論述 》統計 · 調查

國際人口普查應用機器學習之契機與挑戰

機器學習在政府統計具相當潛力，各國人口普查主要應用在資料分類及註碼，而檢誤、插補及圖像分析大部分則仍處於研究測試階段，其發展經驗或可作為我國辦理相關調查參考。

周元暉（行政院主計總處國勢普查處研究員）

壹、契機

國家統計機構主要任務是產生及發布具公信力之統計資料，聯合國文獻「政府統計之機器學習應用（Machine Learning for Official Statistics）」指出在使用者對於相關性（relevant）、及時性（timely）、可及性（accessible）及豐富性（detailed）統計資料需求日增之環境下，探索機器學習如何提升資料品質，以及如何因應及提供更好統計服務厥爲關鍵。受新冠疫情（COVID-19）影響，資料品質更側重於即時性，機器學習爰扮演要角，尤其對於需勞力密集及重複性之作業，可藉由自動化而更有效率地完成。機器學習雖然具有優勢，惟實務上仍有侷限性，應用上必須考量下列 6 項要件：1. 符合業務需求；2. 達到預定資料品質；3. 具附加價值；4. 表現穩健；5. 考量道德及法律；6. 堅實科學基礎。

聯合國文獻中亦指出機器學習在政府統計中可應用於資料分類及註號、檢誤、插補及圖像分析，在各國人口普查作業應用上以分類及註號最爲普遍。本文除介紹新加坡、加拿大、法國最近一次人口普查應用機器學習在行職業註號的經驗，並分享我國人口普查在機器學習上之應用，另就澳洲、義大利等國之政府統計，在檢誤、插補及圖像分析導入機器學習所獲成效及面臨之挑戰，期能對我國政府統計的資料處理方式有所啓發。

貳、各國人口普查導入機器學習之情形

爲了解常住人口之工作型態，精確判定其從事行職業，人口普查以文字資料方式蒐集

相關資訊，其中「行業」填報內容係指受訪者工作場所之主要經濟活動，「職業」則爲受訪者所從事之職稱及主要工作內容、職責，文字描述判斷後，再依統計標準分類進行註號。各國普查當局多應用機器學習提升文字答項資料處理效能，以下就新加坡、加拿大、法國及我國導入機器學習的相關經驗進行探討。

一、新加坡

新加坡人口普查行職業註號方式，依序採自動化批次註號（batch coding）及人工註號（manual coding），無法使用批次者才由人工介入處理。2020年普查行職業註號作業首次導入機器學習應用技術，並運用2020年普查資訊系統（Census 2020 IT System），及人力資源部所辦理之歷次全面勞動力調查（the Ministry of Manpower's Comprehensive Labor Force Survey）資料訓練學習，將答項置入模型，依循國家標準行職業分類預測註碼。

機器學習應用包括4項步驟（圖1及圖2）：

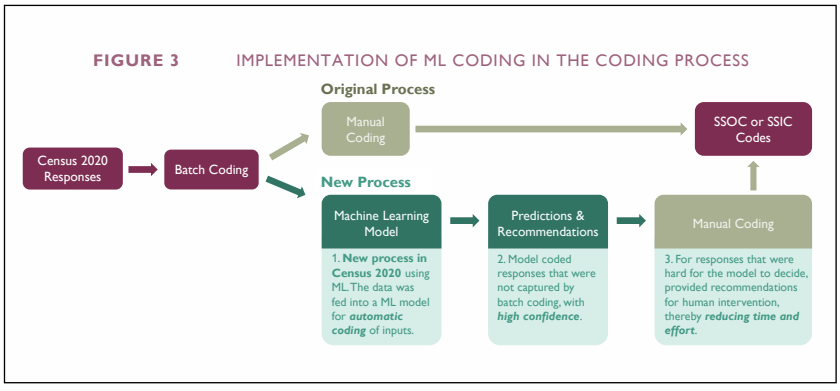
（一）清理資料：預先將資料整理爲標準格式，以符模型使用。

（二）訓練及選擇模型：神經網絡（Neural Network, NN）模型，模仿人腦處理資訊之過程爲最佳執行模型。

（三）產生預測註碼（predicted codes）：就國家標準行職業分類5位註碼計算分數，獲得最高分之註碼被選爲預測註碼；分數差者，則改以人工註號處理；預測註碼分數介於中間者，由註號人員參考建議註碼進行註號。

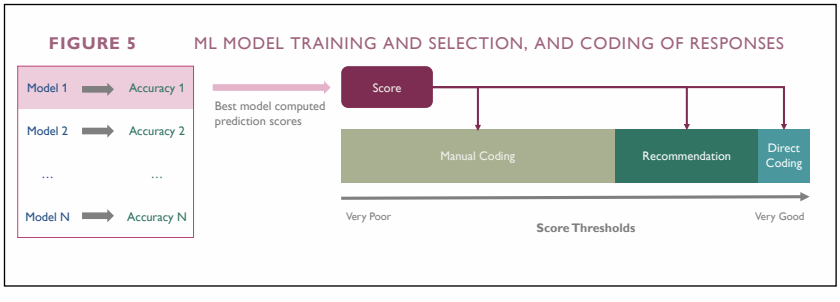
（四）評估品質：定期檢查及給予回饋以改進模型成效。

圖1　新加坡普查行職業之機器學習應用



資料來源：Coding of SSOC/SSIC in Census 2020 using Machine Learning, Statistics Singapore Newsletter, Issue 2, 2021, p17, Singapore Department of Statistics.

圖2　新加坡機器學習模型之訓練、選擇與答項註號作業



資料來源：Coding of SSOC/SSIC in Census 2020 using Machine Learning, Statistics Singapore Newsletter, Issue 2, 2021, p18, Singapore Department of Statistics.

就成本效益而言，新加坡在 2000 年系統之批次註號率僅 6-7％，2010 年升至 18％，2020 年將機器學習模型整合於註號系統後，行業及職業註號率已分別升至 75％及 30％，並節省約 5,600 個工時（man-hours），時效大幅提升。惟隨新興產業之發展，透過人工審核來管控品質仍無可避免。

二、加拿大、法國

加拿大人口普查作業中，文字答項之註號過去大部分係透過對照檔以自動化方式完成，少部分由人工處理。爲提升普查資料品質，統計局試圖研發新技術取代人工作業，並於 2021 年人口普查首次透過自然語言演算（the fasttext natural language processing algorithm）處理，將機器學習應用於文字答項之註號。引入相關技術後，時效大幅提升，減省 690 萬筆人工註號作業，並節省約 9,300 萬元（400 萬加幣）。

法國人口普查自 2004 年起改以市鎮人數爲門檻分省按年辦理滾動式調查（Annual Census Survey, ACS）取代傳統全查方式，以 5 年之中間年爲資料蒐集基準年產製統計結果，其行職業註號分爲自動化及人工處理，目前人工處理約占 12％。2020 年法國修訂職業標準分類，並訂於 2024 年辦理按年普查調查（ACS）時使用，考量舊有之自動辨識模式已無法因應新的分類方式，因此導入機器學習模型。2021 年先進行大規模人工註號，用以訓練模型及驗測資料，最終採 2 層神經網絡演算（two-layer neural network algorithm）模型，驗測結果顯示行職業辨識準確度提高。

三、我國

我國 2020 年人口及住宅普查輔助行職業註號系統，係蒐集按月人力資源調查、工業及服務業普查相關調查結果及勞健保等公務登記資料，建立相關文字及對應註號之辭庫，導入審核系統以自動辨識行職業，提供參考註碼建議，提升編碼效能。

另爲估計普查中間年區域別常住人口及編製常住人口就業失業統計，運用 2020 年人口及住宅普查結果，並整合多元公務檔案資料，分爲 80％訓練樣本及 20％測試樣本，透過拔靴取樣（Bootstrap Sampling）產生不同的訓練資料集，建置隨機森林（Random Forest）模型，用以估計個人常住縣市，經驗證模型估計之正確率可達 8 成以上。

參、機器學習在檢誤、插補及圖像分析上之應用

由於機器學習應用在檢誤、插補及圖像分析更加繁瑣及複雜，目前各國多在測試及試行階段，其初步成果如下：

一、資料檢誤及插補

檢誤之機器學習演算是從過去檢誤結果學習而來，包括識別各項目間之邏輯及內容是否合理，從而改進機器學習模型中之檢誤規則。英國及義大利統計局研究發現，因爲機器模型可以處理大量資料，所以檢誤過程相對有效，但爲更新訓練資料及評估成效，短期內並無法減省成本。

再就插補而言，理論上機器學習模型應可改善插補模型，並達到下列預期目標：1. 讓插補值接近眞值；2. 插補順序之準確性（ranking accuracy）；3. 插補值之分配與眞值一致；4. 推論參數具不偏性及有效性；5. 插補值具合理性（imputation plausibility）。比利時、德國及波蘭統計局研究發現，與傳統插補模型比較，機器學習模型部分變數因無須事先轉換處理，更能即時產生結果，惟須注意只有在假設符合的前提下，才能得到良好效果。

義大利統計局爲插補居民基本登記（the Base Register of Individual, BRI）個別居民之教育程度，使用多層次感知器（the Multi Layer Perception, MLP）演算法建置神經網絡模型進行測試，相較於傳統對數線性（log-linear）模型，導入機器學習確具時效性。但以2018年普查資料驗測插補值之分配，傳統模型仍優於機器學習模型，因而決定不發布插補結果。我國人口普查已結合公務登記資料，不但讓問項更加精簡且資料正確性提升，並使用機器學習方法建立常住人口模型，定期推估市縣別常住人口，以銜接兩次人口普查間常住人口資訊。

二、圖像辨識及分類

圖像辨識及分類傳統以人工方式進行，隨機器學習導入得以自動化。澳洲統計局爲提升地址登記（Address Register, AR）之品質及處理每季上萬筆尚未分類的新增地址，與國家科學及工業研究院（the Commonwealth Scientific and Industrial Research Organization, CSIRO）合作，建立自動化圖像辨識模型（the Address Register Automated Image Recognition, AIR），該模型運用卷積神經網絡模型（deep-learning convolutional neural network）處理大量圖像資料，將地址區分爲居住用、建造中、空地、商用、高密度人口、地理編碼不完整等6類，除處理時效提升，並減省大批人力，爲提升其應用效益，須持續訓練模型，建置備援模型。

美國普查局爲維護各項調查所需地址主檔（Master Address File, MAF），針對全國住宅地址持續辦理地址巡檢作業，基於預算及成本考量，於普查中間年重新設計地址巡檢作業，從衛星影像圖判定及驗證地址異動部分，2020年普查75%地址已於室內巡檢完成，較上次普查縮減65%實地巡檢作業，成效相當顯著。我國近2次普查使用地理空間圖資輔助普查區界限劃分，至於圖像分析中衛星圖及全球定位資料之應用，仍應審愼評估個資之保護及資料品質之可靠度。

肆、未來挑戰

人口普查中行職業機器學習模型發展最普遍且成熟，其應用成效與辭庫完整性及是否定期維護有很大關聯。新加坡行職業辭庫除了歷次普查及家戶調查資料、最近一次國家行職業標準分類（含敘述及註碼）以外，還包括大量行政紀錄（administrative sources）。法國經驗顯示模型準確度會隨時間經過而遞減，所以定期訓練資料爲重要關鍵。加拿大發展

過程中也面臨諸多挑戰，包括資料處理作業及時程因導入機器學習而必須作相應調整，且原普查之資料處理方式已行之有年不易推動創新，另機器學習應用含括多項專業知識及技術，須進行跨領域合作。

此外，在主觀性相對較高之資料處理作業則挑戰相對較高，如檢誤及插補。圖像分析之機器學習應用則須視一國圖資之完整性而定，如美國除了建物之座落地址以外，還包括地理編碼、位址特性及與其他建物之關係、經緯座標及地理屬性等描述。此外，還須考慮圖資使用權限之適法性，並與地方政府、企業密切合作，以厚實相關技術之基礎。原則上，機器學習之發展需要完整架構及道德考量，並確保在資料處理過程中不侵犯個人隱私，且無傳統社會、性別成見及文化刻板印象。人口普查規模龐大，機器模型的建置，不僅需要技術、專才及預算，還須符合成本效益，並獲得主管機關之支持，始能推動與發展。

伍、結語

就資料相關性、正確性、及時性及成本效益而言，機器學習在勞力密集、重複性高且穩定之作業極具效益，新加坡、加拿大、法國及我國經驗即為實例。至於在檢誤、插補及圖像分析方面，機器學習仍在研究測試階段，未來在人口普查應用上還有一段長遠路要走。由各國經驗顯示，資料總會出現新形態或特例依賴人工解決，即使將其納入訓練資料，現階段機器學習仍難以完全取代人類智慧。我國所辦各項基本國勢調查及抽樣調查，亦應跟隨時代腳步，積極融入人工智慧，以提升效率及品質，開創統計調查新紀元。

參考文獻

1. Machine Learning for Official Statistics, United Nations Economic Commission for Europe, United Nations (UNECE), Geneva, 2021.
2. Coding of SSOC/SSIC in Census 2020 Using Machine Learning, Chen Tian Min, Teo Zhiwei, and Chan Wen Chang, Statistics Singapore Newsletter, Issue 2, 2021, p16-20, Singapore Department of Statistics.
3. Census of Population 2020-Administrative Report, Chapter 4 Data Processing and Dissemination, Singapore Department of Statistics, 2021.
4. Integrating Machine Learning Techniques to 2021 Census of Population Coding, Observatory of Public Sector Innovation, Organization for Economic Cooperation and Development (OECD), 2023.
5. Machine Learning for Coding Occupations in the Census : First Lessons from Experiments to Production, Theo Leroy, Lucas Malherbe, Tom Seimandi, Elise Coudin, National Institute of Statistics and Economic Studies (INSEE), 2023.
6. Imputation of the Variable "Attained Level of Education" in Base Register of Individuals, Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona, Toti, Deigo Zardetto, 2020.
7. An ML Application to Automate an Existing Manual Process Through the Use of Aerial Imagery, Daniel Merkas, Debbie Goodwin, 2020.❖