論述 》統計 · 調查

交通統計大數據首次應用分析省思

大數據（巨量資料）為近年熱門議題，國內外各界成功應用案例如雨後春筍般出現。交通部統計處甫於去（103）年以國道高速公路計程收費資料，完成首次交通統計大數據分析。本文主要藉此實作經驗，省思政府統計大數據分析效益與關鍵因素，俾為日後精進之參考。

饒志堅[1]（交通部統計處副處長）

壹、前言

行政院毛院長上任後不久，媒體出現這段標題：「毛治國祭出科技三箭：開放資料（Open Data）、大數據（Big Data）、群眾外包（Crowd Sourcing）」，所以如何有效運用愈來愈多的資料，做成有用的決策，已經成爲民間及政府一起努力的目標。

其實，早在院長擔任交通部長時期即非常重視 Open Data 與 Big Data 之應用，交通部統計處 4 年前即撰文「公開政府資料以配合快速創新的腳步」（主計月刊第 666 期），呼籲加速公開政府資料，並率先開放 9 項去識別化之統計調查「原始資料」（Raw Data，非統計表），於政府資料開放平台（Data.gov.tw）供各界自由下載應用，以實際作爲回應全球趨勢。

去（103）年國道高速公路由計次改爲計程收費，交通部統計處旋即將每月超過 5 億筆（35GB）之資料，利用中華電信雲端伺服機，以 Hadoop 分散式資料平行處理技術，完成國道塞車型態、服務區旅次特性，及國道易飆車路段三篇分析，提出具體建議供業務單位參考應用，並分別於交通部、主計總處、交通大學、新北市政府等單位之研討會中發表

（內容參見主計月刊第 710、714 期），跨出政府統計交通大數據分析的第一步。

經過這一段期間的實務運作，面對愈來愈夯的大數據分析議題，聽了許多專家的演講，心中沉澱累積一些對政府（交通）統計的看法，爰以「交通統計大數據首次應用分析省思」爲題，與大家分享。

貳、何謂大數據

資料量多大才叫大數據？前述高速公路計程收費每月 35GB，一年約 0.4TB 的資料就可稱大數據嗎？從 2001 年美國研究機構率先提出大數據 3V 特性，即資料數量大（Volume）、產生速度快（Velocity），與複雜多樣性（Variety）後，迄今未見公認的 3V 門檻值。其實這不難理解，隨 IT 技術日新月異，就算是 PC 的處理能力也愈來愈強，早年的「大數據」，現在看來可能已不算甚麼！所以維基百科定義大數據時，僅以「涉及的資料量規模巨大到無法透過人工，在合理時間內達到擷取、管理、處理、並整理成爲人類所能解讀的資訊」概念式表達，所以它的「數量大」、「產生快」與「成分雜」都是相對的，沒有絕對標準。

回到高速公路計程收費的例子，一個月 35GB 的資料量已無法在一般 PC 運算處理（拉不動），就算勉強可以也耗時甚鉅，必需以多個伺服器平行處理，才能在「合理時間內 ... 整理成爲人類所能解讀的資訊」，以現時處理能力來看，堪可稱「大」。次就產生速度而言，平均每分鐘超過 1 萬筆的資料被記錄，日積月累、持續不斷，勉可稱「快」。惟最後就記錄項目觀之，除年月日、時分秒外，僅包括車型及通過之門架編號，屬傳統結構化資料，並不複雜；但未來若結合其他資料來源，如沿路天候狀況（降雨量）、車禍地點（A1 及 A2 類交通事故）、施工路段時間、匝道儀控及交通管制資料，甚至 Google 搜尋景點熱門度即時變化統計，增加數據廣度，將有助預測塞車路段之精準度，提升應用價值。這也是後來有人提出新的大數據 4V 特性，也就是在原有 3V 上面，再加上「有價值（Value）」一項，因爲如果沒有應用價值，分析再大、再快、再雜的資料也是毫無意義的。

綜此，所從事的數據分析究竟是否屬於「大數據」分析，一點都不重要。重要的是它有沒有應用價值；如果有價值，就算只是在 PC 上跑出來的結果，一樣值得讚許；如果沒有價值，千萬不要死抱大數據不放，徒作虛工。話雖如此，不可否認的，具 4V 特性的大數據分析的確帶給我們前所未有的重要發現與應用，在國內外工、商業界，這類案例已多如牛毛，無庸贅述。這也提醒我們，政府統計重視並發展大數據分析有其必要性。

論述 》統計 · 調查

參、政府統計大數據分析之效益

爲什麼說政府統計應該要作大數據分析？從國道高速公路計程資料研究中，發現大數據分析有幾項優點。

一、大數據分析可補抽樣調查之不足

由於大數據是將全部資料納入分析，鉅細靡遺，不會將稀少性但有意義之個案排除。例如，在高速公路易飆車路段研究中發現，少部分車輛飆車時速超過 150 公里，雖輛數占全部不到千分之一，卻有其規律性。若以傳統抽樣調查分析，此類稀少性樣本甚難被抽中，就算眞被抽中，也要視爲特異值排除在外，因爲若不排除，當反推母體（乘以抽樣係數）後，就會被扭曲高估。惟此類樣本若被拿掉，就看不出易飆車路段、時段等特性，無法做出相對應防治措施，減少資料應用價值。

類似情形在政府許多統計調查中也有可能發生。例如，以衡量國人所得差距倍數之家庭收支調查而言，很難抽到所得極高的金字塔尖端人士（因爲機率關係），就算抽到也很難保證受查戶願意被調查（因爲隱私權關係）；更重要的是，就算他願意接受調查，並且據實填報資料，政府也不敢用！因爲如果採用，稀少性樣本被抽樣係數擴大後，當年的所得差距倍數會被不當高估，無法解釋。次年，很可能因沒抽到稀少性樣本，所得差距倍數掉下來，又會被社會大眾抨擊。雖然按統計學理，政府將特異值排除是合理正確的作法，但長期將這些金字塔頂端的高所得家庭排除在外，難免有低估所得差距之嫌。過去或因大數據分析技術尚不成熟，或資料來源所限，不得不如此；但展望未來，政府統計應有更多大數據思維，納入相關去識別化的財稅資料，改進作業方式，以呈現更眞實的所得分配狀況。

二、大數據分析可細緻化政策應用層面

大數據的特點就是量大，量大的好處是可作非常細緻的分類，且分類後仍然有足夠的個案數作分析。例如，在高速公路易塞車路段研究中，可將全臺灣地區國道 300 多個匝道（起訖點）作任意交叉分類，再按不同時段分，至少有 200 萬種排列組合，據此了解平日、假日、尖峰、離峰、長短程、不同地區之塞車型態，擬訂個別因應對策。

若以傳統公務統計來看，通常是公布大分類的平均值，但這對政策應用非常不方便，有時甚至會產生誤導，因爲平均數將兩邊極端值均化，看不出眞正的變化。若要產製細緻化公務統計報表也不太可能，因爲每月多達數萬張報表。就算產製，也無法進行分析，因爲報表數量太多，

看得眼花撩亂，審視曠日廢時，其中需要合併運算時，繁雜程度更是難以想像！但這些透過大數據分析，都不再是問題：藉由平行運算（Parallel Computing）技術，解決資料量大拉不動的問題；藉由商業智慧（Business Intelligence）的視覺化圖表下鑽功能，解決了初步資料過濾（挑圖）的問題;藉由資料探勘（Data Mining）技術，可建立模型，強化分析深度。大數據資料在手中，無論直看、橫看、斜看、倒著看，愛怎麼看，就怎麼看，如此才能找出問題核心，而這是傳統公務統計難以做到的事情。

統計之目的不在於製造數字，而在於「檢驗政策、支援決策」，面對多元化社會，細緻化政策應用是必然的方向，善用大數據對政府統計功能之發揮，絕對有幫助。

三、善用大數據可簡化相關統計調查減少擾民

大數據資料通常是由於管理之需要而生，例如國道計程收費資料、公車及遊覽車之GPS定位資料、交通各業營收資料...。這些資料雖非因統計目的建置，但政府統計若能善用之，則可取代部分調查項目，簡化調查表式，撙節相關支出，如此也降低擾民程度，一舉兩得。美國麻省理工學院曾利用網路商品售價巨量資料編製「每日網上價格指數（Daily Online Price Index）」，將之與「消費者物價指數（CPI）」比較，發現前者不僅編製成本低，指標領先敏感度更優於後者，殊值政府統計工作者參考。在強調Open Data的時代，如何蒐集並運用大數據，是政府統計應積極努力的方向。

肆、政府統計大數據分析之關鍵

政府大數據分析已成爲施政趨勢，國外已有許多成功的案例，國內政府統計在這方面尚有很大進步空間。要做好政府統計大數據分析，實務上提出三方面供參考。

一、與資訊人員充分溝通，取得資安信任

大數據分析作業流程包含資料的取得、建置、分析、討論、成果發布，其間涉及資訊人員、統計人員、業務單位，甚至外部社會公眾彼此的合作與互動，絕非單純購買軟硬體與資料分析而已。其中資料的建置方式與權限管理，在日益講求隱私保護與資訊安全的時代，更是重要。從事大數據分析前應先考量本身資訊處理能力可否勝任？系統架構是否與所在單位資訊系統融合？資安控管是否已考量無虞？這些先與資訊人員充分溝通協調，取得資安信任，有助後續作業順利進行。

二、與業務單位密切合作，共同解析大數據

做好政府統計大數

論述 》統計 · 調查

據分析，需要三方面的配合：資訊技術（Information technology）、統計方法（Statistic methodology），與專業知識（Domain knowledge）。對政府統計人員而言，第 1 項略有涉獵，要充實較不困難（當然需要經費支援）；第 2 項較爲熟悉，但仍需精進；第 3 項最爲缺乏，應充分與業務單位密切合作，共同解析大數據，方能提升分析效率，避免閉門造車。交通部統計處與部裡所屬機關（如運研所、港務公司、高公局等）平日即合作密切，未來將更積極共同進行大數據分析。

三、爭取長官重視，積極投入培養分析人才

大數據分析對政府統計而言是相對陌生的領域，如何實際操作，仍處摸索階段。但由於政府統計人員多已具備基礎分析能力，要再深入精進分析技術（如資料探勘），並不是太困難的事。人才是業務成功推動的關鍵，除做中學、學中做，自我訓練成長，以實際成果贏得長官重視外，爲加速提升大數據分析質量，實應投入更多資源積極培養優秀分析人才，充分發揮大數據功能，成爲政府施政利器。

伍、結語

最近幾年，大數據突然成爲社會顯學，報章雜誌媒體爭相報導，氣勢越炒越熱。當然我們也知道，大數據分析不是萬能的，成功案例固多，失敗（無意義）的數據分析也是不乏其數。但當前這股對大數據的熱衷，至少喚醒大家對「數據分析」的重視，講求用數字科學來解決問題。就身爲政府統計工作者的角度來看，除樂見其成，更也自我惕勵，盼望未來政府統計在大數據分析上出現更多成功案例，利國利民。

註釋

1. 本文文責由作者自負，不代表服務機關意見。

參考文獻

1. 張嘉玲（2014.12.23），毛治國祭出科技三箭：開放資料、大數據、群眾外包；數位時代，http://www.bnext.com.tw/article/view/id/34796。
2. 饒志堅、伍家志、張富凱（2011），公開政府資料以配合快速創新的腳步，主計月刊，666，80-87。
3. 郭昌儒（2015），探勘交通統計大數據（Big Data）－高速公路易壅塞路段概況分析，主計月刊，710，78-85。
4. 郭昌儒（2015），首創巨量分析技術（Hadoop）探勘交通大數據，主計月刊，714，95-98。
5. 維基百科，大數據，http://zh.wikipedia.org/wiki/%E5%A4%A7%E6%95%B8%E6%93%9A#cite_note-17 。
6. 姜澍（2014.04），大數據時代下的政府統計，http://www.stats.gov.cn:82/tjzs/tjsj/tjcb/dysj/201405/t20140506_549561.html。❖