論述 》統計 · 調查

精進技術發展爲資料創新鋪平道路

政府統計的創新發展，統計技術一直扮演重要角色，資料革命所形成資料生產範式之變革，促進資料創新的契機，傳統統計技術適用性受到衝擊，國際間關注相關技術的研發及應用，重視創新思維，以因應數位時代的快速發展，而能創造政府統計新局。

羅國華（行政院主計總處國勢普查處研究委員）

壹、前言

政府統計機構長期以來作爲成功的資料供應者，深受社會大眾的重視及期許，近年來隨資料革命而形成資料來源及類型的益趨多元複雜，以及各界對及時、準確與可信之統計資料需求的不斷增長，促進政府統計創新思維，各種新技術呈現前所未有的蓬勃發展，相關成果擴增統計應用範疇，亦對傳統資料生成範式及作業處理技能產生影響，所遵循的政府統計規範受到檢驗，國際間關注政府統計資料創新，持續推動各項相關技術之研發及應用，成爲政府統計發展重要的挑戰。

貳、國際間促進資料創新概況

面對益趨複雜而艱困的統計生成環境，所須技術層次勝於以往，涵蓋的知識領域更爲廣泛，主要國際統計組織相繼成立工作小組，協同合作積極進行相關的研究及推展，例如：聯合國歐洲經濟委員會（UNECE）的現代化小組、聯合國統計委員會（UNSC）的大數據全球工作組等，建立虛擬網路協作平台，分享知識及經驗，促進交流活動，並經由沙盒（Sandbox）試作[1]，逐步建立可行之通用作業模式。另有些國家統計局爲能掌握契機而進行組織調整，以能善用新技術提升作業效能，進而強化

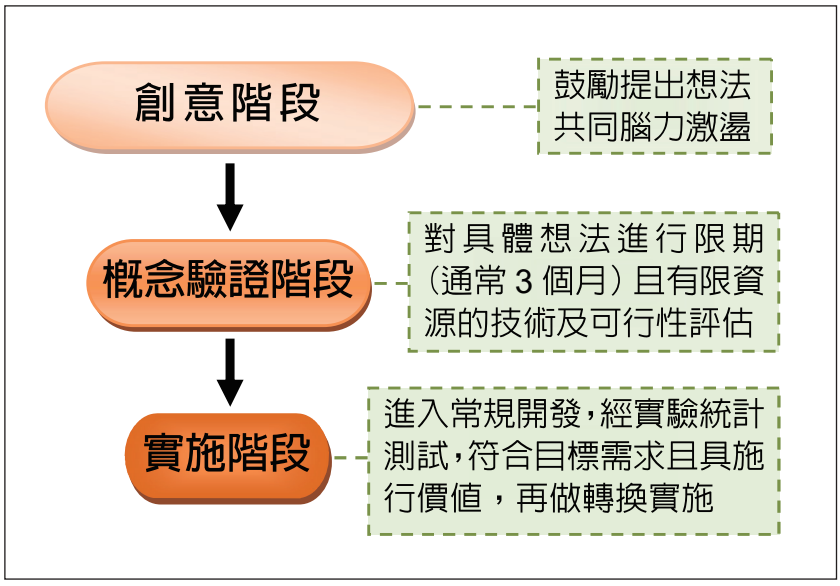
政府統計。

一、荷蘭

荷蘭統計局於 2012 年最早成立創新實驗室（Innovation Lab），此概念隨後亦爲加拿大、英國與瑞典等國家統計局複製。實驗室具備良好的電腦作業環境，鼓勵人員不拘於理論或實踐，共同進行腦力激盪。以三階段漏斗法（圖 1），從粗略想法的充實而形成清晰的項目提交，再進行限時且資源有限的概念驗證（Proof of concept，POC），成功的 POC 才會開發稱爲「測試產品」的實驗統計，經檢查確認符合目標需求，且具有施行價值，再轉換爲成熟的政府統計。並不期望全部想法及試作都能成功，每階段約有一半項目因不適合而停止，有些項目則會與外部單位合作發展。初期由該局人員兼任，逐步累積解決問題的經驗，過程曾遭遇技術、方法、法規及人才等方面的挑戰。

鑑於次級資料源（登記資料、大數據 / 非機率樣本、多源資料等）在政府統計之價值日益重要，該局於 2016 年成立大數據統計中心（Center for Big Data Statistics），團隊由多類資料科學專家組成，其使命是基於新方法，創建及改進政府統計資料，並不限於大數據的應用。推動上建立外部合作夥伴的協作及專業交流，加強局內部門的支持及參與，積極培育資料科學人才，並對文字探勘（Text Mining）、機器學習、深度學習（Deep Learning）、人工智慧（AI）、合成資料及可視化等技術主題進行前沿（Cutting-edge）研究，數年來已有超過 30 項的實驗產品，進行資料源的穩定性檢查、驗證方法、操作需求評估等，須保證結果是可靠且有價值，才會在政府統計中實施。

圖 1　三階段漏斗法



資料來源：作者自行繪製。

二、英國

英國統計局於 2017 年成立資料科學園區（Data Science Campus），核心人員爲資料科

學專業人士，其設立目的在探尋新資料源的使用，開發新的工具及技術，並發展資料科學能力，以建立新的理解並改善公共利益的決策。

研發流程是從構想或發現階段，到交付有效解決方案階段，再進入移交試作階段，終至完成及生產階段。該園區已完成 20 個研究項目，涵蓋經濟、運輸、金融、能源、健康、圖像、合成資料等方面，運用的技術及工具，包括機器學習、自然語言處理、電腦視覺、地理空間、合成方法等。另爲提升人員在資料科學素養能力及經驗的延續，積極進行各級人才培育計畫，並加強與學術界、工商業界及國際組織的合作夥伴關係，以擴增影響力，促進更多的公共利益。

三、瑞士

瑞士統計局於 2015 年設置工作小組，負責大數據主題；2017 年成立臨時的資料科學實驗室，並啓動 5 個試驗項目；2021 年正式創建資料科學能力中心（Data Science Competence Center），其目的爲培育人員知識與技能，加強多領域合作夥伴關係，並應用適當的方法、技術和實踐，以解決複雜、非結構化及資料豐富的問題，而能增進公共利益。

該局爲促進資料利用，另開發協作平台供參與單位共用，並對試驗項目報告從文件、目標、附加價值、驗證品質及符合該局品質標準等 5 個方面進行評估，爲移轉生產而做準備。試驗項目側重於應用新的統計技術及機器學習方法，包括文字探勘、航空影像識別、分類編碼自動化、小區域估計等。

參、資料創新中的技術發展

資料革命使資料成爲用之不竭的資產，傳統基於機率樣本及推論變量關係的統計技術，對於次級資料源的新型態資料，受到適用性的侷限，政府統計亟須創新思維，善用新技術開發新資料源、整合多源資料集及改進資料生成方式，以建構靈活的資料生產範式，擴增資料價值。關於新技術發展涵蓋多方面領域而具跨學科知識，其所運用的演算法及統計方式，雖有異於傳統政府統計方法之處，惟仍應適當遵循聯合國「政府統計基本原則」，這是發展上須重視及面對的課題。

一、機器學習（Machine Learning）

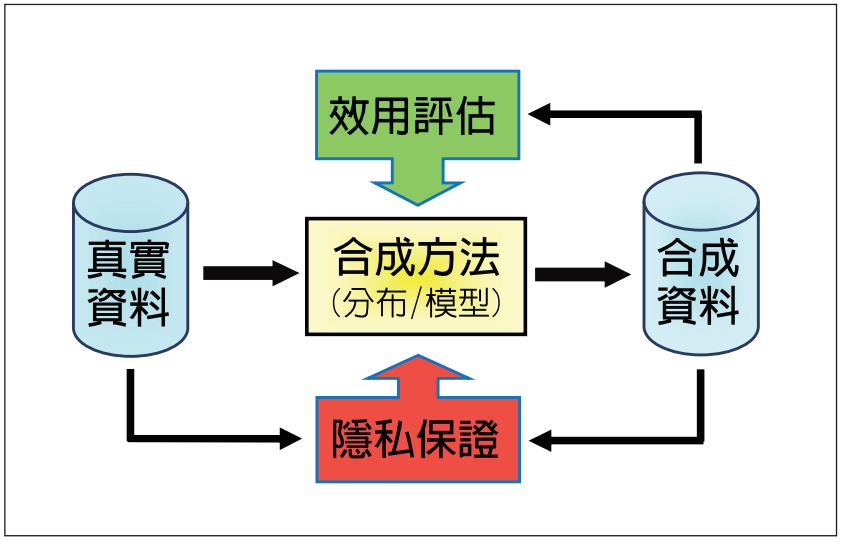
機器學習是利用電腦透過演算法，可自動分析資料中的變動模式及隱藏訊息，利用訓練資料學習以獲得其中之規律，進而評估適當預測模型，並用測試資料驗證其準確性。本質上，機器學習具有分析複雜資料的能力，對異常值及錯誤資料的敏感性較低，無須過多的程式撰寫，通過自動化程序，當資料量大時更能創建穩健可靠的模型，提高資料生成

效率及預測結果，在多種類型資料處理上具有很大潛力，為先進國家統計局逐漸重視的新技術。

傳統統計模型重視變量間的關係程度及其推論解釋性，通過信賴區間、顯著性檢驗或其他檢驗方式，以評估模型的合理性及誤差。機器學習的演算法中雖涉及許多統計學概念，但對輸入資料與預測結果之關係，未做嚴格的假設及解釋，而其過程的複雜性常被認為是令人難以理解的黑匣子，受到透明度的質疑；至其有別於傳統的方法，在統計品質上，對聯合國「政府統計品質保證架構」的適用性仍須檢驗。鑑於透明度及品質保證是政府統計的核心價值，UNECE 現代化小組於 2020 年底提出相關研究結果，建議結合原有架構加強考慮，從可解釋性、準確性、及時性、再現性及成本效益等 5 個面向補充定義，並加以詮釋說明，以能完整涵蓋並作為依循。

二、合成資料（Synthetic Data）

圖 2　合成資料生成系統



資料來源：作者自行繪製。

合成資料具有多種用途，對政府統計而言，主要是一種隨機生成的虛擬資料，其目的在替代眞實資料以擴增大眾應用，並能有效保護眞實資料之隱私。合成資料的生成方法（圖 2），主要是基於眞實資料的統計特徵，運用生成技術產生與其有相似分布並具高度相關的人工資料，但在個別訊息層面，其敏感性資料均被虛擬資料所取代，而能大幅降低披露風險。生成技術已有久遠歷史，但政府統計的應用始於 2003 年美國普查局對「收入及計畫參與調查（Survey of Income and Program Participation，SIPP）」資料生成部分合成資料集，並於 2007 年開始供大眾使用。基於資料開放需求及隱私風險考量，近年來有些國家統計局積極進行相關研發，UNECE 現代化小組為促進合成資料方法的統計完整性及有效實踐，亦正進行工作組研討作業，就合成資料的用途類型、推薦方法、披露風險、試驗建

議等項目，於 2021 年底提出研究結果。

政府統計資料之釋出，傳統運用統計披露控制（Statistical Disclosure Control，SDC）技術，依資料特性進行適當的概化或抑制（Generalization or Suppression）措施，以避免識別特定個體而達到保護隱私目的。對於完整資料集，SDC 並無法有效避免隱私披露風險，因而採管制室作業予以限制性使用。合成資料虛擬眞實資料，原則上幾無披露風險，但其挑戰主要在於從眞實資料中複製所有必要的特徵相當複雜，過程中亦可能輕忽遺漏造成不一致，並因個案或有不同，因此合成資料須經嚴謹的效用評估，驗證其替代有效性。

三、多源資料整合（Multi-source Data Integration）

多源資料整合係將來自統計調查、行政登記、大數據、空間訊息及私營部門紀錄等兩種以上來源資料之結合。整合方法分爲紀錄連結（record linkage）及統計匹配（statistical matching）兩大類，利用鍵值（key）、機率模式或統計值等連結對應。資料整合具改善既有統計程序，減少訪查負擔，以及增進資料涵蓋面之契機，歐盟統計局於 2006 年開始進行相關研究及推展計畫，現已成爲歐盟國家重要的統計生產方式。

傳統調查資料是依其目的經過「設計」而獲得，有統計理論的支持與理解；資料整合來自各有其利用目的不同之資料源，面臨的挑戰主要在資料取用的持續性及對資料內涵的熟悉度，以及對資料準確度、一致性、及時性、可比性及遺漏等品質要求，應有可行的處理方式，並能重複且穩定地產生結果。

肆、創新技術應用實例

機器學習應用於政府統計正逐步增加，UNECE 亦積極研討其遵循之統計規範及應用方式。加拿大統計局開發機器學習方法，從證券管理單位取得各上市公司提交的年度財務報表中，識別及擷取財務變量訊息，每年可及時且自動處理 7 萬個文件檔作業，大幅減省原需人工作業時間，並提升資料正確性。瑞士統計局運用機器學習技術，驗證行政資料及調查資料的品質及可靠性，可自動識別潛在的錯誤模式，不僅縮短作業時間並提高資料品質，驗證結果較原方式發現的錯誤增加 29%，且預測準確率超過 94%。

合成資料目的在替代眞實資料，並具有相似的資料分布及性質，能兼顧保護資料隱私及擴大資料開放應用。美國普查局線上查詢系統 OnTheMap 是一個結合統計資料、街廓地圖及統計圖表的系統，可提供小地區統計供大眾查詢，該系統利用合成技術，監測資料敏感變量，以保護分類上稀少資料之隱私。英國統計局建立勞動力調查的合成資料，提供使

用者熟悉資料的結構及測試，以簡化在安全環境之實際使用。

多源資料整合所帶來時效、成本、細緻度及涵蓋面等多方效益，已成爲國際間關注應用的方法，多源統計生成是重要的研討課題。美國普查局多年來致力整合的雇主與住戶縱向動態系統（LEHD）及人口普查縱向基礎建置計畫（CLIP），深獲好評並受到廣泛應用。我國近年來亦積極運用資料整合方法，創編多元薪資統計及國人赴海外工作人數統計等多項成果，擴增統計應用。

伍、結語

政府統計的創新發展，統計技術一直扮演重要角色，資料革命使資料複雜度增加，亦提高了處理的技術層次，並形成技術的多元及動態發展，有些尚處萌芽正持續精進中。面對資料生產範式的變革，新技術別於傳統方法所產生的質疑，UNECE 提出研究建議及詮釋，爲符合聯合國「政府統計基本原則」奠立基礎，相關技術發展所面臨的挑戰更勝以往，須不斷學習以保持競爭力，各國推展經驗可作爲借鏡，適當調整組織功能及加強交流合作，重視人員培育及人才引進，評估適用技術積極進行研發及應用，爲政府統計的永續創新鋪平道路。

註釋

1. 沙盒（Sandbox）試作：意爲一種擬眞的獨立環境，提供新方法的有限試作，其風險僅限於該環境中。各組織運作上是由有意願之會員國於其國內試行，並將經驗成果分享各國。

參考文獻

1. Ioannis Kaloskampis (2019). Synthetic data for public good. UK ONS Data Science Campus.
2. Machine Learning Team (2018). The use of machine learning in official statistics. UNECE HLG-MOS Machine Learning Project.
3. Machine Learning Team (2020). A Quality Framework for Statistical Algorithm. UNECE HLG-MOS Machine Learning Project.
4. Statistics Netherlands (2019). Interim review 2017-2019: Three years of Center for Big Data Statistics at Statistics Netherlands.
5. Swiss Federal Statistical Office (2020). Data Science Competence Center.
6. UK Data Science Campus (2021). https://datasciencecampus.ons.gov.uk/.
7. 羅國華（2018），政府統計資料整合的機遇與挑戰，主計月刊，756 期，84-89 頁。❖