼員將分析文本進行量化的編碼，再進行統計分析。大數據分析則是運用類似的方法進行文字探勘，差別在於大數據分析方法是設計出一套演算法進行斷字與斷句，並且將正負面語意例字與例詞寫入程式中訓練演算法自動進行判斷。然而內容分析方法在方法論上的客觀性也遭到挑戰，包括編碼員的訓練、分析單位、信

效度甚至於研究的可重製性都備受質疑（游美惠，2000）。大數據分析所運用的演算法是使用電腦軟體進行分析，政府應用大數據精進公共服務與政策分析之可行性研究可減去因編碼員差異導致的差異，而演算法程序、斷字斷句邏輯與判斷正確性都是可能存在的問題。

本計畫為主計總處開發一套客製化的 Text Mining 平台，讓同仁可以簡易的使用文字探勘，瞭解貼近業務及政策需要，隨時應變並修正施政方針。

第二節、Text Mining 平台設計

此次平台設計上主要為 4 大塊，分別是 Crawl（爬文）、Text Mining（文字探勘）、Text Mining_Auto（自動分析）以及 Facebook（臉書）四大功能，如圖 71 所示。在首頁的部分除了功能的基本介紹外，亦提供了相關平台的超連結，如 Fanpage。

圖 71 客製化 Text Mining 平台-首頁

一、Crawl（爬文）

提供了新聞網站 ETtoday 的爬文功能，使用者僅需輸入關鍵字、爬文頁數以及存檔路徑，即可獲取相關文章，如圖 72。

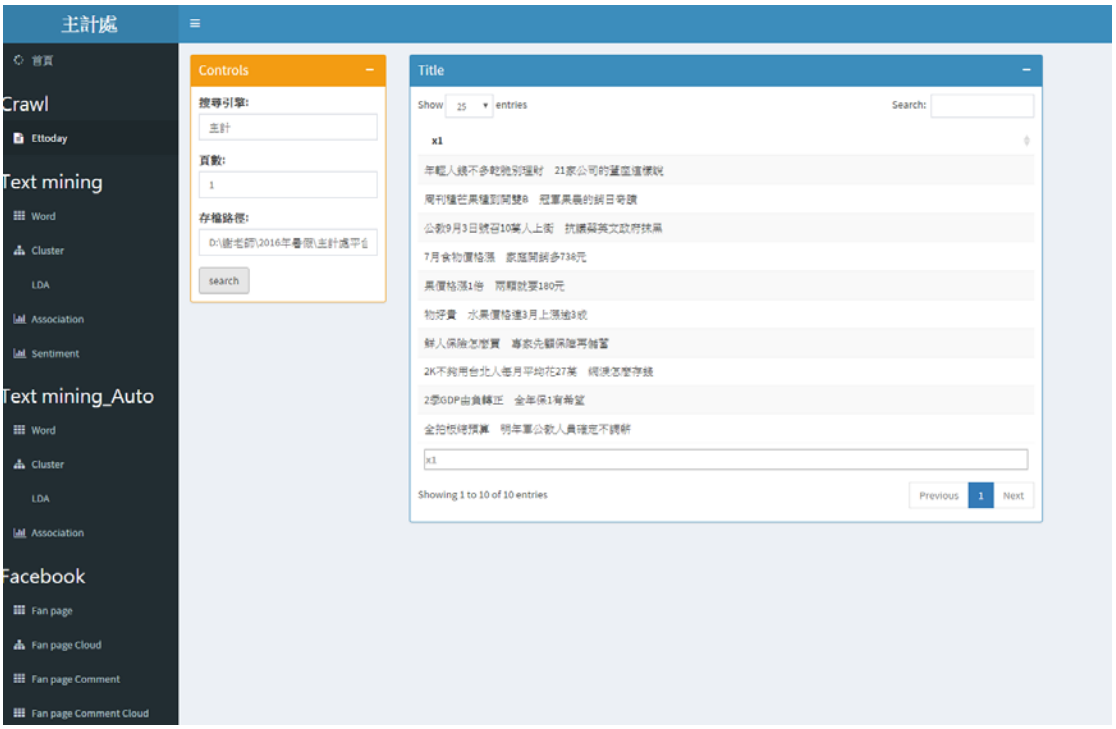


圖 72 客製化 Text Mining 平台-Crawl 分析頁

二、Text Mining（文字探勘）

主要提供使用者自行匯入文檔進行分析的便利性上，使用者僅需將文檔匯入後即可進行後續一系列分析。

(一)Word 分析（文字雲）

依照收集到的文檔資料中，文章詞彙分析機會自動分析所有文章內有哪些字詞，並且統計有哪些字詞出現的頻率高，頻率高的字詞在文字雲中的字體就會較大，反之頻率低的字詞就會較小，如圖 73。

圖 73 客製化 Text Mining 平台-Word 分析頁

(二)Cluster 群聚分析

是指將文字文件自動地分成幾個群集。因此，文件同屬在一個群集內的相似度會較高，而群與群之間的相似度就比較低，如圖 74。

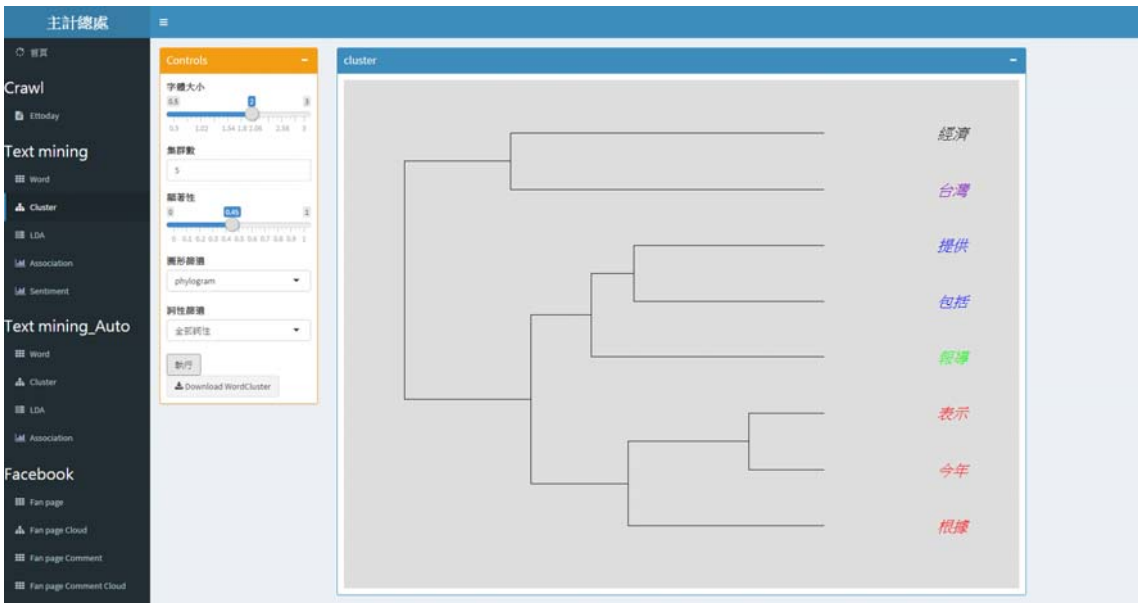


圖 74 客製化 Text Mining 平台-Cluter 分析頁

(三)LDA 脈絡分析

指將文字文件自動地分成幾個群集。因此，文件同屬在一個群集內的相似度會較高，而群與群之間的相似度就比較低，如圖 75 至 77。

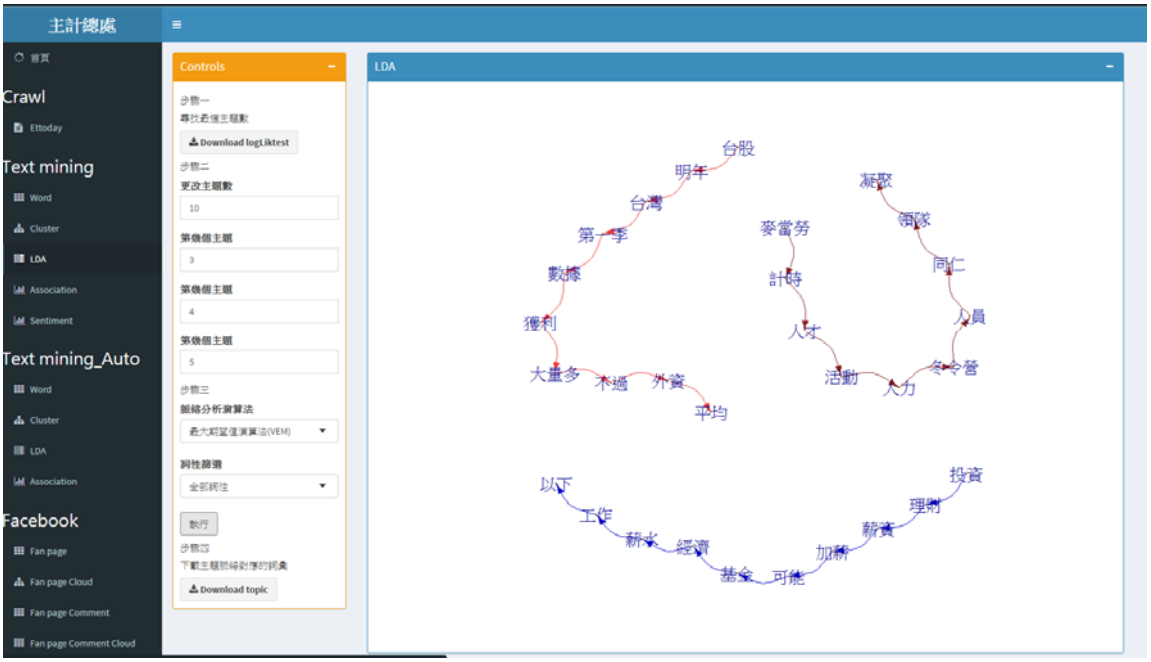


圖 75 客製化 Text Mining 平台-LDA 分析頁

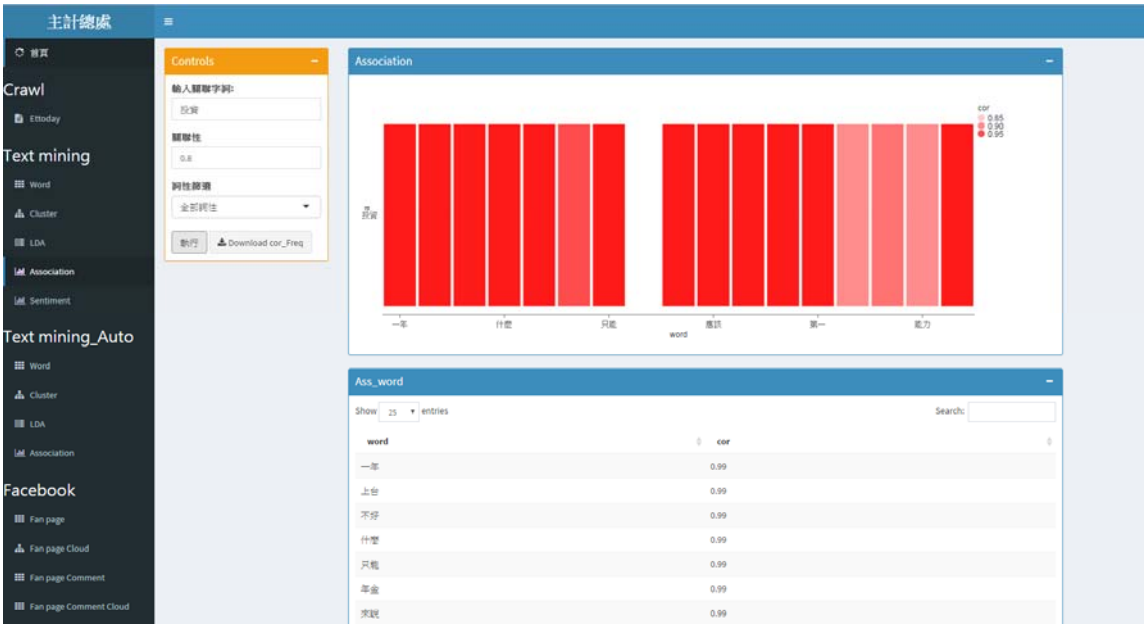


圖 76 客製化 Text Mining 平台-Association 分析頁

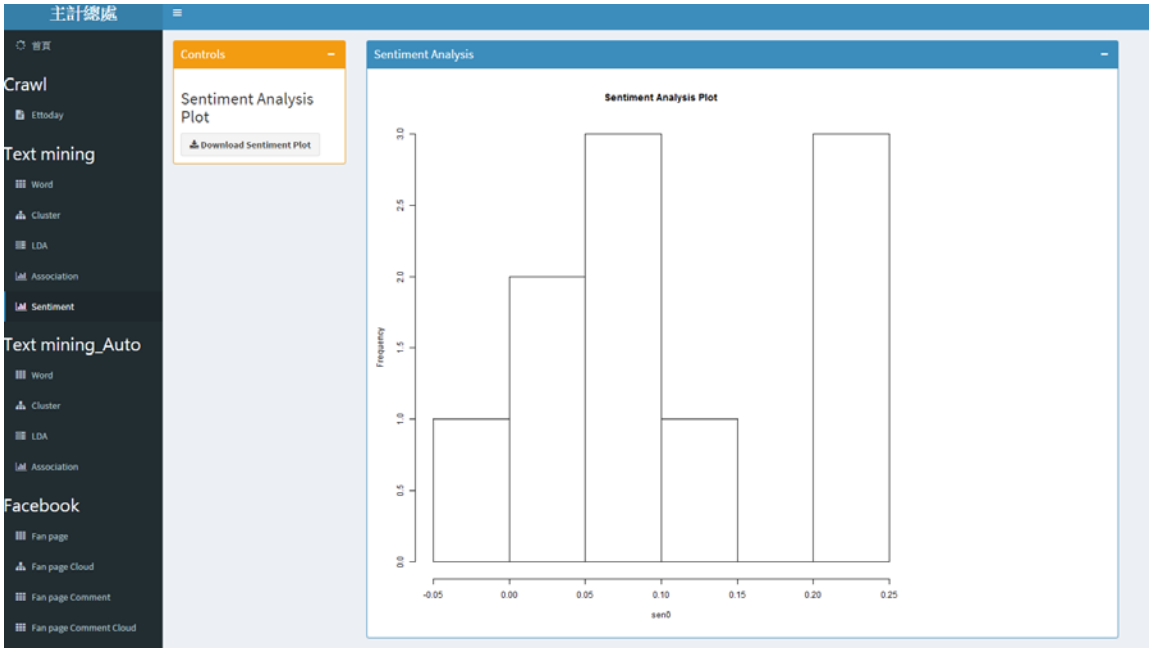


圖 77 客製化 Text Mining 平台-Sentiment 分析頁

三、Text Mining_Auto（自動分析）

可以直接對第一功能 Crawl 抓取下來的文章進行分析，無須另外匯入文檔，如圖 78 至 81。

圖 78 客製化 Text Mining 平台-Sentiment 分析頁

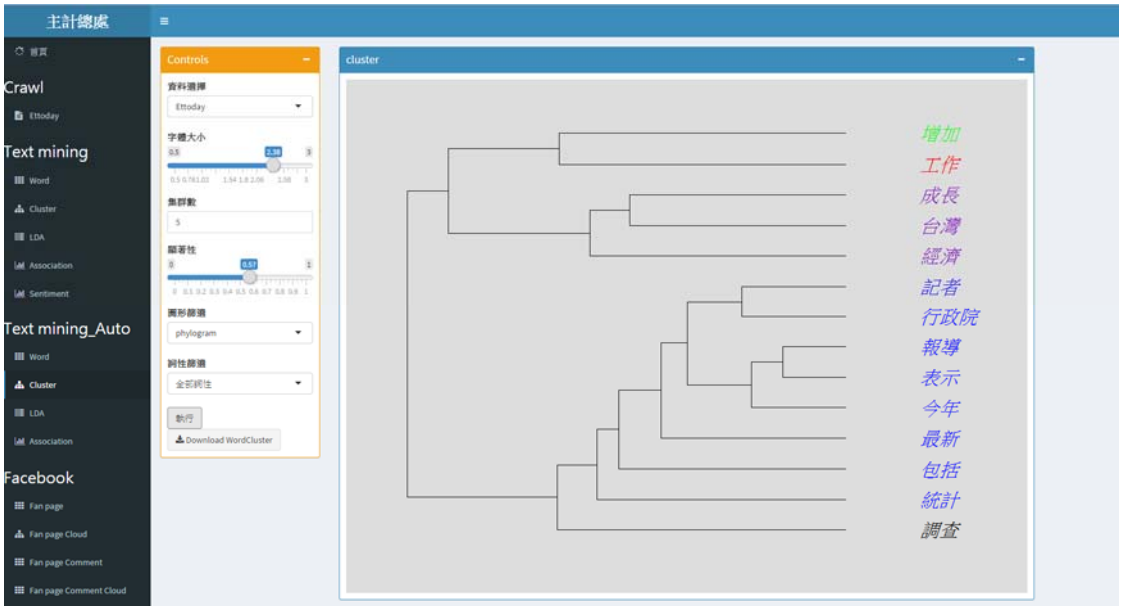


圖 79 客製化 Text Mining 平台-Sentiment 分析頁

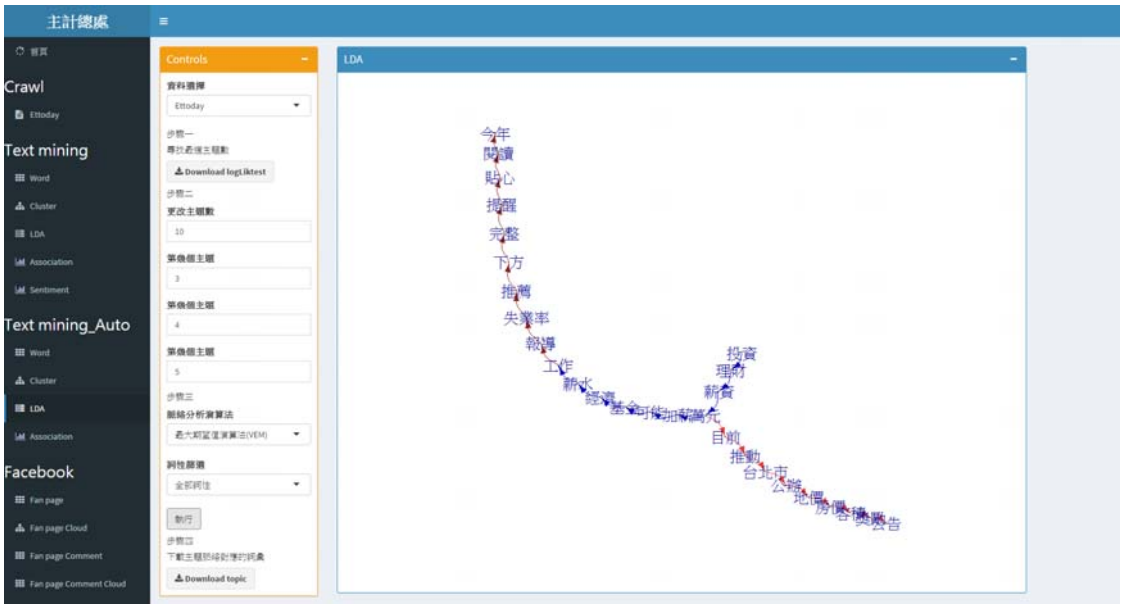


圖 80 客製化 Text Mining 平台-Sentiment 分析頁

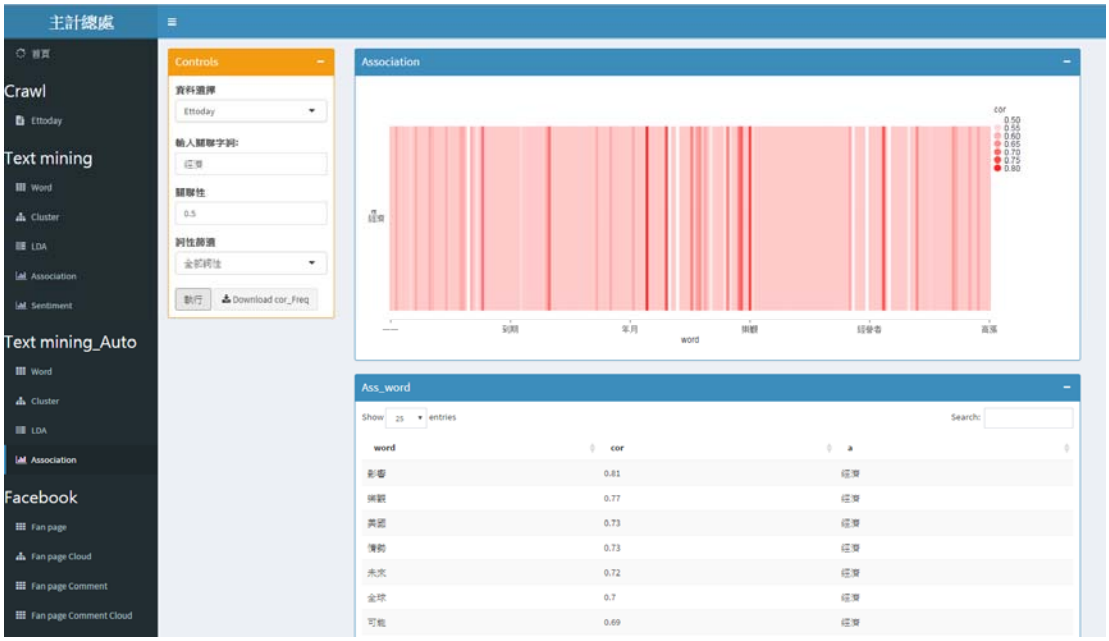


圖 81 客製化 Text Mining 平台-Sentiment 分析頁

四、Facebook（臉書）

可以分別對粉絲團小編的貼文以及每篇文章下粉絲的流言進行分析，進而探討對特定議題下，小編們及一般粉絲的想法是否有差異，更深入的瞭解民眾的想法，如圖 82 至 83。

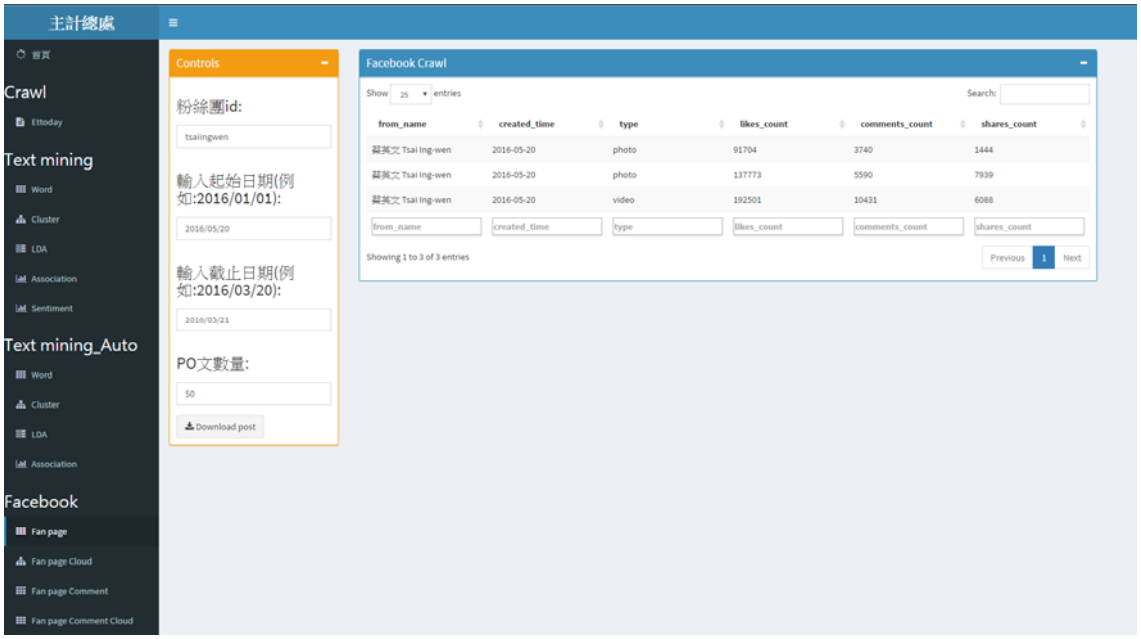


圖 82 客製化 Text Mining 平台-Sentiment 分析頁

圖 83 客製化 Text Mining 平台-Sentiment 分析頁

透過此平台可以快速獲取客觀的新聞稿文章以及主觀的民眾發言貼文，整合兩者不同面向的分析更能清楚的反應出媒體將消息傳給決策者時是否存在偏差，而決策者獲取資訊後有哪些不同的解讀，如此一來在民眾輿情的部份能夠掌握更多的徵兆。

第三節、客製化 Text Mining 平台應用效益

台灣民意調查的發展至今已有 50 年，各種民意調查的結果，透過媒體的宣傳，已貼近一般大眾的生活，甚至影響著大眾的各種決策行為。民意調查的方式也持續改良中，但仍不脫尋找對象並求其回饋資訊的方式。近年來文字探勘技術的成熟，開始可以大量擷取、分析非結構化資料，利用網路資源作為民意調查的原始資料，已成為可思考的方向。國內從 2012 年開始，Big Data 的話題持續發燒，許多應用也隨之而生，尤其這幾年 Facebook、Plurk 的興起，加上手機網路的盛行，使用者可以隨時隨地的發表自己的意見，試想，網路使用者將各種不同的意見資訊，主動放在網路的 Big Data 之中，而我們只要利用自動化的技術，就可以把這些非常大量的資訊，轉為我們想要的調查分析，這為文字探勘技術的應用，帶來了無限的可能性。而主計總處為瞭解國內外經濟情勢、產官學意見與輿情民意，更需要從社群網站、BBS、Web 頁面、新聞張貼、電子郵件訊息，以及預決算書與統計調查內文字說明等非結構化文字資料來源進行文字探勘，提供決策訊息解讀與分析。因此，若是可以客製化 Text Mining 平台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，主計總處即可以迅速有效掌握民眾輿情或文字資料意涵。

因此，透過本計畫針對主計總處需求，開發一套客製化的 Text Mining 平台，提供非結構化的文字探勘功能，其功能包括爬網（獲得的文本資料）、分析訊息得到的詞頻基本特質、建立詞頻關聯表與正負向情緒判別，並進行進行字詞的重要性分析、關聯性分析、集群分析、文章的正負向情緒分析、語意分析、脈絡分析…等。讓主計同仁可以簡易的使用文字探勘，瞭解並因應業務及政策需要，隨時應變需求，針對施政方針進行分析。例如，運用文字採勘進行提出預算計畫前後 100 天爬文進行輿情分析比較研究、爬文聽取更多民眾對於「經濟發展面」與「就業市場面」的需求、預算與

計畫端結合文數字對於計畫效能面加入追踪管考機制等。本平台操作簡易直覺，未來將可提供主計總處延續擴充功能並推廣至全國主計人員使用，創造更多資料應用價值效益。

第八章、成果摘要及推行建議

第一節、本計畫成果摘要

政府財務之處理，必須透過行政、主計、公庫及會計審核等組織，以達成相互配合及分立制衡之功能，而主計功能的發揮，更是促使國家有限資源合理分配及有效運用的重要關鍵。因此，若能整合歲計、會計、統計資料，運用大數據分析技術，選擇相關研究議題，建構示範案例及資料模型，產生對施政有參考價值之研究結果，增進主計業務相關政策規劃應用能力。本研究目的是期能藉由完成「公務檔案串連跨資料整合、「歲計資料整合應用」、「資料平台跨域整合」及「客製化 Text Mining」等 4 項關鍵核心議題的示範案例，經運用大數據分析技術，整合歲計、會計、統計資料，以及資料樣態特徵萃取後，建構出預測與資料加值應用模型之研究成果及後續研究建議。

在第二章文獻探討，本研究報告依序介紹大數據的發展歷程與定義、針對國外主計領域大數據分析研究現況與發展趨勢、大數據分析規劃及分析方法現況與發展趨勢、工具等進行評述，以作為後續推動主計資料大數據分析之借鏡參考，其中並簡要介紹了大數據分析應用案例現況，以作為本總處推動之參考。

於第三章研究設計，本研究計畫係「從大數據（Big Data）思維與角度出發，分析技術則以機器學習、資料採礦 CRISP-DM 及統計等演算法為本」的概念，嘗試利用資料映射（Data Mapping）技術將政府其他資料庫如：政府歲計會計資訊管理系統（GBA）、臺閩地區工商及服務業普查資料、全民健保資料檔資料、財政部營利事業資料、科技部產業創新調查資料...等，以及外部資料如：全國達康資料、全國意向資料...等，進行跨機關及跨領域資料串連，以利後續資料採礦分析之用。

經過期中專家交流座談會議與審查，本研究計畫參酌各方建議及整合文獻與實務作法，以「公務檔案串連跨資料整合、「歲計資料整合應用」、「資料平台跨域整合」及「客製化 Text Mining」等 4 項關鍵核心議題的示範案例發展模型雛型，包含了四大模組：（1） 村里常住人口推估模型；（2） 歲計估算預測模型；（3） 科技跨域整合資料庫模型；（4）客製化 Text Mining。並於後續第四章至第八章等四個章節當中，分別示範各模組之發展與應用程序。各模組之發展歷程簡述如下：

一、村里常住人口推估模型

隨著訪問調查日漸困難，若本計畫可以採用創新大數據應用方法，為戶口普查資料進行更細部的推估模型建立，可擴大普查效益。因此本案例採用內政部民國 105 年 7 月人口資料、主計總處民國 99 年人口普查資料與全國達康的民國 105 年 6 月常住人口資料，進行村里常住人口推估模型之建立。透過與普查年資料銜接，以函數映射方法估計到一定期間各縣市村里的人口常住資料，掌握調查時點最新人口數量及分布，藉由具有所需欄位的另一個相關資料庫，依其關連推估出目標資料庫所需欄位之值，以增進後續分析之可行性與價值。

首先，利用全國達康的民國 105 年 6 月的常住人口資料，計算出每個村里在其所在區中所佔的比例，再依此比例去加權內政部的民國 105 年 7 月的鄉鎮區人口資料，建立推估各區村里常住人口資料之模型。本研究並彙總上述成果，以視覺化分析工具設計互動分析圖表；本案例成果就政策效益而言可以提供小地區人口及分布狀況，以為政府規劃產業及均衡城鄉發展政策之參考。就企業經濟效益而言，可供廠商經營選址應用，提高企業經營成效。就統計效益而言，可提供學術團隊研究人口、社會變遷及城鄉發展所需。

二、歲計估算模型

我國歲計支出在法律義務支出占比近 7 成，歲出結構僵化導致政府財政支出未能隨景氣循環彈性調整改善。本示範案例是建立歲計預算模型可以即時並且充分反映政府整體的財政狀況，以及減少不必要的資源浪費之情形，除了預測之外，也可以用來監測政府預算，通過建立基本建模的數據庫來分析預算偏離的情況。

先以法務部人事費用途別資料為主，機關部分編號大致上可分為以下幾類，法務部本部、全台監獄單位、全台地方法院檢察署等，因有許多機關別，先以機關編號 23001000（法務部本部）當範例。分析主軸為以下兩點：1.找出預算-決算（借-貸）差異最大的用途別及其原因；2.利用決算之資料及指數平滑法進行預算的推估。經分析發現，比較二級用途別及三級用途別之預算決算差異結果相似，皆是以法定編制人員待遇及退休退職給付這兩個用途別為差異較大。此部分原因可能為職員調派以及在年度終了時預算員額與實際員額有落差，導致法定編制人員待遇預算編製過多，而退休退職給付因預算編製不足導致是常需要從其他用途別流用預算來彌補不足之缺額；由此可知，歲計估算與實際員額編列具有高度相關性，有礙於目前無法順利取得法務部實際員額詳細資料以致未能進行更精準的分析。

接著利用指數平滑法推估後所得之未來兩年預算之推估，為了與實際值符合，所以先以民國 96~ 103 年推估民國 104 年，利用民國 104 年的真實資料與推估出的值算出權重，並且對後來推估出來的民國 105、106 兩年度之值乘上此權數做修正。在導入預測模型後，善用並且發揮預算預測之功能，第一個是預測未來期間的預算，第二個是預測單位預算可能執行的結果。在預測未來期間的預算時，對未來各個財政目標進行分析和預測，實現中長期預測目標的長遠規劃，為預算規模和預算規劃提供決策的支持，也為預算編製提供參考之依據。在預測預算單位的可能執行結果方面，須

要根據各預算資金單位的執行情況及發展趨勢進行年度的比較，為預算調整和適時的糾正偏誤措施提供可行的依據。

三、科技跨域整合資料模型

本研究嘗試將過往執行的企業面的科技議題相關的調查，以資料映射（Data Papping）方式進行資料庫的串聯進行「科技跨域整合資料」。本研究採用第 12 次（民國 100 年）工商及服務業普查做為主資料庫，以及民國 100 年的第三次產業創新調查作為輔助資料庫，用函數映射方式將創新調查資料映射至工商普查的資料庫中，以擴增工商普查資料庫之應用價值及提升科技跨域資料之整合分析能力。

本次嘗試推估各縣市的研發經費，從輔助資料庫映射及推估出技術創新活動與花費及國內各縣市研發經費。首先利用工商普查之營業額、民國 100 年的各項技術創新活動的總花費佔營業額之百分比、公司內部的研發活動（R&D，包含 R&D 設備與建築的資本支出）花費百分比與委託其他公司或機構研發花費百分比來找出全國各公司的研發經費，並將全國研發經費依「民國 100 年全國各縣市之企業家數比例為權重」推估至各縣市研發經費。再以 RandomForest 隨機森林演算法建立模型，其正確預測率最高，且此演算法能夠針對類別變數及連續變數進行建模與預測。

科技跨域整合資料庫在未來可以繼續收集、儲存與管理國內主計總處與科技部各項調查所收集之企業研發與創新資料，未來提供多元資料整合供應與加值應用服務。

四、客製化 Text Mining

由於主計總處為瞭解國內外經濟情勢、產官學意見與輿情民意，常需從社群網站、BBS、Web 頁面、新聞張貼、電子郵件訊息，以及預決算書與統計調查內文字說明等非結構化文字資料來源進行文字探勘，提供決策

訊息解讀與分析。

本計畫則針對主計總處需求，開發一套客製化的 Text Mining 平台，提供非結構化的文字探勘功能，其功能包括爬網（獲得的文本資料）、分析訊息得到的詞頻基本特質、建立詞頻關聯表與正負向情緒判別，並進行字詞的重要性分析、關聯性分析、集群分析、文章的正負向情緒分析、語意分析、脈絡分析...等。讓主計同仁可以簡易的使用文字探勘，瞭解並因應業務及政策需要，隨時應變需求，針對施政方針進行分析。例如，運用文字採勘進行提出預算計畫前後 100 天爬文進行輿情分析比較研究、爬文聽取更多民眾對於「經濟發展面」與「就業市場面」的需求、預算與計畫端結合文數字對於計畫效能面加入追踪管考機制等。

第二節、推行挑戰與建議

為促進整合歲計、會計、統計資料，運用大數據分析技術，本研究計畫選擇相關研究議題，建構示範案例及資料模型，產生對施政有參考價值之研究結果，增進主計業務相關政策規劃應用能力。本研究已實地與主計人員訪談收集意見，以及透過相關資料庫發展四大應用範例，受限於時間與人力或仍未不足之處；但透過小規模之雛型開發與修正，將有利於可行方案評估與縮短後續系統開發時程。主計資料大數據分析之研究在主計資訊系統環境下之推行建議與未來發展方向，彙整說明如下：

一、村里常住人口推估模型

本次研究目前已進行到全村里的常住人口推估值，為期能創造更多加值應用，未來本研究團隊建議，可以再結合其他資料例如 Wifi 使用分佈，以產出更多細部資料欄位如宅戶數等，以使模型更為精緻，應用廣度更加多元；另一方面，為驗證推估資料品質，未來將朝與台電及自來水公司之開放資料結合，建立相互驗證機制方向研究，以提升資料品質與精確度。

二、歲計估算模型

由於人事費預決算金額與實際員額變動、考績晉級具有高度相關性，而 GBA 系統僅有人事費預決算金額資料，為使歲計估算模型更加精準預測，未來研究建議以主計總處為例，整合主計總處相關人事或薪資系統資料，分析瞭解其關係變化，並找出人事費預算經費樣態（pattern），強化模型預測之精準度。

三、科技跨域整合資料模型

如本研究中因為所得到的工商普查資料所釋出的第一類資料檔中，並沒有包括企業所在縣市之變數，故無法實際從映射後的資料估算各縣市的研發經費，僅能使用企業家數比例的次級統計資料，來推估各縣市的研發經費。又如本研究原本預計從其他政府單位取得研發相關資料以驗證本研究映射資料之成效，但受限於該單位因個資法問題，無法提供原始資料及攜出以供研究使用，使得本研究在執行上多所困難。故本研究受取得資料的細緻及不同資料庫本身架構及特性的差異，真正在資料映射上的驗證問題，只能以各項欲映射的次級統計資料進行趨勢的驗證。

對於未來研究上，建議資料的豐富性及細緻度仍是影響本研究成效最重要的關鍵因素。同屬於政府單位的不同部門，若能攜手提供更完整且詳實的資料，才能創造共贏的科技跨域資料整合的期待，免除各單位各自受資料短缺之苦，使資料價值完整發揮。

另外，本研究團隊建議，可增加資料驗證與提升分析資料品質議題（含探討至每個資料處理步驟的信度與效度）、非結構資料及去識別化資料之跨域整合議題，以使資料品質獲得確保，以及跨域整合應用更加彈性多元。

四、客製化 Text Mining

由於本計畫之客製化 Text Mining 平台，是就現有主計總處需求，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；未來本研究團隊建議，未來主計總處可延續平台模組擴充功能並推廣至全國主計人員使用，創造更多資料應用價值效益。

參考文獻

一、中文文獻

1. Anderson, Chris. (2013)。自造者時代：啟動人人製造的第三次工業革命。(連育德, 譯者) 天下遠見出版。
2. Barabási, Albert-László. (2013)。爆發：大數據時代預見未來的新思維.。北京: 中國人民大學出版社。
3. MacCormick,John. (2014)。改變世界的九大演算法：讓今日電腦無所不能的最強概念。經濟新潮社。
4. Kahneman, Daniel. (2012)。快思慢想 Thinking, Fast and Slow。天下文化。
5. Mayer-Schonberger,Viktor., & Cukier, Kenneth. (2013)。大數據. (林俊宏, 譯者)。台北市: 天下文化。
6. Nate, Silver.(2013)。精準預測：如何從巨量雜訊中，看出重要的訊息？台北市: 三采文化。
7. Steiner, Christopher(2014)。演算法統治世界 Automate this：how algorithms came to rule our world。
8. 胡世忠. (2013)。雲端時代的殺手級應用~海量資料分析。台北市: 天下雜誌出版。
9. 涂子沛(2013)。大數據：數據革命如何改變政府、商業與我們的生活。香港: 中和出版。.
10. 譚磊(2013)。New Internet。大數據採礦。北京: 電子工業出版社。
11. 工商時報(2013.11.28)。海量資料 助零售業決策。工商時報。
12. 甘芝萁(2014.2.19)。賽門鐵克 2014 預測：社群媒體仍是犯罪寵兒。自由時報。
13. 李孟熹(1985)。現代商業經營要訣。作者自行出版。
14. 吳展旭(2004)。台灣綜合零售業治理組織型態選擇之研究。國立中山大學經濟學研究所碩士論文，高雄市。
15. 吳凱琳(2012.2.13)。紐時：資料分析人才，未來最搶手。天下雜誌。
16. 呂愛麗(2014.3.26)。超級電腦「華生」再升級 準備幫你看病！財訊雜誌，447。

17. 林俊宏譯(2013)。大數據。台北市：遠見天下文化出版股份有限公司。
18. 林麗娟(1997)。質化研究取向與讀者資訊尋求。資訊傳播與圖書館學。4(1)，52-59。
19. 胡世忠(2013)。雲端時代的殺手級應用：海量資料分析。台北市：天下雜誌股份有限公司。
20. 胡幼慧(1996)。質性研究：理論、方法及本土女性研究實例。台北市：巨流圖書公司。
21. 康文柔(2013.11.12)。光棍節瘋網購 交易額 1700 億。中國時報。
22. 教育部電子報(2012.12.6)。美國大學因應企業需求紛設「海量資料分析」碩士學程。541 期。
23. 張偉翰(2011)。探討物聯網議題下 RFID 技術發展與未來應用趨勢。國立交通大學科技管理研究所碩士論文，新竹市。
24. 張慈映(2010)。2010 醫療電子產業趨勢。工業技術研究院。
25. 許蒨棱(2012)。Big Data：顧客需求導向經營策略的驅動力。經濟部商業發展研究院，未出版。
26. 梁嘉勝(2013)。以 Hadoop 為平台－結合異質資料庫與 Hive 之加速查詢應用。國立東華大學資訊工程學系碩士在職專班碩士論文，花蓮縣。
27. 黃亦筠(2012.4.18)。巨浪來襲 政府、企業的下一場戰爭。天下雜誌，495，50-55。
28. 黃俊龍(2012)。有關 Big Data 的二三事。NCP Newsletter，42，9。
29. 黃瑞琴(2001)。質的教育研究法。台北市：心理出版社。
30. 葉至誠(2000)。社會科學理論。台北縣：揚智文化事業股份有限公司。
31. 趙郁竹(2012.3)。萃取大數據 IT 方案再進化。數位時代電子雜誌，214，84。
32. 劉麗惠(2013)。大數據時代全面來臨 發展新商業模式 掌握先機。貿易雜誌，268，60-61。
33. 鍾倫納(1992)。應用社會科學研究法。台北市：台灣商務印書館。
34. 鍾張涵(2013.10.15)。大數據商機 台灣準備好了嗎。聯合晚報。
35. 韓培爾(1998)。社會科學方法論：量化與質化 Q＆A。台北市：風雲論

壇出版社。
36. 嚴勻希(2013)。大數據與雲端儲存之專利佈局與研發方向分析。國立台灣科技大學專利研究所碩士學位論文，台北市。
37. Ho, A. (2010)。新版個資法即將上路，NOPAM Themis 為您的電郵資安把關。綠色運算股份有限公司。
38. IBM (2012.10)。藍色觀點 BLUE VIEWPOINT。台灣國際商業機器股份有限公司，44，8-9。
39. 邱莉燕、鄭婷方(2013.1.7)。看見未來 5 分鐘 Big Data 大數據正在改變生活・創造新生意。遠見雜誌，319。
40. 時雪峰、陳萍秀、劉豔磊(2005)。科技文獻資訊檢索與利用(2 版)。清華大學出版社。
41. 曾仁凱、黃國蓉(2012.10.3)。趨勢科技張明正：海量資料 科技界未來黑金。經濟日報。

二、外文文獻

1. Autor, David. (2014). Polanyi's Paradox and the Shape of Employment Growth. Re-Evaluating Labor Market Dynamics.
2. Brat, Eric., Clark, Paul., Mehrotra, Pranay., Stange, Astrid., & Boyer-chammard,Celine. (2014). Bringing Big Data to Life_Four Opportunities for Insurers. BCG.
3. Brat, Eric., Heydorn, Stephan., Stover, Matthew., & Ziegler, Martin. (2013). Big Data：The Next Big Thing for Insurers? BCG.
4. Byrne, J., & Pattavina, A. (2006, 9). Assessing the Role of Clinical and Actuarial Risk Assessment in an Evidence-Based Community Corrections System: Issues to Consider. Federal Probation.
5. Chassaing, Thierry. , DiGrande,Sebastian., Field,Dominic,, & Rose,John. (2013). Delivering Digital Satisfaction: U.S. Consumers Raise the Ante. BCG.
6. Davenport, Thomas. (2012). Data Scientist The Sexiest Job of the 21st Century. Harvard Business Review.
7. Davenport, Thomas. (2013 年 12 月). Analytics 3.0. Harvard Business Review, 頁 66-67.
8. Duhigg, C. (2012. February 19). How Companies Learn Your Secret. The New York Times, MM30.
9. European Technology Platform on Smart Systems Integration (2008). Internet of Things in 2020 – ROADMAP FOR THE FUTURE. Ontario, Canada: Continental Automated Buildings Association.
10. Kelly, J. (2014). Big Data Vendor Revenue and Market Forecast 2013-2017. Marlborough, MA: Wikibon.
11. Lohr, S. (2012. February 12). The Age of Big Data. The New York Times, SR1.
12. Page, L. (2012. April). 2012 Update from the CEO. Google.
13. Patton, M. Q. (1990). Qualitative Evaluation and Research Methods.

Newbury Park: Sage publications.
14. Russom, P. (2012. January 23). Big Data Analytics: 2012 New Year's Predictions. Renton, WA: The Data Warehousing Institute.
15. Scherer, M. (2012. November 7). Inside the Secret World of the Data Crunchers Who Helped Obama Win. New York: TIME.
16. The Economist intelligence Unit (2013). The hype and the hope: The road to big data adoption in Asia-Pacific. USA: Hitachi Data Systems.
17. The World Economic Forum (2012). Big Data, Big Impact: New Possibilities for International Development.Geneva Switzerland: The World Economic Forum.
18. Bryant, R. E., Katz, R. H., & Lazowska, E. D. (2008). Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. Washington, DC: Computing Community Consortium.
19. Davenport, T. H., Barth, P., & Bean, R. (2012. July 30). How 'Big Data' Is Different. Cambridge, MA: MIT Sloan Management Review.
20. Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. San Francisco, CA: Operating Systems Design and Implementation.
21. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011. June). Big data: The next frontier for innovation, competition, and productivity. New York: McKinsey & Company.
22. Mayer-Schonberger, V., & Cukier, K. (2013). BIG DATA: A Revolution That Will Transform How We Live, Work, and Think. New York: Houghton Mifflin Harcourt.
23. McAfee, A., & Brynjolfsson, E. (2012. October). Big Data: The Management Revolution. Harvard Business Review, 62-63.
24. Mills, P. M., & Ottino, J. M. (2012. January 30). The Coming Tech－led Boom. New York: The Wall Street Journal.

三、網路部分

1. http://asmarterplanet.com/blog/2013/12/ibms-5-5-future-computers-will-learn.html，Hamm, S.，IBM。
2. http://blogs.hbr.org/2013/07/five-roles-you-need-on-your-bi/ ， Ariker, M., McGuire, T., & Perry, J.，Harvard Business Review。
3. http://mashable.com/2012/10/12/big-data-infographic ， Al-Greene, B. ，Mashable Business。
4. http://newsroom.fb.com/company-info/，FACEBOOK。
5. http://wired.tw/2013/07/18/big_data_2/index.html，Leonard-chien。
6. http://www-01.ibm.com/software/tw/data/bigdata/，IBM。
7. http://www.alcatel-lucent.com，阿爾卡特朗訊。
8. http://www.apache.org，Apaches Software Foundation。
9. http://www.bnext.com.tw/article/view/id/30561，陳芷鈴，數位時代。
10. http://www.brickmeetsclick.com，Brick Meets Click 公司。
11. http://www.businessweek.com/articles/2012-10-18/big-dairy-enters-the-era-of-big-data，Grobart, S.，Bloomberg Businessweek。
12. http://www.ctimes.com.tw/DispNews-tw.asp?O=HJY2KAYKK0QSAA00NA，蔡維駿，CTIMES。
13. http://www.dgbas.gov.tw，行政院主計總處。
14. http://www.digitimes.com.tw/tw/cloud/shwnws.asp?cnlid=16&cat=10&cat1=10&id=0000354595_R956NDK88TYYR86XLHH56 ， 李 茂 誠 ，DIGITIMES。
15. http://www.epochtimes.com/b5/14/4/5/n4123873.htm，徐傑，大紀元。

16. http://www.forbes.com/，Forbes。

17. http://www.green-computing.com，綠色運算股份有限公司。

18. http://www.ibm.com/news/tw/zh/2013/07/09/v534639i06471f97.html ，IBM。

19. http://www.idc.com，IDC。

20. http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx，國際電信聯盟(International Telecommunication Union)。

21. http://www.mckinsey.com，McKinsey & Company。

22. http://www.mckinsey.com/insights/marketing_sales/competing_through_data_three_experts_offer_their_game_plans，McKinsey & Company。

23. http://www.moneydj.com/KMDJ/News/NewsViewer.aspx?a=abd850f8-c7d9-4bc7-bf61-9ffd3154c253，鄭盈芷，MoneyDJ 理財網。

24. http://www.moneydj.com/topics/bigdata/，MoneyDJ 理財網。

25. http://www.ncc.gov.tw，國家通訊傳播委員會。

26. http://www.nytimes.com，The New York Times。

27. http://www.runpc.com.tw/content/content.aspx?id=109816 ，陳兆寧，RUN!PC。

28. http://www.runpc.com.tw/news.aspx?id=101133，施鑫澤，RUN!PC。

29. http://www.sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm#toc，美國證券交易委員會。

30. http://www.splunk.com，Splunk。

31. http://www.storageforum.tw/2014/01/blog-post.html，台灣企業儲存技術論壇。

32. http://www.taiwannews.com.tw/etn/news_content.php?id=2362902 ， 鍾 榮 峰，中央社。

33. http://www.time.com/，TIME。

34. http://yowureport.com/?p=10542，林宗立，YOWU REPORT。

35. http://www.stat.gov.tw/ct.asp?xItem=1135&ctNode=525&mp=4，中華民國統計資訊網。

36. http://www.ettoday.net/news_search/doSearch.php?keywords=%E4%B8%BB%E8%A8%88%E8%99%95&x=0&y=0，Ettoday 新聞雲。

附錄一、創新、就業、分配—應用大數據，培育新世代主計 3.0 人才研究成果分享發表會


大數據時代-主計3.0人才培育座談會
臺北醫學大學管理學院暨大數據研究中心
行政院主計總處

附錄二、主計資訊應用研討會研究成果分享發表會