專題

運用大數據技術創編區域常住人口統計

為突破常住人口統計因普查每 10 年更新一次之限制，及因應當前社經變遷之需求，爰運用人口普查與多元巨量資料，導入人工智慧、機器學習、資料探勘等大數據分析工具，創編區域常住人口統計並定期提供應用。

楊麗華、林姿吟、黃宇葶（行政院主計總處國勢普查處科長、專員、科員）

壹、前言

常住人口統計係國家政策釐訂及資源配置之重要資訊，舉凡居住政策、公共設施、交通建設、產業發展、醫療資源及文化活動等，均可參考人口特性結構分布，進行區域規劃，以符合實際需求。每 10 年舉辦一次之人口普查爲常住人口統計資料來源，由於更新週期較長，致普查中間年相關資料付之闕如，僅能應用戶籍資料。民國 80 年以前戶籍人口和常住人口差異不大，但隨著社會及經濟發展，因爲工作、就學、福利措施等因素，使得實際居住地和戶籍不一致的情形漸增。依 99 年人口及住宅普查結果，戶籍人口中已逾 2 成並未住在設籍地，近年普查試驗調查結果則接近 3 成，顯見戶籍人口資訊已無法充分支援區域政策所需。爰此，爲因應社經快速變遷，亟須突破常住人口統計 10 年更新一次之限制，運用大數據概念及相關統計技術，以普查年資料爲基準（benchmark），整合跨部會公務登記資料，建立估計模式，以常川提供區域別常住人口資訊。

貳、作業方法

爲定期產生常住人口資訊，本作業導入人工智慧（Artificial Intelligence, AI）及相關統計技術，以 99 年人口及住宅普查統計結果爲基準，蒐集同時期各部會相關公務登記資料，建立常住人口估計模型，並規劃建置普查資料庫，茲說

明作業方法如下：

一、整合戶政登記、入出境紀錄及外國人相關登記資料，運用人口變動要素合成法（Cohort Component Method），編算全國常住人口。

二、蒐集各機關公務登記有關人口常住地資料，以99年普查資料爲基礎，採用相關分析（Correlation Analysis）及決策樹（Decision Tree）CHAID法（Chi-square Automatic Interaction Detector），解析及篩選重要變數。

三、導入AI隨機森林法（Random Forest）建模，以拔靴取樣（Bootstrap Sampling）及機器學習（Machine Learning）方式，建立人口常住縣市估計模型。

四、結合每年模型估計結果及家戶面統計調查最新資料，常川更新「普查常住人口資料庫（Census Resident Population Micro Database）」，以動態提供常住人口資訊。

參、編算全國常住人口

爲編算全國常住人口總數，本作業運用公務登記資料，依常住人口定義，按月採用基期累算法編算，按年採用基準日編算法校正。

一、按月基期累算法

基期累算法亦即人口變動要素合成法（Cohort Component Method），是目前各國最常使用於人口推估的方法，其做法係以前一期推估之全國常住人口數爲基礎，運用相關公務檔案，產生當期出生、死亡及國際遷徙等自然及社會增減變動要素，按月依流動情形計算當期之全國常住人口總數。公式如下：

$P_{t+1} = P_t + B_{t,t+1} - D_{t,t+1} + NOM_{t,t+1}$

P_t：t期末之常住人口推估數

P_{t+1}：t+1期末之常住人口推估數

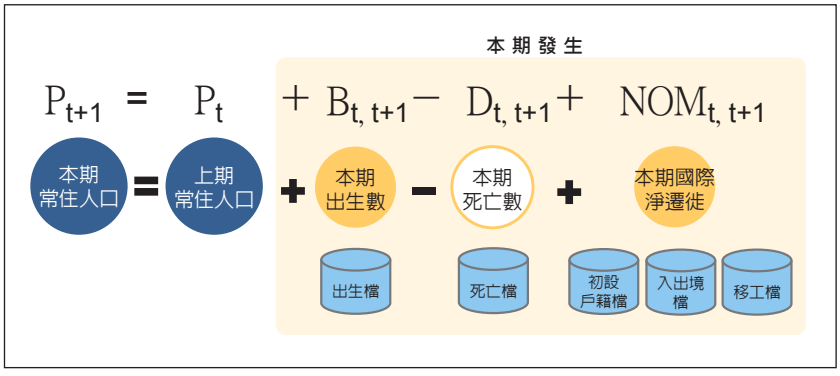
$B_{t,t+1}$：t,t+1期發生之出生數

$D_{t,t+1}$：t,t+1期發生之死亡數

$NOM_{t,t+1}$：t,t+1期之國際淨遷徙人數

本法係以「99年人口及住宅普查」常住人口爲基期資料，當期實際出生、死亡人數運用出生、死亡登記檔產生，而當期國際淨遷徙人數則以遷入之常住人數、初設戶籍人數及移工人數，扣除遷出之常住人數，主要運用檔案爲入出境資料檔，依推計日計算個人一年內停留境內及境外日數是否達183日判定，資料處理流程參見圖1。

圖1　基期累算法資料處理流程圖



資料來源：作者自行繪製。

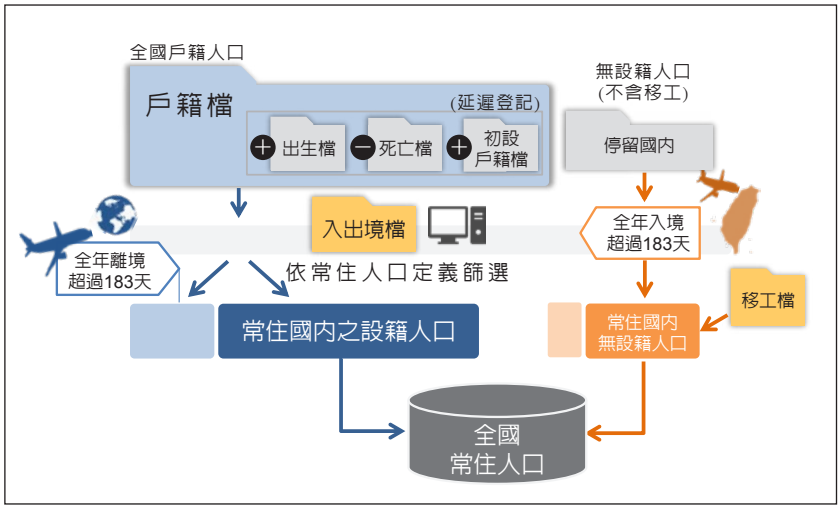
二、按年基準日編算法

依據人口及住宅普查定義，常住人口係指標準時刻實際居住在國內已達或預期達 6 個月以上之所有本國籍、外國籍、大陸地區（含港澳）人口。基準日編算法係運用推計日之全國戶籍人口，扣除全年居住國內未達 183 天之設籍人口，加上全年居住國內達 183 天之無設籍人口。基本公式如下：

> 全國常住人口＝全國戶籍人口－非常住國內之設籍人口＋常住國內無設籍人口

由於各種公務檔案均有延遲登記情形，本編算法推計日之全國戶籍人口係以戶籍檔爲基準，爰運用出生、初設戶籍及死亡登記檔推計，增加基準日前尚未登記之新生兒、初設戶籍人口，扣除尚未通報之死亡人口，至是否常住國內則再連結入出境資料檔判定，其中非常住國內之設籍人口以篩選全年累計停留境內未達 183 日或累計離境達 183 日者推計，另常住國內無設籍人口則依外國人及大陸港澳地區人民入出境檔，篩選全年累計停留境內達 183 日之推計，資料處理流程參見圖 2。

圖 2　基準日編算法資料處理流程圖



資料來源：作者自行繪製。

肆、常住人口估計模型

爲提供區域別常住人口資訊，本作業採用大數據分析工具，建立個人常住縣市估計模型，以下分別說明變數篩選流程、建模方法及普查資料庫建置內容。

一、變數篩選

模型建立所需之變數，係經由蒐集各種公務檔案，並評估各種檔案資料項目內涵後，透過資料清理（Data Cleansing）、資料整合（Data Integration）、資料轉換（Data Transformation）等資料前置處理（Data Preparation），並運用統計相關分析（Correlation Analysis）及決策樹 CHAID 法選取重要變數。

茲以就學、工作及就醫地相關變數之資料處理爲例說明（下頁圖 3）：

（一）就學相關變數篩選

彙整 4 個年度教育程度通報檔資料，依年齡與教育程度合理性，篩選出 15 歲以

上在學者，透過連結大專校院及高級中等學校各校區地址，再依就讀學校之校區縣市與戶籍縣市距離之通勤機率判定，產生就學縣市作爲模型投入變數。

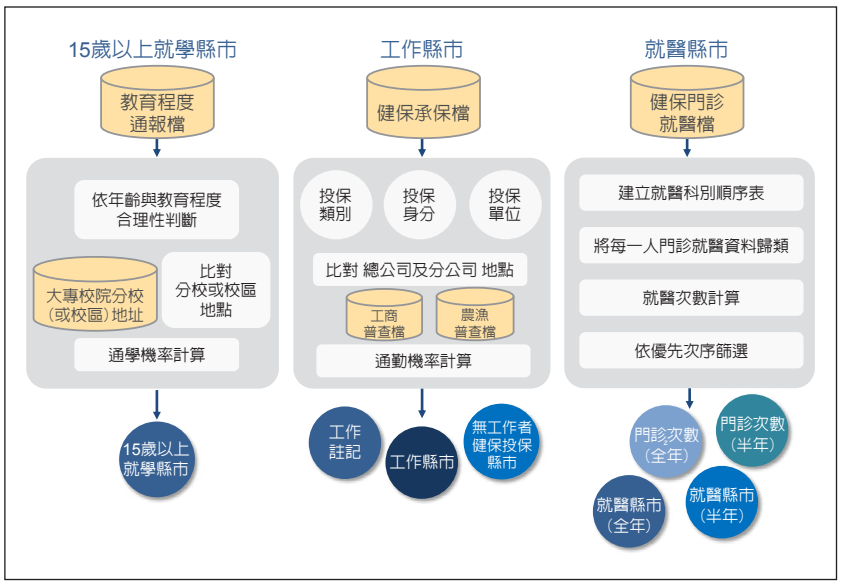
（二）工作相關變數篩選

依健保承保檔之投保類別及投保身分，區分有無工作，有工作者之投保單位依規定可由總公司或分公司辦理，因此，多營業場所之公司，員工實際工作縣市可能與投保單位縣市不同，爰透過連結工商普查檔及農業普查檔，可建立多個營業場所縣市資料，再與戶籍縣市距離之通勤機率判定，產生工作註記及工作縣市等變項。

（三）就醫相關變數篩選

運用99年普查資料與同時期健保就醫資料，分析每個縣市之人口就醫行爲，採用決策樹CHAID法建立就醫科別順序表，並分析診所與醫院層級、門診次數及就醫縣市等變數，以篩選與常住縣市關聯性較高之變數。

圖 3　就學縣市、工作縣市及就醫縣市相關變數資料處理流程圖



資料來源：作者自行繪製。

二、模型建立

（一）隨機森林法

隨機森林（Random forest）是由Leo Breuman發表一種機器學習的演算法，將其在1996年提出的Bagging機器學習理論與Tin Kam Ho在1998年提出的隨機子空間方法相結合。原始隨機森林演算法中分類器爲CART（Classification and Regression Tree）樹，透過Bagging演算法進行組合學習，並在CART樹生長時隨機選取變數進行分裂。隨機森林演算法與傳統的決策樹相比，有更強的泛化能力與分類效果。

（二）模型建立與評估

以最新人口及住宅普查調查資料爲樣本，其中以80%爲訓練樣本，20%爲測試樣本，以訓練樣本連結各公務檔案，作爲建模基準檔，其中設籍註記、常住縣市分別爲反應變數，經由變

數篩選流程，將個人基本資料以及各種可能爲常住地之縣市等項目作爲模型投入變數。

採用隨機森林建模，係透過拔靴取樣（Bootstrap Sampling）產生不同的訓練資料集來建構各個分類器（亦即森林中的決策樹），當使用 Bootstrap 方式生成訓練集時，原始樣本中有一部分資料（低於 40％）不會出現在訓練集中，這些資料便稱爲 OOB（Out-Of-Bag）資料，運用 OOB 的誤分類率可以直接評估模型效果。

三、規劃建置普查資料庫

爲擴大普查應用範疇，規劃以普查資料爲基礎，運用每年模型估計結果，並蒐集個人縱貫面及橫斷面資料（Individual longitudinal & cross-sectional data），並結合相關公務資料及各項家戶面抽樣調查資料，建置普查常住人口資料庫（Census Resident Population Micro Database），在符合資訊安全規範下，提供統計、分析及探勘等應用。

伍、結語

運用大數據技術創編區域常住人口統計之作業，其革新性及預期應用效益如下：

一、普查年蒐集基礎資料，非普查年定期推估常住人口

爲突破常住人口統計 10 年更新一次之限制，建立「普查年蒐集基礎資料，非普查年定期推估常住人口」機制，以普查年資料爲基準，整合跨部會公務登記資料，運用大數據及人工智慧技術，創建常住人口估計模型，規劃於非普查年提供區域別常住人口統計，大幅縮短常住人口資訊產生週期，提升資料時效，及時供爲資源配置及區域規劃之參據。

二、提升區域別調查結果推估品質，增進常住人口統計應用價值

運用模型估計之區域常住人口，作爲各項家戶面抽樣調查區域統計之推計基礎，確實反映區域別就業統計及家庭型態等重要家戶面調查指標，提升調查資訊應用效益，後續透過資料庫可動態觀察並呈現常住人口資訊，增進常住人口統計價值。❖