出國報告（出國類別：研究）

赴美國研習地址母體檔作業方法及差分隱私技術

服務機關：行政院主計總處
姓名職稱：謝博行專員、林昌諒專員
派赴國家/地區：美國
出國期間：**113** 年 **9** 月 **21** 日至 **113** 年 **9** 月 **26** 日
報告日期：**113** 年 **12** 月 **9** 日

摘要

人口普查的結果是國家政策擬定的重要參據，因此資料品質之良窳至為重要。然隨資通訊技術蓬勃發展、調查環境日益艱困，又個人隱私保護的需求日趨強烈，調查相關作業有必要做出相應之調整。藉由本次赴美與普查局及費城地區辦公室首長及專家之交流，了解美國在地址母體檔案的維護及隱私保護技術的實務經驗，對我國家戶面調查所需之抽樣底冊及對民眾之隱私保護承諾都極具參考價值。

本出國報告之內容主要分為二個主題，第一個主題是「地址母體檔（MAF/TIGER）的建置與維護」，其基礎來自地址主檔（MAF）與拓撲整合地理編碼和參照系統（TIGER）的兩項資料的整合，用以支持美國人口普查和相關家戶面調查的作業基礎。該系統不僅管理住宅單元和地理位置數據，還通過多層次更新策略（如利用美國郵政服務的遞送序列檔案、地方政府合作及地理空間支持計畫）確保數據的準確性與完整性。此外，美國普查局開發的辦公室內地址清查（IOAC）與實地地址清查（IFAC）技術，通過高效的影像審查及現場實地判定，有效降低了普查成本並提高數據品質。

第二個主題為「差分隱私（Differential Privacy）作業及隱私保護技術」，近年美國普查局面對數據重建與重新識別的威脅，逐步引入更高標準的隱私保護技術。自早期的資料遮蔽與資料交換到最新的差分隱私技術持續精進，差分隱私已成為目前核心隱私保護技術，不僅具有數學可證明性，還能根據數據需求靈活設置保護參數，可有效防止統計數據被還原為個人資料。2020 年美國人口普查已首次應用差分隱私，在統計資料中添加噪音，藉此平衡數據的準確性與隱私性，確保公開統計數據能兼顧安全與實用性，相關經驗值得我國借鏡。

目次

摘要 i
目次 ii
表目次 iv
圖目次 v
第一章 研習目的.....1
第二章 地址母體檔（MAF/TIGER）的建置及維護.....3
第一節 地址母體檔的發展.....3
一、MAF/TIGER 資料庫概述.....3
二、MAF/TIGER 系統的演進.....4
第二節 MAF 資料庫架構及有效地址篩選.....5
一、MAF/TIGER 系統的資料庫架構.....5
二、有效地址篩選原則.....7
第三節 MAF/TIGER 資料更新.....7
一、MAF/TIGER 資料來源與更新機制.....7
二、MAF/TIGER 系統資料驗證及更新流程.....8
第四節 地址清查作業.....10
一、辦公室內地址清查（In-Office Address Canvassing, IOAC）.....10
二、實地地址清查（In-Field Address Canvassing, IFAC）.....19
第三章 差分隱私（Differential Privacy）作業及隱私保護技術.....24
第一節 美國隱私保護技術發展.....24
一、隱私保護重要性.....24
二、隱私保護歷程.....24
三、隱私保護新挑戰.....25
第二節 差分隱私（Differential Privacy, DP）技術.....27
一、差分隱私概述.....27
二、差分隱私定義.....28
三、美國揭露避免系統.....31
第三節 TopDown Algorithm（TDA）演算法.....32
一、P.L. 94-171 重新劃分選區數據摘要檔案.....32
二、TopDown Algorithm（TDA）運作原理.....32
三、TDA 結果補充說明.....39
第四節 SafeTab 演算法.....41
一、詳細人口與住宅特徵檔案（Detailed Demographic and Housing Characteristics File）A.....41
二、SafeTab 運作原理.....41
三、誤差範圍衡量.....44

第五節 PHSafe 演算法 45
一、補充人口與住宅特徵（Supplemental Demographic and Housing Characteristics, S-DHC）檔案 45
二、PHSafe 運作原理 46
第六節 差分隱私應用 50
第四章 心得與建議 52

表目次

表 1-1　研習日程表 2
表 2-1　互動式審查處理結果 13
表 2-2　互動式審查前後資料處理結果 13
表 2-3　抽樣底冊維護計畫處理之活躍區塊地址數 14
表 2-4　集體住所／臨時地點（GQ/TL）審查處理結果 17
表 2-5　LUCA 地址驗證（LAV）處理結果 18
表 2-6　資料品質控管（QC）特徵與違規記點對照表 21
表 2-7　BCU 清查及分層抽樣後複查工作量 22
表 2-8　BCU 抽樣計畫容許錯誤數 22
表 2-9　BCU 分層抽樣後複查結果 23
表 3-1　不同揭露避免方法的比較 28
表 3-2　2020 年人口普查重新劃分選區檔案變數分類 34
表 3-3　人口普查街廓群組的噪音添加範例 35
表 3-4　後處理說明 36
表 3-5　使用 TDA 統計結果不合理的情形 40
表 3-6　各類表格之人口數門檻 41
表 3-7　美國的新加坡人加入噪音範例 43
表 3-8　家戶總人口資料表截斷範例 47
表 3-9　僅在最內層的單元格加入噪音 48
表 3-10　2020 年美國人口普查產品所使用的重要避免揭露技術 50

圖目次

圖 2-1　MAF and TIGER databases.....5
圖 2-2　地理空間支援系統合作夥伴計畫提供的資料地區分布.....8
圖 2-3　辦公室內地址清查時序.....10
圖 2-4　互動式審查工具 BARCA 畫面.....11
圖 2-5　比對當前影像及基準影像顯示建物增加.....12
圖 2-6　比對當前影像及基準影像顯示建物減少.....12
圖 2-7　各地區未地理編碼比率.....15
圖 2-8　IOAC 結果被標記為「未解決」的地址被指派給 IFAC.....19
圖 2-9　編列員及訪查設備.....20
圖 3-1　美國人口普查隱私保護歷程.....24
圖 3-2　資料庫重建結果與人口普查資料比對一致率.....25
圖 3-3　隱私損失與精確度關係.....29
圖 3-4　拉普拉斯與高斯分布.....29
圖 3-5　隱私損失預算（ε）－調節噪音大小的控制旋鈕.....30
圖 3-6　TopDown Algorithm.....33
圖 3-7　標準地理區域層級.....37
圖 3-8　避免揭露系統地理區域層級.....38
圖 3-9　TDA 堪薩斯原住民地區範例.....39
圖 3-10 自動調整設計.....42
圖 3-11 PHSafe 和後處理流程.....46

第一章　研習目的

人口為國家基本要素之一，其組成、素質、分布、發展及遷徙均關係著國家發展與社會福祉。根據聯合國統計，全球約有 9 成以上國家於西元 2020 年前後辦理人口普查，我國人口及住宅普查（以下簡稱人口普查）迄今已順利完成 7 次，有助於瞭解全國人口之質量、家庭結構、就學就業及住宅使用狀況，供為政府研訂施政計畫、規劃國家建設發展之主要參據。

隨著經社環境變遷，各界對統計調查相關資料時效之要求逐漸提高，但同時國家經濟發展也促使家庭型態改變，並推升民眾個資保護意識。民眾越發注重個人隱私及自身權益，調查環境漸趨艱困，全面性普查之成本日益提高，聯合國於西元 2000 年人口及住宅普查原則與建議中即提出，普查除辦理全面訪查外，亦可整合運用公務登記與調查資料方式辦理，尤其強調規劃時應審慎考量成本、資料品質及作業可行性等問題，我國爰於 2010 年普查起，改以公務資料為基礎，無法自公務檔案產生之常住人口特徵及分布、家戶結構、遷徙及健康照護等項目則輔以調查蒐集。2020 年賡續採「公務登記及調查整合式普查」方式辦理，整合相關公務登記檔案，以取代或簡化部分調查項目，並結合地理資訊系統與抽樣技術等方法，提升普查辦理效率與資料品質。

美國普查局辦理普查經驗豐富且技術卓越，在地址母體檔維護及隱私保護技術上居領先地位，爰赴美研習地址母體檔如何運用相關資料在辦公室內辦理判定作業、實地判定時所使用的軟硬體設備，以及資料隱私保護技術的概念、應用情形，進而精進我國普查作業方式與普查資料品質。

在美期間承蒙美國普查局（Census Bureau）國際經濟合作部門 Mr. Godo Seri 安排參訪，針對地址主檔之建置、更新方法及差分隱私避免揭露技術，與美國普查局 Stuart Irby, Chief、Kyra Linse, Survey Director 及 Michael B. Hawes, Senior Survey Statistician 等人進行會談，就人口母體資料庫進行討論；另與費城地區辦公室（Philadelphia Regional Office）Rosa M. Estrada, Assistant Regional Director 及 Mario Matthews, Assistant Regional Director、Robert Stabs, Decennial Specialist 等人就地區辦公室運作及實地資料蒐集作業進行會談。本次赴美研習，承蒙駐美國臺北經濟文化代表處經濟組吳秘書俊逸協助安排行程並聯絡美方，使參訪過程順利圓滿，獲益良多，特致謝忱。

表 1-1　研習日程表

時間、地點	研習議題	簡報主講人
9 月 23 日 **美國普查局**	地址主檔（MAF）建置與更新	Stuart Irby, Chief 地理部門
	美國社區調查（ACS）、當前人口調查（CPS）及人口普查（Population Census）	Kyra Linse, Survey Director 當前人口調查與美國時間利用調查小組
	差分隱私（DP）作業與避免揭露技術	Michael B. Hawes, Senior Survey Statistician for Scientific Communication 研究與方法學部門
9 月 24 日 **美國普查局費城地區辦公室**	費城地區辦公室組織架構情形	Timothy Maddaloni, Program Coordinator SIPP, DAAL, RHFS 費城地區辦公室
	實地資料蒐集管理經驗交流	Ian Manners, Program Coordinator CE, NHIS 費城地區辦公室
	特殊普查計畫（Special Census Program）概述	Robert Stabs, Decennial Specialist 費城地區辦公室

本報告分為四章，除本章外，第二章為地址母體檔建置及維護，第三章為差分隱私作業及隱私保護技術，第四章為心得與建議。本報告內容資料來源為研習單位提供之參考文件、普查局官方網站公開之文件、統計圖表與圖檔，以及普查局官方社群媒體公開之圖片。

第二章　地址母體檔（MAF/TIGER）的建置及維護

第一節 地址母體檔的發展

美國地址母體檔（MAF/TIGER）係整合「地址主檔（Master Address File, MAF）」資料庫、「拓撲整合地理編碼和參照系統（Topologically Integrated Geographic Encoding and Referencing, TIGER）」，而建立的一個完整且準確的地址框架，成為人口普查和相關調查計畫的母體資料庫。

一、MAF/TIGER 資料庫概述

美國普查局的地址主檔（MAF）是一個動態的資料庫，為美國社區調查（American Community Survey, ACS）和人口普查的主要地址來源，涵蓋美國和波多黎各的已知住宅單位（Housing Units, HUs）、集體住所（Group Quarters, GQs）及部分非住宅單位的地址清單，主要用於管理地址資訊，每筆紀錄包含門牌號碼、街道名稱、城市、州和郵遞區號、地理編碼、單元的物理特徵、與其他單元的關係、住宅或商業狀態、緯度和經度坐標，以及反映資料更新來源和歷史的異動紀錄。

拓撲整合地理編碼和參照系統（TIGER）是由美國普查局開發的全國性數位空間資料庫，用於所有空間、地理和住宅地址資料的國家儲存庫。與大多數地理資訊系統（Geographic Information System, GIS）資料庫不同，係專為維護地理單位間一致的關係而設計，包含點位特徵[1]（Point Features）、線性特徵[2]（Linear Features）、區域特徵[3]（Areal Features），提供詳細的地理描述，對於統計數據的蒐集和彙整相當重要。

將 MAF 每個地址與 TIGER 資料庫中的地理位置連結後，即可在地圖上進行定位，並標註居住單元的類型，如住宅、集體住所或臨時居所。兩者會定期更新，以反映真實世界中地址和地理特徵的變化。截至 2024 年春季，MTdb（MAF/TIGER database）已包含了 289,394,149 個地址紀錄、195,569,377 個 MAF 單元（MAF UNIT）及 134,403,571 個 MAF 結構點（MAF Structure Points, MSPs），顯示 MTdb 龐大的資料量。

---

[1] 如自然地形特徵、人造結構及居住（機構）場所等。
[2] 如道路、河流、海岸線及鐵路等。
[3] 如行政區界線（legal boundaries）、普查區塊（block）及水域等。

二、MAF/TIGER 系統的演進

在 1970 年以前，人口普查的地址編列及資料蒐集是以一體化作業（all-in-one operation）的方式進行的，編列員（Listers）在編列地址的同時蒐集調查資料；1970 年人口普查改變辦理方式，首次在蒐集資料前先建立地址清單，在編列地址作業完成後，才將問卷郵寄到地址進行資料蒐集；到了 1990 年人口普查使用的住宅地址清單來自「地址控制檔案（Address Control File）」，由地址清單和地圖所組成，主要資料來源是透過美國郵政服務（United States Postal Service, USPS）的遞送順序檔案（Delivery Sequence File, DSF）定期更新，這些地址清單即是地址主檔（MAF）的前身。

到 1990 年代後期，為辦理 2000 年的人口普查，美國普查局建立了 MAF 的初步版本，當時 MAF 和 TIGER 資料庫是各自獨立的資料庫，具有不同的軟體、法律價值和架構。2000 年人口普查後，為確保人口普查數據在正確的地理位置進行處理和統計，地理部門（Geography Division, GEO）制定了 1 項計畫，將這 2 個資料庫整合為 1 個，稱為 MAF/TIGER 系統（MAF/TIGER System, MTS）。此外還引入了相關改進措施，包括「人口普查地址在地更新（Local Update of Census Addresses, LUCA）」、「區塊清查（Block Canvassing）」以及「地址編列（Address Listing）」等更新和驗證地址主檔的作業，允許部落、州和地方政府審查和改進美國普查局的地址清單、普查局員工走訪選定的區域，驗證地址清單的準確性和完整性，並且新增未列入清單的新地址，大幅提高了地址母體檔的資料量及可靠性。

美國地址母體檔的維護和更新是一個不間斷的過程，隨著時間推移，系統也經歷了重大的變化。2000 年人口普查後，按月辦理美國社區調查（ACS）取代原本人口普查中用於蒐集詳細人口和住房特徵的「長表格」，成為一項持續辦理的家戶面調查。為滿足使用者對最新地址資料的需求，普查局採取多層次的更新方法，包括每年至少 2 次使用來自美國郵政服務（USPS）的遞送順序檔案（DSF）來更新地址主檔。然而，對於 DSF 覆蓋不足的地區，普查局藉由「社區地址更新系統（Community Address Updating System, CAUS）」，集中處理非城市地址密度高或住宅單位（HU）可能增長的普查區塊，透過實地驗證提供準確的地址更新，為人口普查及相關調查提供可靠的地址框架。

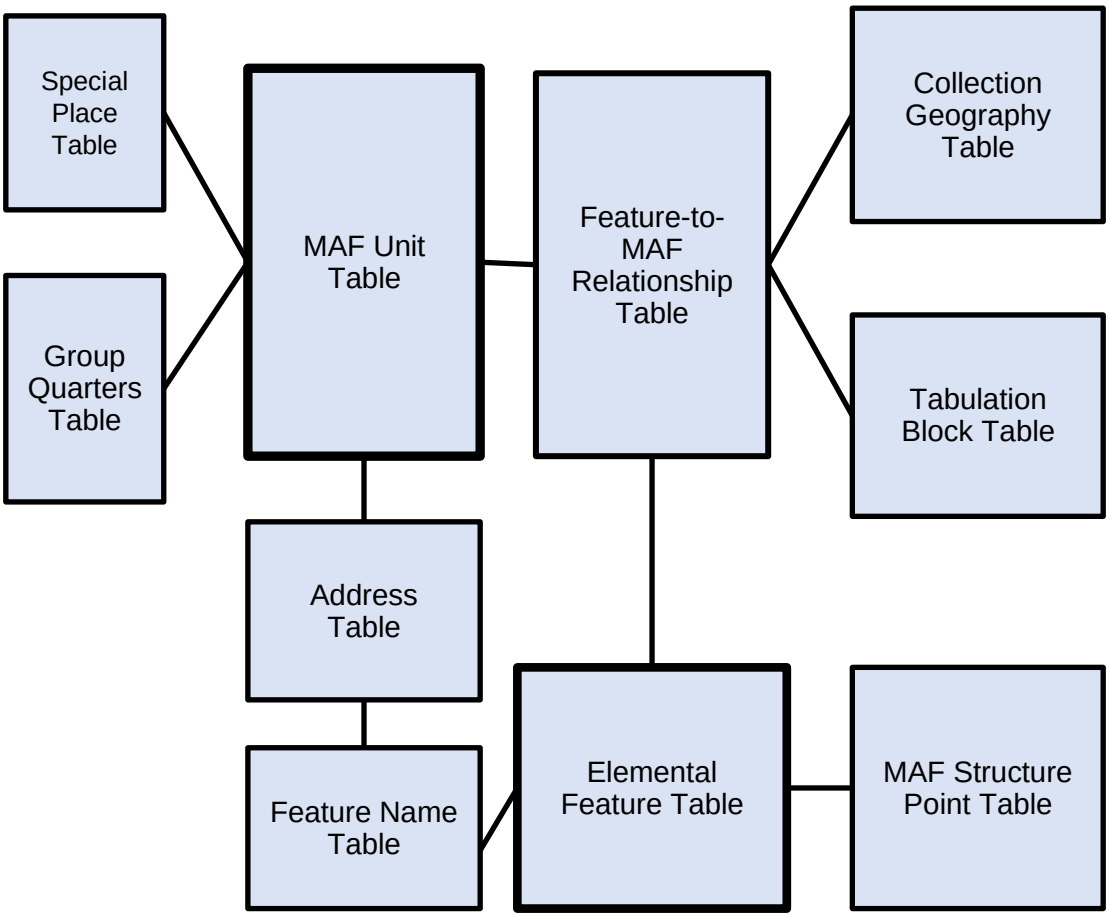
2010 年人口普查之後，普查局為支援各類調查和人口普查，除繼續使用 DSF 和 CAUS 更新 MAF/TIGER 系統之外，加入地理空間支援系統合作夥伴計畫（Geographic Support System Initiative, GSS-I）與地址清查（Address Canvassing, ADC）共同作業，以提供準確及完整的地址、地理特徵和邊界。

第二節 MAF 資料庫架構及有效地址篩選

一、MAF/TIGER 系統的資料庫架構

MAF/TIGER 系統（MTS）包括應用軟體和資料庫，是全國空間、地理和住宅地址資料的儲存庫，用於人口普查和調查資料的蒐集、統計、發布、地理編碼以及地圖製作等。在 MTS 中的個別紀錄稱為 MAF 單元（MAF Unit），與其關聯的資訊分別儲存在其他關聯表中（圖 2-1）：

圖 2-1　MAF and TIGER databases



Source: U.S. Census Bureau.

- MAF UNIT：代表個別紀錄的唯一 ID。表示單元分類（如住宅單位（HU）、集體住所（GQ）、非住宅單位等）、最新紀錄狀態（有效的 HU/GQ、重複、不適居住等），以及 2000 年、2010 年、2020 年是否被列入人口普查的歷史數據。
- ADDRESS（地址）：包含實際的地址資訊（地址/門牌號、單元號、USPS 郵遞區號）以及唯一 ID，允許 ADDRESS 紀錄串接 MAF UNIT 紀錄和

FEATURE NAME（特徵名稱）紀錄。

- GROUP QUARTERS（集體住所）：包含集體住所/臨時地點 MAF UNIT 的特定資訊，例如集體住所/臨時地點的名稱、聯絡資訊，以及一些關於容量（床位數等）和唯一 ID 的數據。
- SPECIAL PLACE（特殊場所）：允許將多個集體住所 MAF UNIT 關連到一個總體的「設施」MAF UNIT。例如，大學宿舍有多個 MAF UNIT 可關連到該大學行政辦公室的 MAF UNIT。
- FEATURE NAME（特徵名稱）：儲存唯一的街道名稱（特徵），這些名稱被 ADDRESS 表和存儲空間特徵的表引用。
- FEATURE TO MAFUNIT RELATIONSHIP（特徵與 MAF UNIT 關係）：包含唯一 ID，用於關連 MAF UNIT 與 TIGER 中的空間特徵，如建築結構點、統計區塊（Tabulation Blocks）和收集區塊（Collection Blocks）之間的串接。
- ELEMENTAL FEATURE TABLE（基本特徵表）：包含代表點、線和多邊形特徵的幾何紀錄，以及關於單一個特徵的詮釋資料（metadata）。
- MAF STRUCTURE POINT（建築結構點）：包含結構點坐標（緯度/經度）的唯一紀錄。
- TABULATION BLOCK TABLE（統計區塊表）：包含描述統計區塊屬性的資訊，例如面積、當前住宅單位數量（來自 MAF）、最近一次人口普查的人口數量以及多邊形的幾何形狀。
- COLLECTION BLOCK TABLE（蒐集區塊表）：包含描述蒐集區塊屬性的資訊，例如面積、當前住宅單位數量（來自 MAF）、最近一次人口普查的人口數量以及多邊形的幾何形狀。
- 「蒐集區塊」與「統計區塊」主要區別在於，統計區塊必須遵循司法管轄區（不可見邊界），而蒐集區塊通常以可見特徵（街道、水體、鐵路等）劃定，便於實地工作人員使用可見邊界進行工作分配。

二、有效地址篩選原則

因應不同的需求，篩選地址主檔單元有不同的過濾標準，例如普查 10 年 1 次的編列與美國社區調查（ACS）會使用不同的過濾標準。一般過濾的條件會著重於篩選地址檔案單元的屬性，包含地理編碼、紀錄類型（住宅單位、集體住所、臨時地點、非住宅）、地址類型（都市型、非都市型）紀錄來源、過去的更新作業（現場操作、USPS、當地合作夥伴），並排除特定紀錄（家庭暴力庇護所、國土安全）。

為了確保調查底冊盡可能完整，首先會設定有效地址的定義，並據此設定條件篩選 MTdb（MAF/TIGER database）中的地址，以美國社區調查（ACS）為例，為了完整地涵蓋住宅單位和人口，因此納入各種類型的地址，如新建單元、未地理編碼的單元，以及「排除在遞送統計之外（excluded from delivery statistics, EDS）」單元，再對單元設定條件篩選地址。

第三節 MAF/TIGER 資料更新

一、MAF/TIGER 資料來源與更新機制

為了確保 MAF/TIGER 系統資料的準確性和完整性，會在 2 次普查的 10 年間持續更新及驗證，美國普查局採用了多種資料來源和更新方法，以確保每 1 個地址都在正確的位置。

美國普查局主要依靠 4 個重要來源進行資料更新，首先是美國郵政服務（USPS）提供的遞送順序檔案（DSF），每半年更新 1 次城市的地址資訊，透過自動化方式，依據門牌號碼、街道名稱、郵遞區號和結構內識別碼，將配對成功之資料識別為相同地址，視為重複紀錄；無法配對者則視為一筆新紀錄，每年增加約 100 萬個新地址。但 DSF 不會用於更新非城市的地址，因為其格式非正規化。

其次是地理空間支援系統合作夥伴計畫（GSS-I），藉由與部落、州和地方政府的合作，蒐集並更新地址和道路數據。第 3 個更新機制是人口普查地址在地更新（LUCA），由各級政府參與審查並提供地址的資訊，2020 年所推動的 LUCA 計畫新增了超過 300 萬個地址。最後是地址清查（Address Canvassing）作業，透過「辦公室內地址清查（In-Office Address Canvassing, IOAC）」及「實地地址清查（In-Field Address Canvassing, IFAC）」進行地址清單的最後確認。

二、MAF/TIGER 系統資料驗證及更新流程

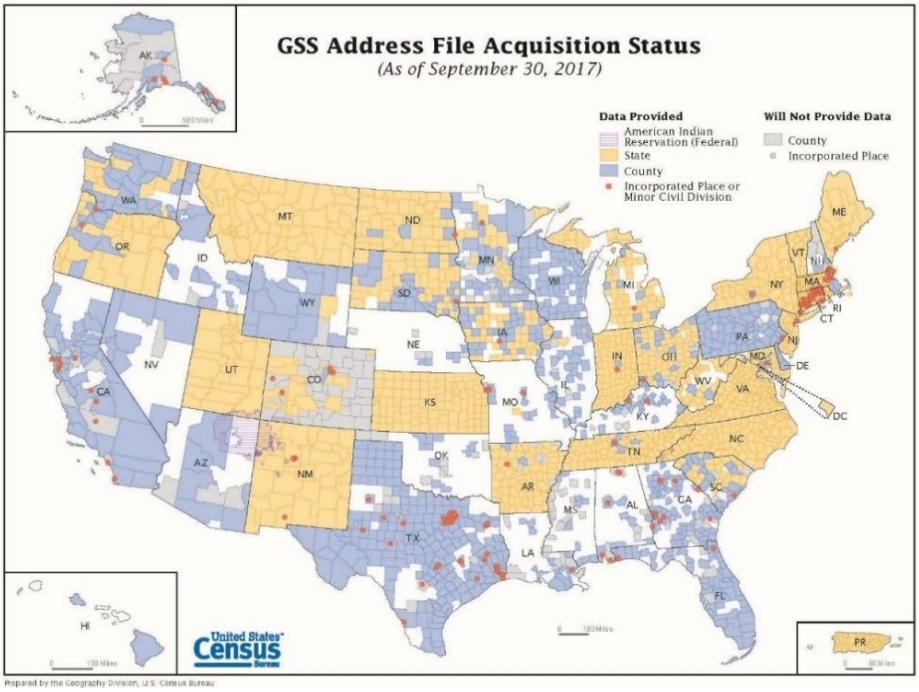
（一）遞送順序檔案（DSF）更新

美國郵政服務（USPS）每年於春、秋兩季提供美國普查局遞送順序檔案（DSF），與地址主檔（MAF）進行比對後，將修正或新增地址更新至 MAF，普查局每半年新增約 50 萬個地址到 MAF 中。此外，利用 USPS 區域改進計畫（ZIP）代碼檔案、可定位地址轉換服務（Locatable Address Conversion System, LACS）檔案，更新區域代碼及進行地址轉換。

（二）地理空間支援系統（GSS）更新

首先由地理空間支援系統（GSS）合作夥伴（部落、州和地方政府）提供最新的地址和道路資訊，地理部門（GEO）就其所提供資料進行地址來源評估（Address Source Evaluation, ASE）是否符合地址和道路資料的內容規範，再透過自動化檢查刪除重複地址，驗證新地址是否實際存在，確認後進行資料整合並更新地址母體資料庫。截至 2017 年 9 月美國普查局已從 1,459 個合作夥伴處取得 100,690,733 筆地址（圖 2-2），其中近 80%地址與母體資料庫比對成功，其中無地址編碼者得以進行註記，並新增 292,471 筆（占 0.3%）新地址，亦有 20%的地址無法成功配對。透過地理空間支援系統合作夥伴計畫（GSS-I）減少地址錯誤和重複，可提高人口普查和調查計畫的效率，從而降低整體成本。

圖 2-2　地理空間支援系統合作夥伴計畫提供的資料地區分布



Source: U.S. Census Bureau.

（三）人口普查地址在地更新（LUCA）

人口普查的結果將影響國會席次分配、聯邦資金分配以及社區規劃，對於部落、州和地方政府的影響重大，因此推動「人口普查地址在地更新（Local Update of Census Addresses, LUCA）」，這是一項依法授權的計畫，讓部落、州和地方政府利用在地知識來驗證和改進普查局的地址清單。參與者會收到包含來自相關審查資料，審查後將更新內容回傳，其中通過 LUCA 驗證的地址，會用於更新 MAF/TIGER。

LUCA 更新方式分為數位和紙本兩種，數位方式的地址清單包含地址、公寓/單元號碼、城市郵寄郵遞區號、集體住所名稱、設施名稱、位置描述、非城市郵寄地址、地址用途、結構經緯度和城市地址標記等資訊，可以在地理更新合作軟體（Geographic Update Partnership Software, GUPS）中輸入更正後的資訊來更新資料。紙本方式的地址清單則包含街道名稱、房屋號碼、單元號碼、普查區號碼、普查區塊號碼以及集體住所標記等資訊。

（四）地址清查（Address Canvassing, ADC）

過去美國普查局為了測量地址清單覆蓋率、更新地址清單等，進行了 MAF 覆蓋率研究（MAF Coverage Study, MAFCS），另為相關家戶面調查（如美國社區調查（ACS）、當前人口調查（Current Population Survey, CPS）、國民健康訪問調查（National Health Interview Survey, NHIS）、收入和計畫參與調查（Survey of Income and Program Participation, SIPP）等）母體名冊更新，辦理「人口統計區域地址編列（Demographic Area Address Listing, DAAL）」，惟隨相關調查改用 MAF 進行抽樣設計，並考量經費的不確定性，MAFCS 及 DAAL 皆已停止辦理。

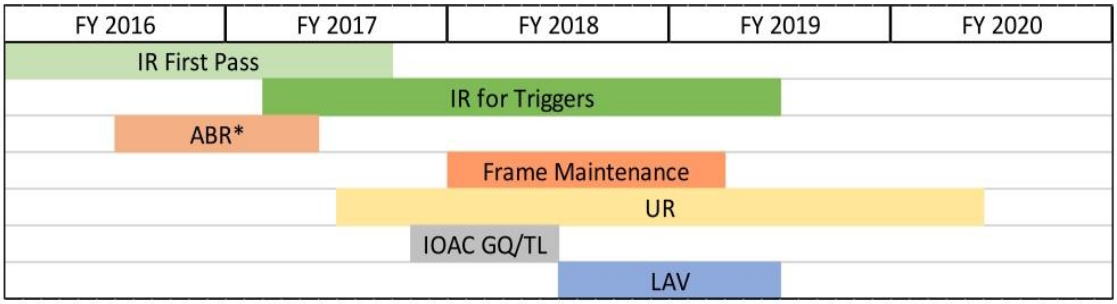
目前地址清查（ADC）作業是透過「辦公室內地址清查（In-Office Address Canvassing, IOAC）」及「實地地址清查（In-Field Address Canvassing, IFAC）」來建立一份完整、準確的地址清單和空間資料庫。IOAC 是一個持續的過程，主要運用影像數據、合作夥伴數據和其他數據源，來識別 MAF/TIGER 系統中資料的錯誤或遺漏，並進行驗證或更新；當 IOAC 無法驗證和更新時，則需要進一步 IFAC，清查後的結果將會被發送到 MAF/TIGER 系統，進行資料庫更新作業。

第四節 地址清查作業

一、辦公室內地址清查（In-Office Address Canvassing, IOAC）

IOAC 是辦理 2020 年普查首次採用的流程，工作人員在辦公室內利用衛星和航空影像、地理資訊系統（Geographic Information System, GIS） 查看器及第三方資料等高品質影像，檢視住宅和非住宅景觀。另運用美國郵政服務（USPS）的資料偵測變化，確保母體資料的完整及正確，大幅減少需要實地清查的地址數量。整個辦公室內清查流程包括互動式審查（Interactive Review, IR）、活躍區塊解析（Active Block Resolution, ABR）、未編碼解析（Ungeocoded Resolution, UR）、集體住所／臨時地點審查（Group Quarters/Transitory Locations, GQ/TL）及 LUCA 地址驗證（LUCA Address Validation, LAV）等（圖 2-3）。

圖 2-3　辦公室內地址清查時序

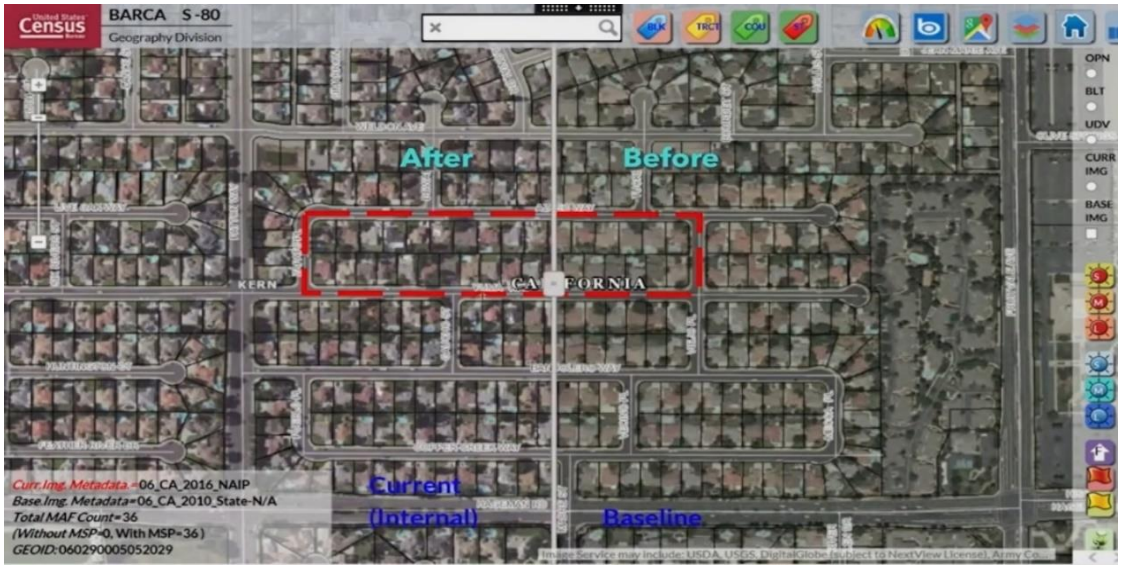


* ABR was discontinued in February 2017, but staff continued to QC the ABR records worked until September 2017.
Source: U.S. Census Bureau.

（一）互動式審查（Interactive Review, IR）

互動式審查（IR）係屬影像審查，是辦公室內地址清查流程中的首站，用於評估當前影像和基準影像之間的變化。審查員以「區塊評估、研究和分類應用程式（Block, Assessment, Research, and Classification Application, BARCA）」為審查工具，如圖 2-4 顯示中線右側為基準影像（2010 年），左側則為當前影像，可以依審查員需求左右移動來比較影像變化。另外運用 MAF 的住宅單位計數、美國郵政服務（USPS）資料、地方/縣/州地理資訊系統（GIS）檢視器以及協力廠商資料，來識別自 2010 年人口普查以來的住宅景觀變化，並判斷區塊是否穩定，從而決定是否需要地址或地理空間更新，或需要進一步的地址清查工作。

圖 2-4　互動式審查工具 BARCA 畫面



Source: U.S. Census Bureau.

透過比較當前影像與 MAF 的「住宅單位計數」，可以將區塊（Block）標記為幾種狀態：

- 地址覆蓋不足（MAF undercoverage）：MAF 列出的地址比圖像中顯示的住房單位少。
- 地址覆蓋過度（MAF overcoverage）：MAF 列出的地址比圖像中顯示的住房單位多。
- 特徵缺失（Missing features）：圖像中顯示的道路在 MAF/TIGER 中不存在。
- 特徵不吻合（Misaligned features）：道路、區塊邊界與圖像不匹配。

另一方面也透過觀察影像上的「建物類型」將區塊分為幾種類型：

- 建成區塊（Built-out）：完全被「住宅結構（residential structures）」覆蓋或被「住宅結構和非住宅土地用途（nonresidential land use）」混合覆蓋之區塊。
- 開放區塊（Open）：有增長空間的區塊。
- 不可開發區塊（Undevelopable）：完全被「非住宅土地用途」覆蓋且不太可能開發之區塊。
- 圖像質量差（Poor Imagery）：由於解析度差、雲層覆蓋或圖像缺失的不完整區塊，導致無法判讀。

區塊評估、研究和分類應用程式（BARCA）會基於上述區塊的審查結果，將每個區塊分類為消極（Passive）、活躍（Active）或暫停（On Hold）三種辦公室內地址清查操作狀態：

- 消極區塊（Passive Blocks）：當前影像中的住宅單位數量與 MAF 數據一致，並且沒有住宅增長或減少的跡象，因此消極區塊不會被分配到現場實地審查。
- 活躍區塊（Active Blocks）：當前影像中的住宅數量位數與 MAF 數據不同（即地址覆蓋不足或地址覆蓋過度），或者當前影像和基準影像之間住宅景觀發生變化，可能為住宅單位的建築物增加（圖 2-5）或減少（圖 2-6），則需要進一步透過「活躍區塊解析（Active Block Resolution, ABR）」流程處理。
- 暫停區塊（On Hold Blocks）：當影像質量不佳（例如景觀受雲層遮擋），無法完成審查，則需要取得額外資訊或更好的影像再進行處理。

圖 2-5　比對當前影像及基準影像顯示建物增加

Source: U.S. Census Bureau.

圖 2-6　比對當前影像及基準影像顯示建物減少

Source: U.S. Census Bureau.

然而由於住宅單位的變化是動態的且在影像中未必可見，在 IR 的審查結束後，區塊的狀態（消極、活躍或暫停）可能會隨著景觀中住宅單位或地址底冊的變化而改變。因此普查局採用一個稱為「觸發器（Trigger）」的流程來識別可能需要重新審查的區塊，例如出現新的或更高解析度的影像，或相關地址檔案更新致地址數量出現改變，顯示可能需要更改區塊分類狀態，則重新進行 IR（IR for Trigger）。首次在 2020 年普查中進行互動審查的 11,155,486 個區塊中，有 13.5%被觸發，其中 73.4%的區塊更新了 IR 狀態（表 2-1）。

表 2-1　互動式審查處理結果

Total number of blocks in IR: 11,155,486 blocks				
Of the 11,155,486 blocks, 13.5% were triggered				
Of the 13.5% blocks triggered	No change in status	*26.6%*		
	Received an updated IOAC IR status	*73.4%*		
	Of the 73.4% blocks that received an updated IR status	Active blocks became passive	*12.2%*	
		Passive blocks became active	*10.9%*	
		Hold for imagery blocks rated active or passive	*53.8%*	
		Of the 53.8% hold for imagery blocks	Became active	*45.5%*
			Became passive	*54.5%*
		Were retriggered	*23.1%*	

Source: U.S. Census Bureau.

全國首次互動式審查，有 16%為活躍區塊（其中 6.4%為覆蓋不足），經過互動式審查反覆處理後，活躍區塊比例降至 9.8%，覆蓋不足率亦降至 6.0%；暫停區塊亦由 12.2%降至 1.6%（表 2-2），大幅減少實地清查作業的數量。

表 2-2　互動式審查前後資料處理結果

	IR "First Pass"		End of IR	
	Percent of Blocks	Percent of Housing Units in the Blocks	Percent of Blocks	Percent of Housing Units in the Blocks
Active	16.0%	27.3%	9.8%	21.6%
Undercoverage	6.4%	11.1%	6.0%	8.9%
Passive	71.8%	55.7%	87.0%	70.1%
On-Hold for Imagery	12.2%	17.0%	1.6%	2.8%
Triggered	n/a	n/a	1.6%	5.5%
Cumulative Number of Block Reviews Conducted by IR Staff				14,360,000

Source: U.S. Census Bureau.

（二）活躍區塊解析（Active Block Resolution, ABR）

活躍區塊解析（ABR）係為解決互動式審查（IR）程序中被識別為「活躍區塊」中的覆蓋問題，審查員使用多個來源重新解析，包括最新的美國郵政服務（USPS）遞送順序檔案（DSF）、圖像、本地地理資訊系統（GIS）或具有街道影像的網絡地圖工具（Google、Bing 地圖）等。解析後「活躍區塊」可能改為「消極區塊」，即屬不需要後續實地地址清查的「已解決（resolved）」狀態；反之，無法在 ABR 解決覆蓋問題，則標示為「未解決（unresolved）」狀態，該區塊就成為實地地址清查的候選區塊。

ABR 最初估計工作量為 170 萬個區塊，由於此法缺乏效率，且受預算限制，ABR 在處理了約 74,500 個區塊後於 2017 年 2 月停止辦理。經 ABR 處理後的區塊「已解決」占逾三分之二，狀態從「活躍」改為「消極」；只剩不到三分之一的「未解決」區塊保持「活躍」狀態。

在 ABR 被中止後，啟動了「抽樣底冊維護（Frame Maintenance）」計畫（2017 年 10 月~2018 年 12 月），目的是利用 ABR 的經驗來維護人口普查母體。與 ABR 不同之處在於這項計畫採用「圖釘標示法（pin-based approach）」，專注於活躍區塊中特定的地址和特徵，此法僅關注區塊中的特定問題（而不是像 ABR 檢查整個區塊）來減少驗證操作，並優先處理可以在辦公室內透過資料就得到解決的區塊，例如覆蓋過度（母體檔的地址比圖像中顯示的住房單位多）較不明顯的區塊、可以透過修正底層地理特徵來解決覆蓋問題的區塊，及需要長途跋涉才能到達的區塊，在沒有可靠來源的情況下，工作人員不會改變其狀態，仍將區塊保留為未解析狀態。

本計畫共處理 2.6 萬個區塊，其中有 79.7%修正為消極區塊，會被用於更新 MAF；另有 20.3%未解決區塊仍然是實地地址清查的候選區塊。以地址數量觀察 2.6 萬個區塊中包含 23.7 萬個地址，最後僅 1.2%的地址無法更新 MAF（表 2-3）。

表 2-3　抽樣底冊維護計畫處理之活躍區塊地址數

Address Outcomes	Number of Addresses	Percentage of Addresses Worked
Addresses Worked** (Including adds and address moved into the block)	237,000	100%
Addresses Verified ("V" or "verify" action)		10.7%
Addresses Updated (Actions other than "V" that updated the MAF)		88.2%
Addresses Rejected (Reject for MAF update)		1.2%

Source: U.S. Census Bureau.

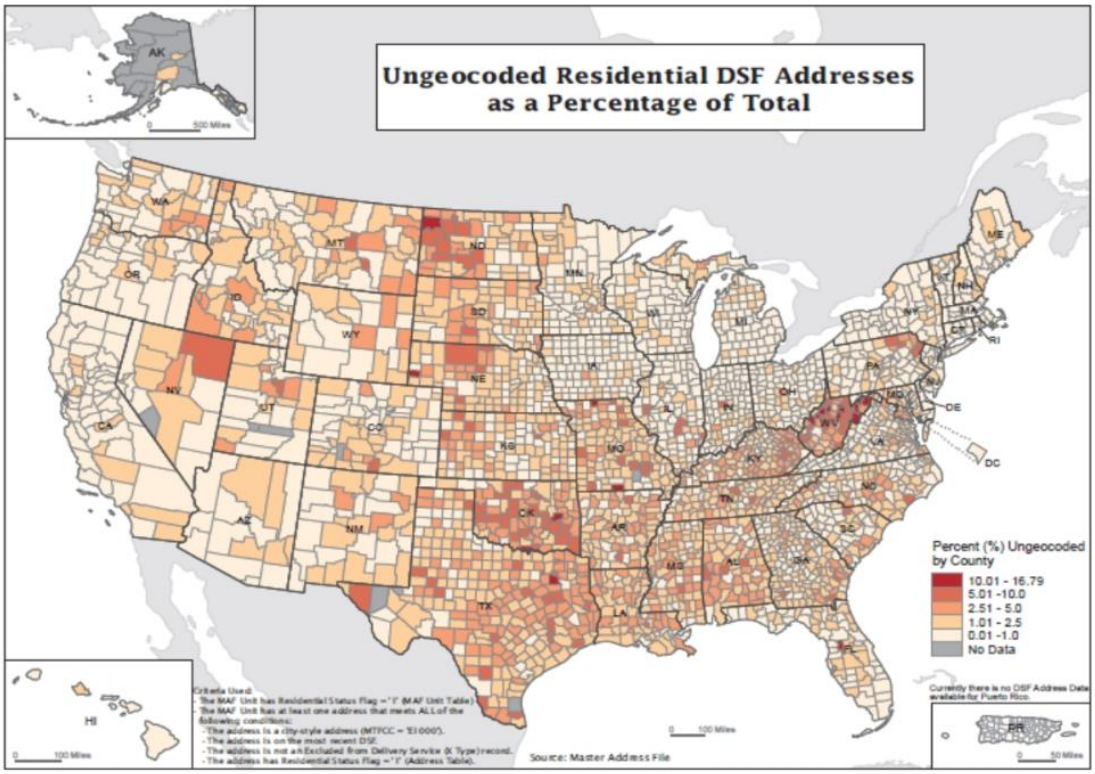
（三）未地理編碼解析（Ungeocoded Resolution, UR）

未地理編碼解析（UR）也是辦公室內地址清查（IOAC）的一個流程，旨在為地址主檔（MAF）中缺少地理編碼（亦即區塊位置）的住宅地址分配區塊位置。主要由國家處理中心（National Processing Center, NPC）的製圖技術員完成，這些工作人員使用配對和編碼系統（Matching and Coding System, MaCS）、地理資訊系統（GIS）檢視器確定每一個地址的地理位置並記錄解析結果。

MAF/TIGER 系統中未地理編碼地址（圖 2-7）主要是因為人口增長所帶來的新住宅開發，這些地址無法由自動地理編碼流程分配到人口普查區塊，主要有以下幾個原因：

- TIGER 中缺少街道特徵，或名稱不正確。
- TIGER 中缺少全部或部分地址範圍，或地址範圍不正確。
- TIGER 中的全部或部分街道特徵資訊（例如街道名稱、郵遞區號）與 MAF 上地址的表示方式不一致。
- 地址雖然在 MAF/TIGER 中，但 MAF/TIGER 地址和美國郵政服務（USPS）的遞送順序檔案（DSF）地址有差異，導致無法以自動流程比對。

圖 2-7　各地區未地理編碼比率



Source: U.S. Census Bureau.

自動地理編碼解析程式會嘗試比對新地址與 MAF/TIGER 系統中具有坐標的地址，或比對 TIGER 中的地址範圍，以取得人口普查區塊代碼。無法自動比對到區塊代碼的地址，將進一步拆分為工作單元分配給分析師處理。分析師首先確定每個地址的地理位置，並透過新增道路和地址範圍、為現有道路添加地址範圍、添加或更正與地址範圍關聯的郵遞區號，以及更正或添加道路名稱等方式更新 TIGER 系統。

在 2020 年人口普查之前，UR 工作小組總共處理了約 285.5 萬個不同的地址紀錄，超過 180 萬個地址成功進行地理編碼。除了對地址進行地理編碼之外，還對 MAF/TIGER 資料庫中的線性特徵（例如道路）做出重大改進，包括添加 38.1 萬餘個線性特徵，並在資料庫中更新了超過 9.6 萬哩的道路。

（四）集體住所／臨時地點（Group Quarters/Transitory Locations, GQ/TL）審查

集體住所／臨時地點（GQ/TL）審查是使用電腦輔助電話訪談（Computer-Assisted Telephone Interviewing, CATI）系統識別、更新和驗證地址主檔中集體住所（GQ）和臨時地點（TL）的地址。最初，GQ/TL 審查的範圍僅限於在活躍區塊解析（ABR）期間所添加的 GQ 或 TL，因為 ABR 工作人員在查核活躍區塊時會添加新的或疑似的 GQ、TL；但是 ABR 在 2017 年初停止使用時，導致 GQ/TL 審查範圍更改為透過電話聯絡或網路查詢 MAF 中的所有 GQ 和 TL，工作量從 2.5 萬個地址大幅增加到 21.2 萬個地址（表 2-4），除了確認目前的 MAF 中的 GQ 或 TL 仍然存在，並盡可能取得 GQ 或 TL 的聯絡資訊。

過程中工作人員發現，透過電話聯絡 GQ 和 TL 比預期的還要困難，例如受訪者常質疑電話的真實性而不願意提供資訊，在完成訪問介紹之前就掛斷電話。即使能夠取得聯繫，一些聯絡人也會表示他們最近已經接受「美國社區調查（ACS）」的訪問，而不願意再次提供資訊，或是擔心違反健康資訊保密法，無法提供資料；另外有些受訪者只會說特定語言，形成語言障礙，或者工作人員自身的專業知識不足，都對 GQ/TL 審查工作造成挑戰。因此 GQ/TL 也在 2018 年 3 月暫停，當時只有 4.4%的 GQ 和 TL 範圍完成了品質控管（表 2-4）。後來美國普查局即將蒐集 GQ 和 TL 資訊的任務改由其他作業完成，例如「服務型場所作業」、「人口普查地址在地更新（LUCA）」和「GQ 事先聯絡作業」。

表 2-4　集體住所／臨時地點（GQ/TL）審查處理結果

Total IOAC GQ/TL Universe: 212,000	Percent of Total Universe	Percent of Total Completed Review
Total Completed Review	12.1%	-
Total Referred	3.1%	25.8%
Total On-Hold	8.0%	66.0%
Total Completed Quality Control	4.4%	35.9%

Source: U.S. Census Bureau.

（五）LUCA 地址驗證（LUCA Address Validation, LAV）

LUCA 地址驗證（LAV）工作主要在審查部落、州和地方政府提交的 LUCA 地址，這些地址未通過與地址主檔（MAF）的自動比對，以及其他指定的紀錄。LAV 的目標是驗證 LUCA 參與者提交的地址是否存在，以及提供的 MAF 版本在空間上是否更準確，並且確認地址紀錄是否屬於 LUCA 參與者或 MAF 指定以外的區塊。

由於 2020 年的 LUCA 參與者提交了超過 2,200 萬個地址，數量是預期的 4 倍，因此地址驗證採用「實體層級抽樣（entity level sampling）」方法，以確保能在既定的時程內，完成所有 LUCA 參與者提交資料的審查。其作業方式係將符合資格的檔案（紀錄在 200 筆以上），由系統隨機選取該檔 20%的紀錄，進行地址驗證的審查工作，只要通過審核之比率達到 80%的門檻，剩餘的紀錄將被暫時接受；如果通過審核之比率未達到 80%的門檻，所有剩餘的紀錄將被拒絕。如果一個實體檔案的紀錄少於 200 筆，則進行全面審查。

LAV 審查員會使用 LAV 生產模組來驗證地址紀錄是否存在，以 LUCA 參與者提供的地址紀錄、地理編碼與 MAF/TIGER（如果存在）進行比對，經過判斷審查員可以採取幾種行動：

- 接受地址紀錄和參與者分配的地理編碼。
- 接受地址紀錄，但將其分配給不同的地理編碼。
- 手動將地址紀錄比對到 MAF 中已存在的紀錄。
- 拒絕地址紀錄。

如果 LAV 審查員確定紀錄是非住宅、不適合居住、位於參與者管轄範圍之外、位於無法開發的位置或不存在，則會拒絕該紀錄。審查員必須有來自可靠且具體的證據才能拒絕該地址紀錄，如果存在任何模糊之處，則應接受該紀錄。整體而言 LAV 審查的紀錄中，通過者占 61%，未通過者占 39%（表 2-5）。

表 2-5　LUCA 地址驗證（LAV）處理結果

Description	Code	Accept/ Reject	Percent of Accepted or Rejected	Percent of Total
Total Number of LAV Records: 861,000				
Address Validated	A	Accept	59.8%	36.6%
Manual Match	L	Accept	6.7%	4.1%
Move	M	Accept	6.7%	4.1%
Provisional Add	P	Accept	26.9%	16.5%
	Total Accepted			**61.3%**
Address Rejected	R	Reject	76.8%	29.7%
Nonresidential	N	Reject	23.1%	8.9%
Outside of Jurisdiction	O	Reject	0.1%	0.0%
Uninhabitable	U	Reject	0.0%	0.0%
	Total Rejected			**38.7%**

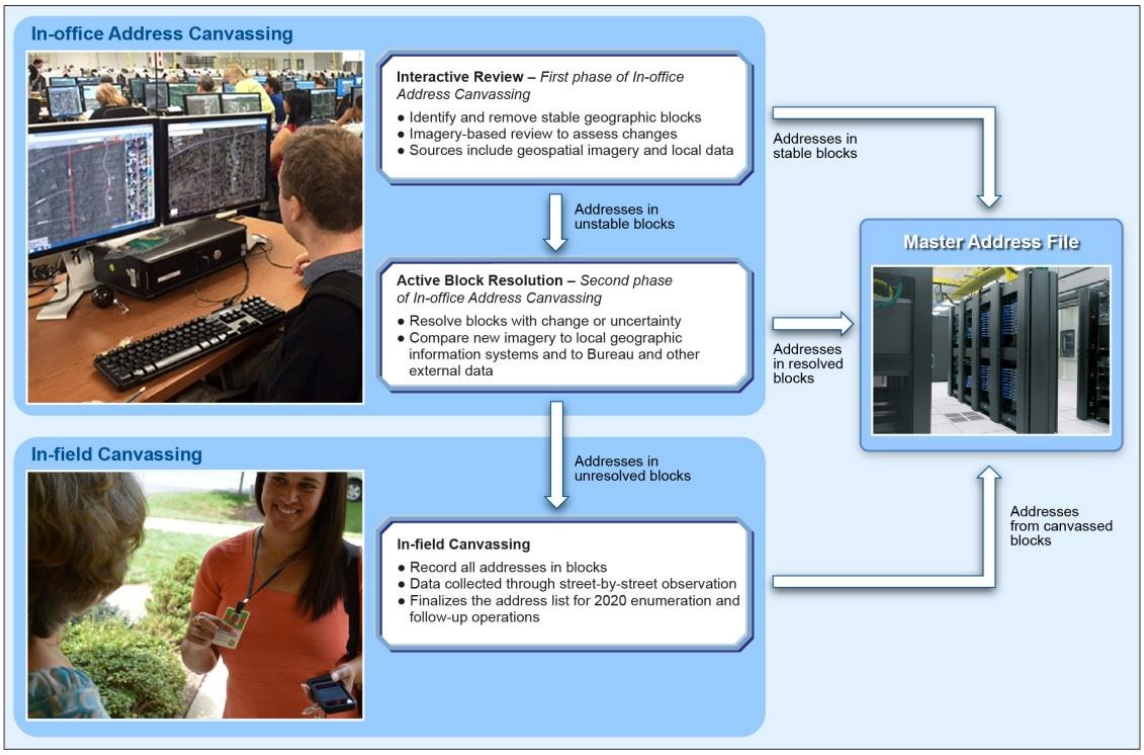
Source: U.S. Census Bureau.

辦公室內地址清查（IOAC）作業處理了大範圍的地址，減少需要實地清查的地址數量，並提高地址主檔的整體品質，2020 年普查 IOAC 成功地將需要實地地址清查（IFAC）的地址數量從 1.4 億個減至 5,004 萬個，地址清單中有 65% 可在辦公室內完成審查，真正需要實地清查者僅 35%，大幅減省人力及成本。

二、實地地址清查（In-Field Address Canvassing, IFAC）

在辦公室完成地址清查後，被標記為「已解決」者用於更新地址母體資料庫，而被標記為「未解決」者，包含「增長、下降、覆蓋過度或覆蓋不足」等在 IOAC 無法確認的區塊，將進一步進行實地地址清查（IFAC）（圖 2-8）。IFAC 是地址清查作業的最後一個步驟，編列員（Listers）須依照被分配的基本蒐集單元（Basic Collection Unit, BCU），識別人們可能居住或停留的每個地點。

圖 2-8　IOAC 結果被標記為「未解決」的地址被指派給 IFAC



Source: United States Government Accountability Office

（一）實地清查地址工作量

在街區通過辦公室內地址清查（IOAC）後，開始進行實地清查作業的規劃及準備工作，首先美國普查局確定地址查驗範圍，這些區域被劃分為基本蒐集單元（BCU），並指派給編列員在一定期限內完成地址查驗工作，工作量會以地理區域叢集的方式指派，以提高作業效率並降低成本。現場管理人員則負責監督編列員的進度、核准時程表並擷取編列員互動數據。

（二）實地資料蒐集作業

2020 年人口普查的實地地址清查 （IFAC）作業中，編列員會攜帶筆記型電腦，並將行動案例管理（Mobile Case Management, MCM）系統和 LiMA（Listing

and Mapping Application）安裝於設備，編列員對分配到的區塊進行現場資料蒐集（圖 2-9）。MCM 系統可以提供案例管理和儀表板，並進行存取、檢視和傳輸作業，另使用 LiMA 來驗證和更新地址資訊。筆電的全球定位系統（Global Positioning System, GPS）可以顯示「您在此處」指示器（“you are here” indicator , YAHI），以便編列員確認自身所處的位置，並可蒐集住宅單位準確的位置坐標。編列員將路面上的實際狀態與 LiMA 上呈現的地址清單、地圖比較後進行更新。

LiMA 係美國普查局開發的應用程式，主要供編列員實地清查地址時使用，必須搭配全球定位系統（GPS）及 Windows 作業系統環境。地理部門（GEO）每半年將 MAF 資料及 MAF/TIGER 資料庫中包含的空間資料子集放置在 LiMA 的 Linux 檔案伺服器上，然後由編列員自 LiMA 使用者端通過虛擬私人網路（Virtual Private Network, VPN）下載。

圖 2-9　編列員及訪查設備

Source: U.S. Census Bureau.

編列員蒐集每一個住宅單位的「結構類型」，例如獨棟房屋、多單元結構、拖車/活動房屋、船、帳篷等，以及「地圖點」表示住宅單位在路面上的大致位置。編列員蒐集地圖點的首選位置是住宅結構的主要入口處，若 GPS 信號無法接收，則可選擇站在次要入口、車道或住宅結構的通道處。

編列員透過輕觸或點擊 LiMA 的地圖來蒐集地圖點，將地點選定在結構的相對位置或 YAHI 區域（GPS 可用時），並將螢幕上被點擊的位置與編列員站立位置的 GPS 定位（無 GPS 時改用 Wi-Fi 及行動網絡位置資訊）進行比較，如果兩點間的距離超出可接受範圍，LiMA 只會提示編列員，但不會被強制中止這次座標蒐集，因為可能有某種合理的原因讓他們選擇此位置。完成後可以檢查工作是否全部完成，確保所有必要的資訊都已蒐集，並且沒有錯誤或遺漏。LiMA 蒐集到的資料可即時傳輸到 MAF/TIGER 系統，完成地址資訊的即時更新。LiMA 成

為 2020 年美國人口普查地址查驗作業中不可或缺的工具，簡化實地資料判定、蒐集過程，提高地址資訊的準確性和完整性，有助於人口普查工作的進行。

（三）資料品質控管（Quality Control, QC）

資料品質控管（QC）的目的是驗證編列員蒐集資料的準確性，在編列員完成指定的 BCU 清查後，從 LiMA 蒐集的數據將傳輸到抽樣、配對、審查和編碼系統 （Sampling, Matching, Reviewing and Coding System, SMaRCS）。由統計研究部門（Decennial Statistical Studies Division, DSSD）使用 SMaRCS 來提供 QC 指標摘要，並以 LiMA 蒐集的數據來分析作業期間編列員的錯誤數、錯誤頻率和錯誤嚴重性。SMaRCS 會依據 BCU 的特徵自動評分（表 2-6），這些特徵表示 BCU 可能有潛在錯誤或編列員可能未遵循正確的程序進行實地清查，例如超過 50% 的地址被標記為「無法進入」或「辦公室更新」，某個 BCU 中「被新增」或「刪除」的地址過多等情形。BCU 會根據違規記點程度被分到高、中、低三個抽樣層之一，其中「高抽樣層」係指違規記點超過 29，「中抽樣層」違規記點介於 20-29，「低抽樣層」違規記點低於 20；違規記點較高的 BCU 被認為錯誤較多，因此更有可能被選中進行品質管制追蹤（實地複查）。

表 2-6　資料品質控管（QC）特徵與違規記點對照表

Condition: BCU Has…	Points	Explanation
More than 75 percent of its addresses in multiunit structures with five or fewer units	10	Listers frequently make errors in multiunits with five or fewer units
More than 22.2 percent of addresses marked as does not exist, duplicate, or exists in fringe	10	Deletes, duplicates, and addresses that exist in the fringe are prone to error
More than 12 percent of addresses added	10	Added units are prone to error
An average strand length (distance between manual and GPS coordinates) minus GPS accuracy more than 20 meters.	10	The lister appears to be far from the units they are listing.
An average strand length (distance between manual and GPS coordinates) minus GPS accuracy between 12 and 20 meters.	5	The lister does not appear to be close enough to the units that they are listing, but not as far away as the previous test.
One or more curbstoning clusters, defined as 6 or more addresses within 7.6 meters (excluding multiunits)	10	The lister appears to be listing many addresses from the same physical location, a sign of falsification.

Source: U.S. Census Bureau.

未達標準而進入品質管制追蹤者，複查員會根據 BCU 內抽樣計畫決定的檢查數量及指定地址進行實地複查，將實際看到的情況與清查結果進行比較，並透過 LiMA 發現錯誤。如果錯誤數量小於或等於可容許的錯誤數量，則 BCU 通過 QC；如果超過容許的數量，則未通過，複查員將清查整個 BCU，以糾正任何可

能的錯誤，藉此來監控編列員的工作品質，並採取適當措施來訓練、觀察或解僱編列員。

2020 年普查前編列員執行約 111.5 萬個 BCU，編列員對無法清查的 BCU 會標記「無法工作（Unable to Work, UTW）」，表示該 BCU 不會進行 QC，也不會被用於更新 MAF；沒有被標記 UTW 的合格 BCU 大約有 1 百萬個，占整體 93.8%，QC 再從其中抽選 12.3 萬個進行複查，複查率約 11.8%（表 2-7）。

表 2-7　BCU 清查及分層抽樣後複查工作量

Status	Low Stratum (5% sample rate)*		Middle Stratum (10% sample rate)*		High Stratum (100% sample rate)		Total
	Count	Percent	Count	Percent	Count	Percent	Count
Completed Production	837,000	75.1%	248,000	22.2%	30,500	2.7%	1,115,000
Eligible for QC	783,000	74.9%	234,000	22.4%	29,000	2.8%	1,046,000
Selected for QC	64,000	52.0%	30,500	24.8%	29,000	23.6%	123,000
Training Period 1*	3,800	73.1%	1,200	23.1%	150	2.9%	5,200
Training Period 2*	22,000	74.6%	6,700	22.7%	800	2.7%	29,500
Normal Sampling	38,000	42.9%	22,500	25.4%	28,000	31.6%	88,500

Source: U.S. Census Bureau.

被選中進行品質管制的 BCU，會依據內含的地址數量，抽選部分地址進行品質確認，複查員會根據抽樣計畫所定的地址數量進行抽查並計算錯誤（表 2-8）。複查員所發現的錯誤如果會導致 MAF 出現涵蓋範圍問題，則列為「關鍵錯誤（critical errors）」，例如，編列員誤刪住宅單位、未添加有效的住宅單位；反之，不會影響 MAF 涵蓋範圍的錯誤則列為「次要錯誤（minor errors）」，這類錯誤通常會使找到地址變得困難，例如未更正不正確的郵政編碼、誤改建物結構類型等，惟若單個地址中有 3 個以上的次要錯誤，則計為 1 個關鍵錯誤。

表 2-8　BCU 抽樣計畫容許錯誤數

Range of BCU Sizes (Number of Addresses)	Sample Size	Number of Allowable Critical Errors
$x \leq 20$	All	0
$20 < x \leq 70$	20	0
$70 < x \leq 105$	35	1
$105 < x \leq 245$	40	1
$245 < x \leq 525$	50	2
$x > 525$	65	3

Source: U.S. Census Bureau.

每個被抽到的 BCU 經過品質管制複查後，依據是否可進行檢查以及錯誤程度，可以被歸類為 4 種結果（表 2-9）：

- 無法工作（Unable to Work, UTW）：複查員無法檢查 BCU 中的任何地址，BCU 內全部地址標記為 UTW。
- 無法完成（Could Not Finish, CNF）：複查員將 BCU 內部分地址標記為 UTW，剩餘能檢查的地址數太少，低於抽樣檢查所需的數量。
- 未通過（Fail）：編列員的錯誤數量超過允許數量。
- 通過（Pass）：編列員的錯誤數量未超過允許數量。

表 2-9　BCU 分層抽樣後複查結果

BCU Status	Low Stratum (5% sample rate)	Middle Stratum (10% sample rate)	High Stratum (100% sample rate)	Total
Completed in QC	64,000	30,500	29,000	123,000
Unable to Work	500	100	100	700
Could Not Finish	550	400	400	1,300
Passed	44,500	18,000	14,500	77,000
Failed	18,500	12,500	13,500	44,500
Percent Failed, unweighted*	29.4%	41.0%	48.2%	36.6%
Percent Failed, weighted (standard error)	32.2 (0.2)	41.8 (0.3)	48.1 (0.0**)	34.8 (0.2)

Source: U.S. Census Bureau.

一旦基本蒐集單元通過品質管制（QC）後，LiMA 就會建立地址更新檔案，透過服務導向架構，將地址更新檔案透過地址接收服務傳遞給地理部門（GEO）的 MAF/TIGER 系統，以進行資料庫更新。

（四）小結

為確保地址母體清單的準確性和完整性。IFAC 扮演關鍵重要的角色，如準確的地址清單可確保普查問卷能夠寄達正確地址，對於未回復郵寄問卷的住戶，普查局會派遣人員進行後續實地訪問，有準確的地址清單可以幫助實訪人員順利找到這些住戶。此外，準確的地址清單是進行人口統計和分析的基礎。

由於辦公室內地址清查作業大幅減少 IFAC 的工作量，相較於 2010 年 IFAC 涵蓋美國大部分地區，2020 年 IFAC 大約只涵蓋 35%的地址，共計花費約 1.2 億美元，其中以人力和交通支出減省最多，分別減少約 3,800 萬及 2,170 萬美元，與 2010 年普查 IFAC 總支出 4.53 億美元（按通貨膨脹調整後約為 5.36 億美元）相較，2020 年 IFAC 成本大幅降低。

第三章　差分隱私（Differential Privacy）作業及隱私保護技術

第一節 美國隱私保護技術發展

一、隱私保護重要性

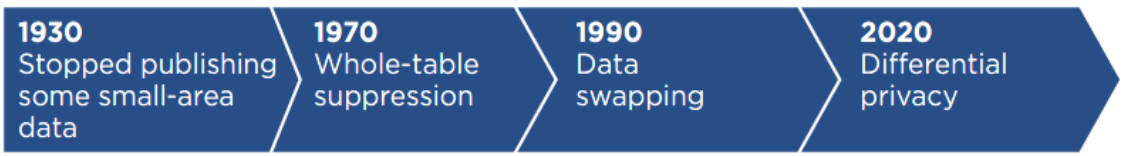
許多商業供應商（commercial vendor）蒐集、出售和發布居住於美國的民眾資料，儘管這些資料具有姓名、地址和出生日期等基本特徵，卻極少能取得人口普查所蒐集到的種族（race）、族裔（ethnicity）及家戶關係（household relationship）等關鍵人口特徵。這些詳細特徵一旦揭露，可能讓詐欺案件、虛假資訊散布，甚至網路暴力變得更加氾濫，其中對弱勢群體（如有色人種社區、同性伴侶、老年人及年幼孩童的父母）影響尤為嚴重。此外，統計結果被解密，也將損害民眾對於政府單位的信任，導致民眾更不願意誠實回答相關調查。

二、隱私保護歷程

1930 年的人口普查為避免外界間接揭露個人資訊，美國普查局停止發布部分小地理區域的統計結果；至 1954 年，隱私保護規定正式納入《美國法典》第 13 章；至 1970 年，美國普查局在特定地區遮蔽部分人口及家庭規模的統計結果表；自 1990 年起，美國普查局開始採用更為複雜的技術—資料交換（Data swapping）來防止資料被揭露，主要係將相鄰地區且具有相似特徵的家庭紀錄進行交換，以增加資料的不確定性。此技術雖防範資料被揭露，但因為普查局沒有公布交換方法的具體資訊，造成資料使用者無法評估這些保護措施對發布資料的影響程度。

在 2000 年和 2010 年的人口普查，美國普查局持續使用資料交換，並使用頂端和底端編碼（Top- and bottom-coding）、空白與插補演算法（blank-and-impute algorithms）、表格及欄位遮蔽（table and cell suppression）等技術，以進一步保護受訪者的隱私；至 2020 年的人口普查，為因應現代數據保護的高標準要求，美國普查局引入差分隱私技術，提供精確且可量化的隱私保護方法，成為隱私保護技術發展的重要里程碑（圖 3-1）。

圖 3-1　美國人口普查隱私保護歷程



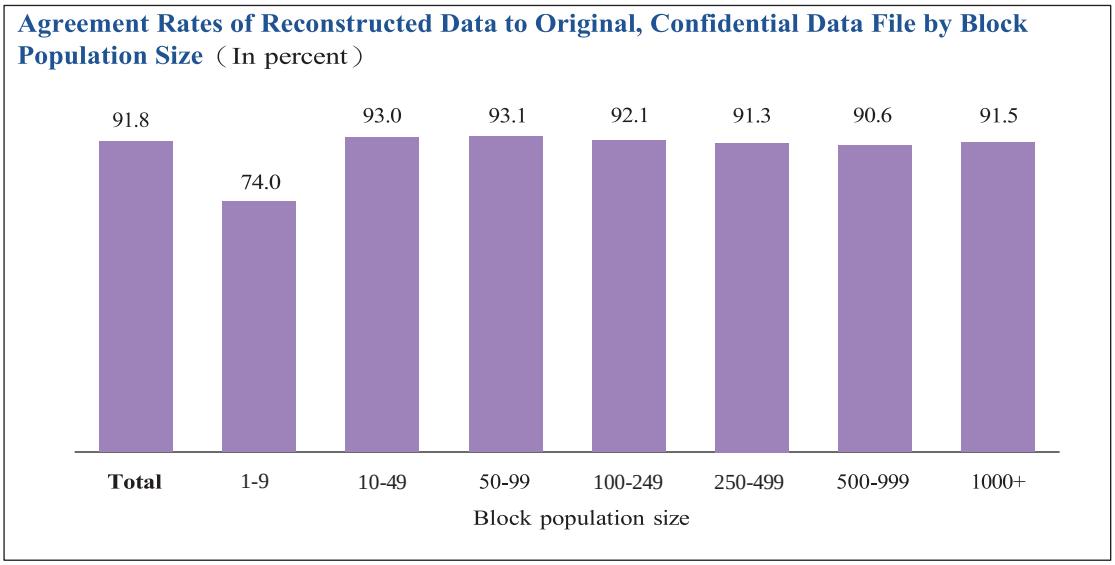
三、隱私保護新挑戰

10 年辦理 1 次的人口普查所釋出之產品數量和複雜性不斷增加，2010 年的人口普查發布約 1,500 億筆巨量統計資料，如果發布的資料是只有幾戶人家的小社區，且大多數人具有相似的特徵，一旦該社區出現某個獨特特徵，其他居民可能很容易猜測出當事人的身分。隨電腦算力提升及各類資料庫數量快速增加，即使在較大地理區域中，擁有獨特特徵的人口被重新識別風險也比以往更高，因而資料保密的技術也需配合升級。

美國普查局發現統計資料越來越容易受到重建資料庫（database reconstruction）和重新識別（reidentification）攻擊，外部機構可能蒐集已發布的統計表，並運用數學模型重建人口普查原始個別資料，再透過人口普查變數或其他個人特徵，與外部資料庫連結，推斷出受訪者的個人隱私資訊。

美國普查局曾使用 2010 年人口普查統計結果表（已經過資料交換等避免資料揭露技術處理）進行一項實驗，結果成功重建超過 3 億筆個人資料；隨後，美國普查局將這些重建的紀錄與 4 個商業資料庫進行比對，識別出居民的年齡、性別、種族、族裔和地理位置等資訊，再比對人口普查原始資料，資料一致率達 91.8%[4]（圖 3-2）。這項資料庫重建模擬的結果震撼美國普查局，因為重建的資料如果被公開，將違反人口普查所承諾的保密規定。

圖 3-2　資料庫重建結果與人口普查資料比對一致率



Source: John M. Abowd and Michael B. Hawes, "Confidentiality Protection in the 2020 US Census of Population and Housing," arXiv:2206.03524, <https://arxiv.org/abs/2206.03524>.

[4] 資料交換技術主要針對人口較少的街廓使用，因此 1 至 9 人的街廓一致率僅 74.0%較低，10 人以上較大街廓一致率則高達 9 成以上。

在 2010 年人口普查中所採用的資料交換方法，主要目的在保護最有可能被識別的個人資料，並非針對資料庫重建和重新識別等類型的攻擊而設計。如果 2020 年人口普查數據延用傳統的揭露避免技術（disclosure avoidance technique），所需的噪音量（noise）將使得這些資料不適合大多數用途。這個問題促使美國普查局的資料管理執行政策委員會（Data Stewardship Executive Policy Committee, DSEP）對 2020 年人口普查的揭露避免技術進行改造—差分隱私（differential privacy）保護機制應運而生。

第二節 差分隱私（Differential Privacy, DP）技術

一、差分隱私概述

2006 年，Dwork 等人提出「ε-差分隱私（ε-differential privacy）」的概念，並以數學方式定義統計資料集中的隱私損失。統計資料集係指在承諾保密的條件下蒐集的一組資料，這些資料僅用於產生統計結果，且不能損害資料提供者的隱私。差分隱私是一種資料隱私保護技術，主要在發布資料或分析結果的同時保護個人隱私。其核心思想是，即使某個人的資料被加入或移除，外界從資料集所能獲得的資訊變化不大，從而防止單一個人的資訊被識別或推斷出來。

差分隱私的原理是在蒐集的資料中添加「噪音」，也就是隨機加上或減去一些小數值，避免外界使用已發布數據的任意組合推斷出特定個人或家庭特徵。就像電視螢幕上看似清晰、明快的圖片，實際上是由數百萬個像素、微小的顏色點組成，如果放大影像，就能夠識別單一像素；添加噪音類似對像素進行微小更改，噪音降低正確識別任何內容的風險，但當影像縮小時則仍然能夠完整呈現整體圖片。

差分隱私（Differential Privacy）相較於資料遮蔽（Data Suppression）、資料交換（Data Swapping）方法，具有保密性（Confidentiality）、準確性（Accuracy）及可用性（Availability）三個主要優勢。在「保密性」方面，差分隱私提供數學上可證明的保護措施，而傳統方法（如資料遮蔽）在處理相關交叉表格時效果不佳，資料交換即使提高交換率也難以避免重新識別；在「準確性」方面，差分隱私透過與內、外部使用者合作建立準確性目標，並進行廣泛參數分析，來確保資料品質，比起資料遮蔽造成的分析偏差，及資料交換可能導致的人口統計扭曲都更有保障；在「可用性」方面，雖然差分隱私為因應保護強度而增加噪音，可能影響部分表格的發布，但不會像資料遮蔽那樣引起使用者對大量數據缺失的不滿，也比資料交換提供更好的保密性。考量這些優勢，美國普查局在實務上選擇採用差分隱私方法（表 3-1）。

表 3-1　不同揭露避免方法的比較

揭露避免方法	保密性 **Confidentiality**	準確性 **Accuracy**	可用性 **Availability**
資料遮蔽 Data Suppression	以特定規則遮蔽欄位數值，因表格間具有複雜的交叉關係，保密效果不佳，可能無法有效保護隱私。	缺失的資料會在分析時造成偏差。	1980年人口普查的資料遮蔽量過大，引起使用者不滿，後來改用資料交換方法。
資料交換 Data Swapping	2010年人口普查的交換率相對較低，無法充分保護受訪者隱私，即使提高交換率也難以防範重新識別攻擊。	可能導致人口統計、種族分析及年齡結構嚴重扭曲。	交換不會限制資料的可用性，只會影響資料的準確性和機密性。
差分隱私 Differential Privacy	提供數學上可證明的保護措施。	美國普查局與內、外部資料使用者合作，共同為數據確立準確性目標，並透過廣泛的分析來設定參數，以達成相關目標。	理論上無須限制發布的資料量，但實務上發布更多資料需要添加更多噪音來保護隱私，就可能影響表格的可用性，因此美國普查局決定不發布某些表格。

資料來源：美國普查局。

二、差分隱私定義

（一）傳統差分隱私（$(\varepsilon, \delta)-$ differential privacy）

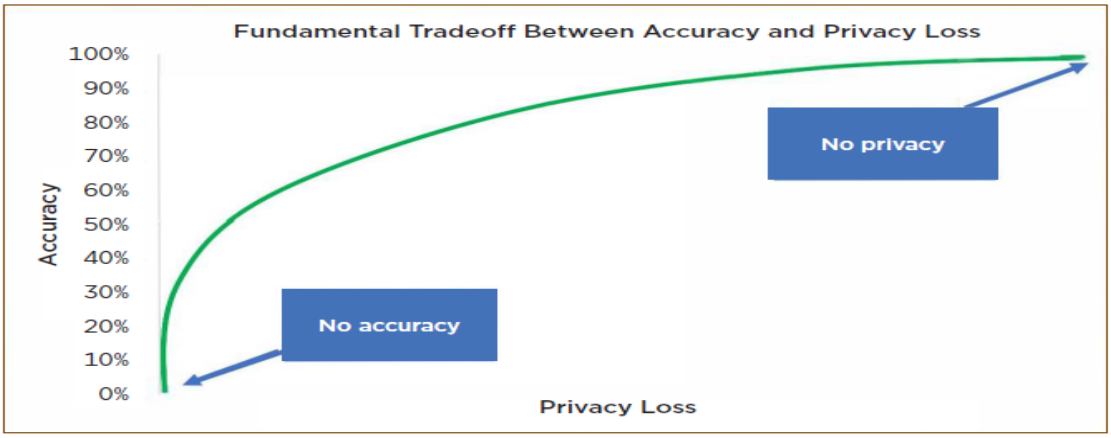
給定$\varepsilon \geq 0$、$\delta \in [0,1]$，如果對於任何兩個鄰近資料集 x、x'（其區別僅差在一筆資料），以及任何輸出子集 $E \subset range$（M），一個隨機機制（Randomized Mechanism）M 都滿足以下不等式，則 M 被稱為 $(\varepsilon, \delta)-$ differential privacy：

$$P(M(x) \in E) \leq e^{\varepsilon} P(M(x') \in E) + \delta$$

其中，ε 係用來衡量一筆資料對查詢結果的影響程度，ε 越小表示 $M(x)$、$M(x')$查詢結果越相近，具有更強保護力，ε 稱為隱私損失預算（Privacy Loss Budget, PLB）；

δ 表允許資訊揭露的機率，用以提供隱私模型設計更多彈性。

圖 3-3　隱私損失與精確度關係



Source: U.S. Census Bureau.

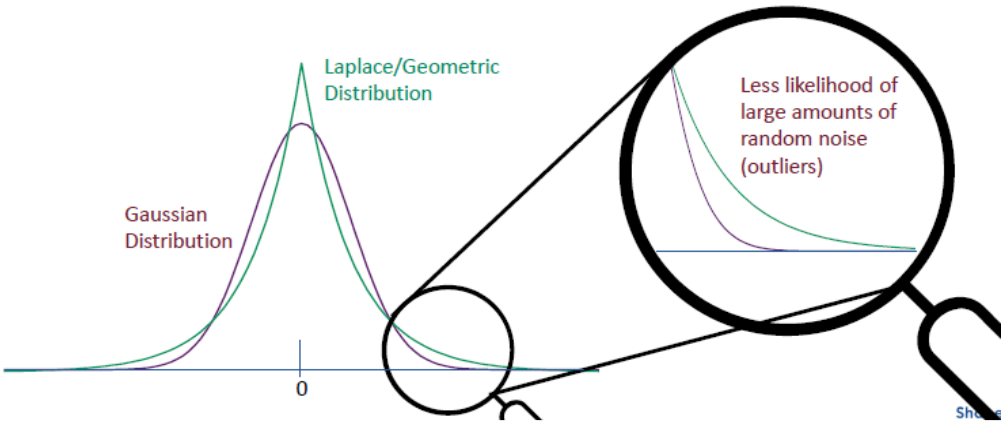
為了滿足差分隱私定義，常見機制為拉普拉斯機制（Laplace Mechanism）或高斯機制（Gaussian Mechanism），其做法係在數據上添加拉普拉斯或高斯分布隨機噪音，這些噪音使得結果不精確，但可控，同時確保隱私並保有統計分析價值。

以高斯機制$M(x)$為例，在查詢結果$f(x)$中添加高斯分布噪音，數學式如下：

$$M(x) = f(x) + noise, \qquad noise \sim N(0, \sigma^2)$$

在相同的隱私損失預算下，使用高斯分布抽取到較大噪音的機率會比拉普拉斯分布低（圖 3-4），添加噪音後的資料相對較為精確。

圖 3-4　拉普拉斯與高斯分布

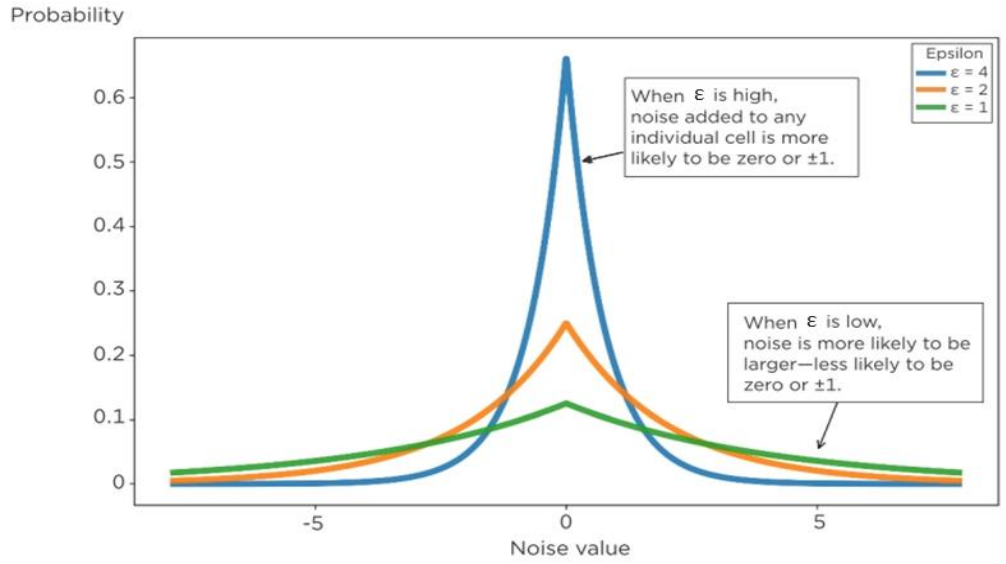


Source: U.S. Census Bureau.

差分隱私使用隱私損失預算（ε）決定加入的噪音程度（圖 3-5），以限定資料集最大揭露風險。隱私損失預算的設定，就像一個控制旋鈕，決定從分布中隨機抽取噪音的範圍。隨著隱私損失預算提高，分布的形狀變得更加陡峭，噪音值

為零的機率提高，統計數值越準確，對隱私之保護力下降。在最極端的情況下，隱私損失預算無限大表示完全無噪音，數值完全準確；反之，隱私損失預算為零表示完全噪音，資料沒有準確性。

圖 3-5　隱私損失預算（ε）－調節噪音大小的控制旋鈕



Source: U.S. Census Bureau.

（二）零集中差分隱私（Zero − Concentrated Differential Privacy）

零集中差分隱私（Zero − Concentrated Differential Privacy, zCDP）是傳統差分隱私的一種強化版本，隱私損失不再僅取決於 ε 與 δ，而是透過一個隱私參數 ρ 來確保隱私性，並引入 Rényi 散度量化兩個分布之間的「差異」，而非機率。特別是需要處理多次查詢時，能使隱私預算控制更加靈活，對於平衡隱私保護與數據實用性有更好表現，為美國普查局目前所採用的方法。

給定$\rho \geq 0$，如果對於任意兩個鄰近資料集 x、x'（其區別僅差在一筆資料），及所有 $\alpha \in (1, \infty)$，一個機制M滿足以下不等式，則 M 被稱為 $\rho -$ zero $-$ Concentrated Differential Privacy：

$$D_\alpha（M（x）||M（x'））\leq \alpha\rho$$

其中，$D_\alpha（P||Q）= \frac{1}{\alpha-1} log （\sum P（E）^{\alpha} Q（E）^{1-\alpha}）$ 是Rényi散度（Rényi Divergence），表示機制M在兩個鄰近數據集上的輸出分布之間的差異程度，α為Rényi散度的階數；

ρ 類似於傳統差分隱私中的 ε，但更加靈活，能夠提供更好的隱私效用。

美國普查局最常被查詢的資料是人口統計，查詢結果必須為整數，爰採用離散型高斯分布作為添加噪音的機制。

三、美國揭露避免系統

差分隱私能模糊資料集中任何人或一小群人的存在與否，同時保留統計價值。不論攻擊者部署何種類型的演算法或外部資料庫，揭露的風險程度不會受到影響，這對於訂定揭露審查標準的透明度具有優勢。美國普查局以零集中差分隱私為基礎，建構避免揭露系統（Disclosure Avoidance System, DAS），用以保護 2020 年人口普查受訪者的回復資料，且為因應不同人口普查資料產品特性，發展了 TopDown Algorithm（TDA）、SafeTab、PHSafe 三種演算法，在人口普查統計表格導入差分隱私技術，有別於一般應用係在個別資料加入噪音。2020 年美國人口普查釋出的個別資料，為經過 TDA 演算法的統計表格生成，並作為人口普查查詢系統的底層資料。

第三節 TopDown Algorithm（TDA）演算法

一、P.L. 94-171 重新劃分選區數據摘要檔案

2020 年美國人口普查中第一個使用差分隱私保護的資料是「PL 94-171 重新劃分選區數據摘要檔案」。此檔案根據《1975 年公共法律第 94-171 號》制定，要求美國普查局在辦理人口普查後一年內，必須交付各州人口變動資料及小地區的人口統計資訊，並依此重新劃分選區，這些數據對選區劃分扮演極為重要的角色。

在 1980 年的人口普查重新劃分選區數據中，美國普查局首次納入管理和預算辦公室（U.S. Office of Management and Budget, OMB）第 15 號指令指定之主要種族群體資料。這些群體包括白人、黑人、美洲印第安人/阿拉斯加原住民、亞洲人/太平洋島民及其他種族，並與族裔（西班牙裔/非西班牙裔）產生交叉分類。另應州立法機構和司法部的要求，1990 年重新劃分選區數據中增添了投票年齡（18 歲以上）統計，以便進行選舉相關分析。到 2020 年，重新劃分選區數據檔案首次包含居住在 7 種類型集體住所中的人口數據，例如懲教設施、學院/大學宿舍以及軍事宿舍。

摘要文件內容包括：

P1.種族（Race）

P2.西班牙裔或拉丁裔、非西班牙裔或拉丁裔按種族分類（Hispanic or Latino, and not Hispanic or Latino by Race）

P3.18歲以上人口的種族分類（Race for the Population 18 Years and Over）

P4.18歲以上人口的西班牙裔或拉丁裔、非西班牙裔或拉丁裔按種族分類（Hispanic or Latino, and not Hispanic or Latino by Race for the Population 18 Years and Over）

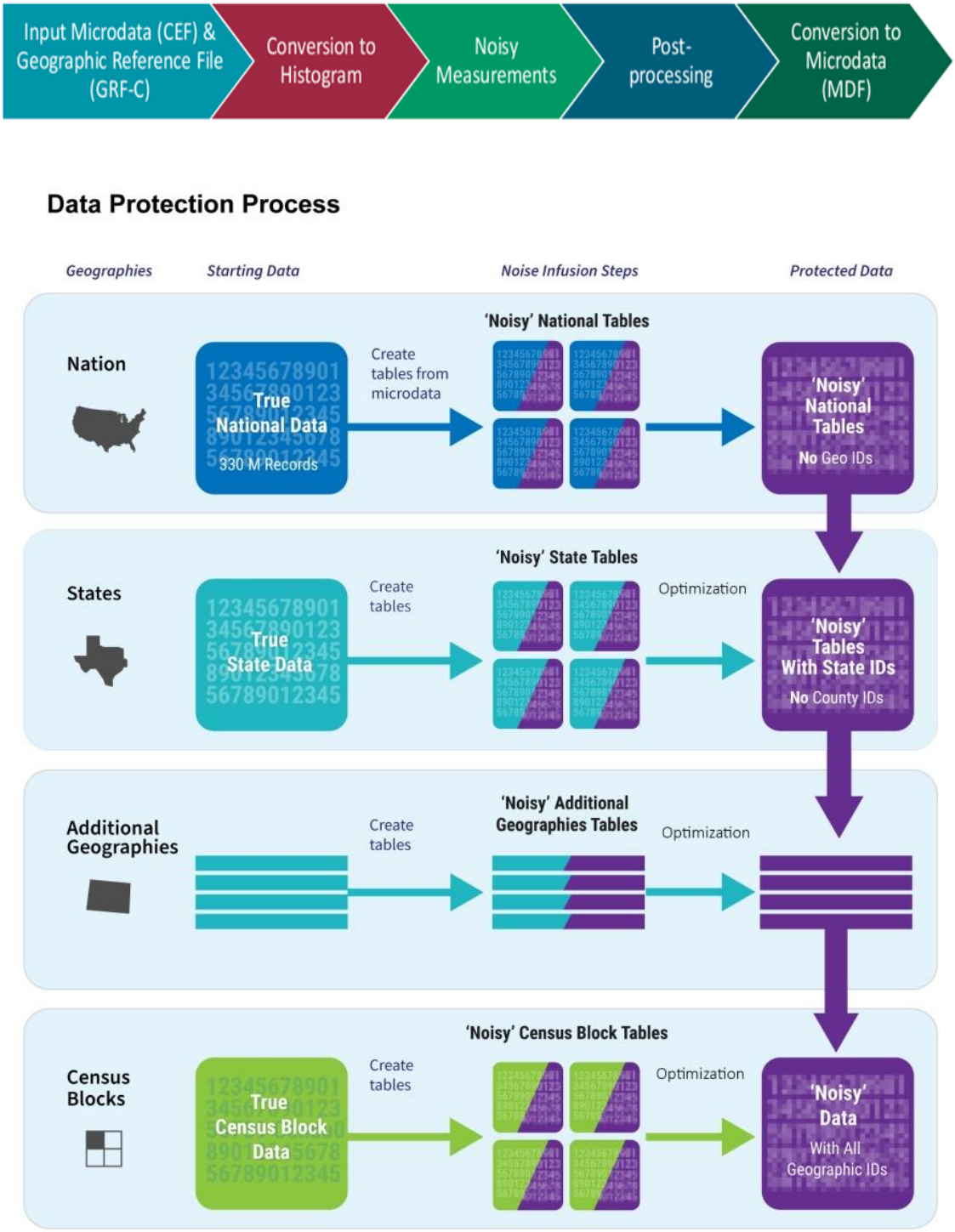
P5.集體住所人口按主要群體類型分類（Group Quarters Population by Major Group Quarters Type）

H1. 居住狀態（Occupancy Status, Housing）

二、TopDown Algorithm（TDA）運作原理

美國普查局避免揭露系統（DAS）用於重新劃分選區數據的重要技術包含使用差分隱私添加噪音與後處理（Post-processing）兩個部分，同時具備這兩種技術的框架稱為 TopDown Algorithm（TDA）。TDA 先為資料添加噪音，再進行後處理，以提升某些統計結果的一致性（例如：確保州內各郡的人口總和等於該州總人口數）。處理過程說明如下（圖 3-6）：

圖 3-6　TopDown Algorithm



Source: Population Reference Bureau.

（一）匯入微數據 CEF 和 GRF-C（Input Microdata CEF & GRF-C）

首先將機密普查編輯檔案（Confidential Census Edited File, CEF）與地理參照檔（Geographic Reference File, GRF-C）匯入避免揭露系統（DAS），其中「機密普查編輯檔案」是經過品質控制處理（例如插補遺失值）的個別普查資料。

（二）轉換成分組統計資料（Conversion to Histogram）

美國普查局根據利害關係人的意見，編製 2020 年人口普查「重新劃分選區數據摘要檔案」統計表（P1-P5, H1），將這些統計表合併為一個更詳細的交叉表，包括每個地理級別（從國家、州，到普查街廓）與所有人口特徵變數類別的交叉統計資料。

2020 年人口普查約有 800 萬個普查街廓（Census Blocks），其中有人居住者約有 600 萬個，每個街廓依種族、族裔、年齡及住宅類型等變數，可能有 2,016 種分類（63 x 2 x 2 x 8 = 2,016，表 3-2），TDA 需要處理超過 160 億筆人口及 1200 萬筆住宅統計資料。

表 3-2　2020 年人口普查重新劃分選區檔案變數分類

變數	分類數量
種族 Race（6 race alone groups; 57 multiple race combinations）	63
族裔 Ethnicity（Hispanic or Latino; Not Hispanic or Latino）	2
年齡 Age（voting age, total population）	2
占用情形 Occupancy status（occupied, vacant）	2
住宅類型 Population in housing units（1 type）or in group quarters（7 types）	8

Note: This table shows the number of categories for each variable, not the publication data layouts.
Source: Population Reference Bureau.

（三）噪音測量（Noisy measurements）

提供高度準確的統計資料會帶來揭露風險，10 年一次的人口普查資料必須在機密性與準確性之間取得平衡，TDA 選擇零集中差分隱私框架，為每個統計結果添加（正或負）整數噪音，以保護個別受訪者的機密，這些統計資料添加隨機噪音後產生的檔案稱為「噪音測量檔案（Noisy Measurements File, NMF）」。

添加到每個單元格的噪音量與其數值大小無關，這意味著小數值的單元格有可能添加較大的噪音量，反之亦然；部分單元格則可能不添加任何噪音，數值因此保持不變。此外，噪音會獨立地添加到每個單元格中的各個特徵上，然而這種獨立性可能導致表格內的數據出現邏輯上的錯誤。以下例來看（表 3-3），街廓 4 加入過多負值噪音，致未滿 18 歲人口數為-1；又如街廓 5 在添加噪音後 18 歲以上人口數為 4，竟高於該街廓人口總數 3 等不合理現象。

表 3-3　人口普查街廓群組的噪音添加範例

街廓	編列人口數			噪音			初步噪音表格		
	未滿18歲	18歲以上	總數	未滿18歲	18歲以上	總數	未滿18歲	18歲以上	總數
街廓1	25	75	100	0	–4	2	25	71	102
街廓2	20	70	90	–3	2	3	17	72	93
街廓3	10	40	50	2	–3	–2	12	37	48
街廓4	1	9	10	–2	1	1	–1	10	11
街廓5	1	2	3	0	2	0	1	4	3

資料來源：美國普查局。

（四）後處理（Post-Processing）

對於加入噪音後產生的不合理情形，尤其是負數結果，會使數據使用者感到困惑，因此需要「後處理」來調整噪音結果，其步驟如下：

1. 不變量（Invariants）

人口普查中的各州人口統計數據被用來重新分配美國眾議院在 50 個州中的席位，某些數值必須確保在噪音添加後不變動，不變量包括「各州、哥倫比亞特區及波多黎各的總人口數」、「各普查街廓的住宅單位總數」與「各普查街廓各類型使用中的集體住所設施數量」。

差分隱私不同於傳統的揭露避免方法，其噪音添加技術提供「可量化」且「可證明」的機密性保證。其中「不變量」的設計相當於將無限大的 PLB 分配給特定統計資料，偏離一般差分隱私做法，因為數據不受 PLB 控制，即無法向受訪者保證「所發布的統計結果得以限制攻擊者推斷某些資訊」，「不變量」弱化控制揭露風險的核心承諾。基於此，普查局限制 2020 年人口普查中的不變量數量。

2. 額外限制條件（Additional Constraints）

除了上述不變量外，TDA 還對所有地理層級應用某些限制條件。這些限制條件包括：

- 人口和住宅數量必須為非負整數。
- 表格中的各單元格之行、列加總後應分別等於其行和列的邊際值（margins），而這些邊際值合計必須等於該表的總人口數。
- 針對特定範圍統計數據必須在表內、表間以及不同地理區域間保持一致。例如不同種族之人口數合計後應等於總人口數，「有人居住」和「無人居住」住宅單位數量之合計必須等於住宅單位總數，州內各郡的人口數總和應等於該州的總人口數。
- 如果某個地理區域內沒有住宅單位且沒有集體住所，則不能將任何「人」分配到該地區。
- 每個集體住所的居住人數至少 1 人。
- 每個住宅單位和集體住所設施裡的人數應少於 10 萬人。
- 在「護理機構/專業護理機構」集體住所類型中，未滿 18 歲的人數應為零。

以表 3-4 為例，每個單元格中加入的噪音可能會導致一些不合理的情形，後處理步驟用來解決加入噪音後的幾個問題，如街廓 4 中未滿 18 歲人口數為-1，必須調整為非負整數；街廓 5 中 18 歲以上人口數為 4，高於該街廓人口總數 3，亦必須修正；最後，檢查所有相關地理區域人口總數的一致性，初步噪音表格人口總數為 257，必須與經過隱私保護的街廓群組總數一致，調整為 254 人。

表 3-4　後處理說明

街廓	編列人口數			噪音			初步噪音表格			後處理人口數		
	未滿18歲	18歲以上	總數	未滿18歲	18歲以上	總數	未滿18歲	18歲以上	總數	未滿18歲	18歲以上	總數
街廓1	25	75	100	0	–4	2	25	71	102	27（+2）	71（–4）	98（–2）
街廓2	20	70	90	–3	2	3	17	72	93	19（–1）	72（+2）	91（+1）
街廓3	10	40	50	2	–3	–2	12	37	48	12（+2）	37（–3）	49（–1）
街廓4	1	9	10	–2	1	1	–1	10	11	0（–1）	11（+2）	11（+1）
街廓5	1	2	3	0	2	0	1	4	3	1（+0）	4（+2）	5（+2）
街廓群組										59	195	254

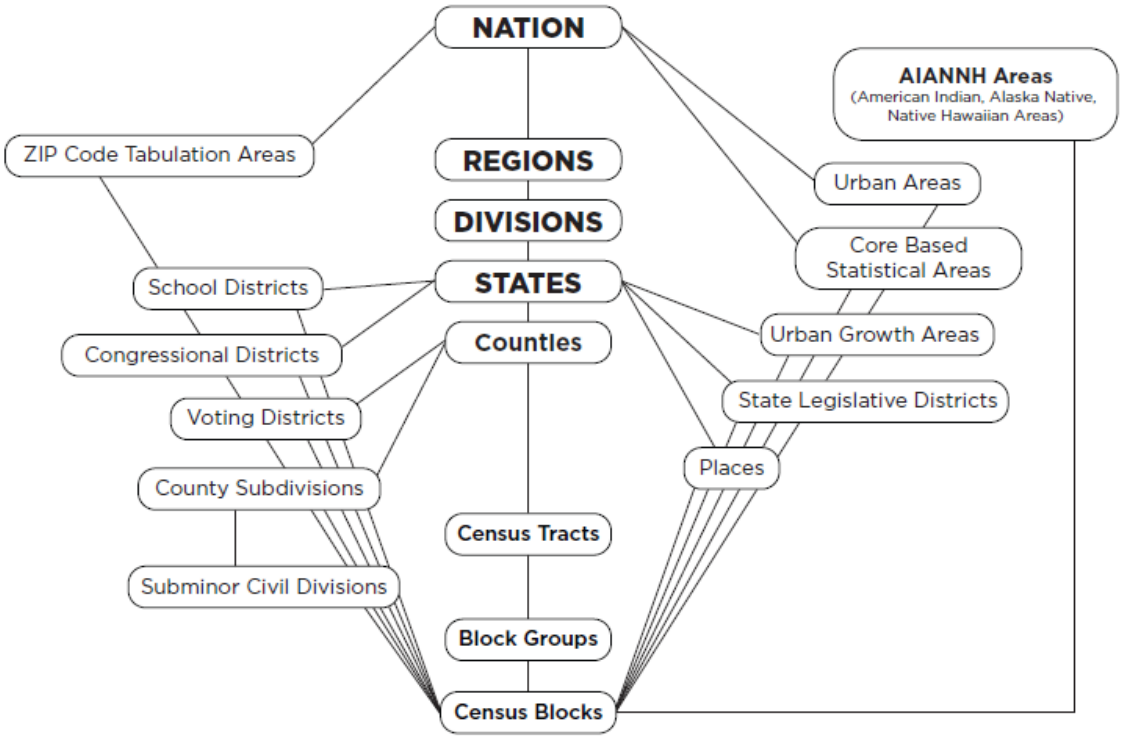
註：（）內數字係「後處理人口數」與「編列人口數」之差值。
資料來源：美國普查局。

因為噪音是從以零為中心的對稱分布中提取的，不會引入偏差（Bias），但具有大量「負數」的統計資料經過「後處理」的結果可能發生扭曲。美國普查局發現小人口區域傾向出現「正向偏差」，即發布的人口數高於原始數據；而較大的人口區域則傾向出現「負向偏差」。例如農村地區經由後處理，人口會從城市轉移到農村，導致農村人口數高於原始數據。

為消除這種偏差，美國普查局重新配置 TDA 參數，實施多層級處理框架，依序在國家層級、州層級，然後是更低的地理層級，確定每個地理層級中各單位的人口數，並且這些統計數據符合上層原定的人口數範圍。這項新的優化程序，可以減少小地理區域和人口子群體中的偏差。隨著被測量的基礎人口規模增大，統計數據的準確性和可靠性隨之提升。

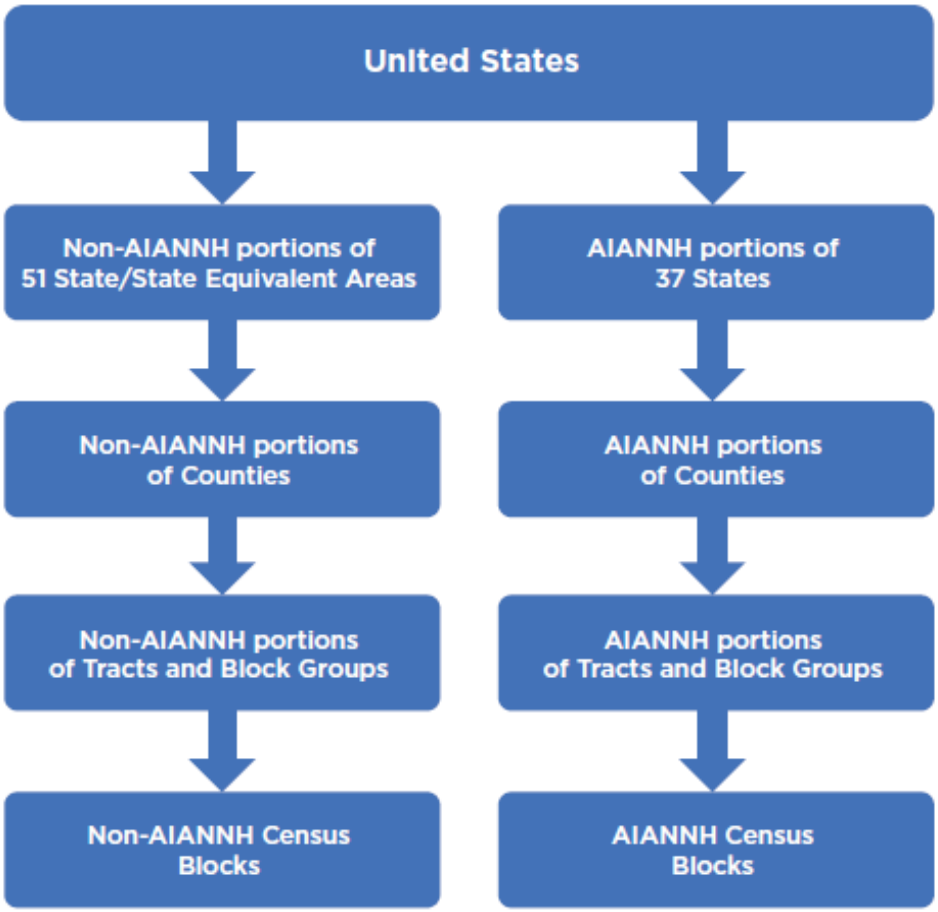
此外，TDA 僅對主幹上的地理單位（圖 3-7 中央區域）進行添加噪音與後處理。然而，某些地理區域並不適合嵌套在標準的地理框架內，例如學區不一定可由人口普查區劃分。為了回應數據使用者對於非主幹地理區域的數據需求，美國普查局對 TDA 所使用的地理層級進行了調整（圖 3-8）。這些改變旨在提升「非主幹」地區的數據準確性，確保能夠滿足不同層級的數據需求。

圖 3-7　標準地理區域層級



Source: U.S. Census Bureau.

圖 3-8　避免揭露系統地理區域層級



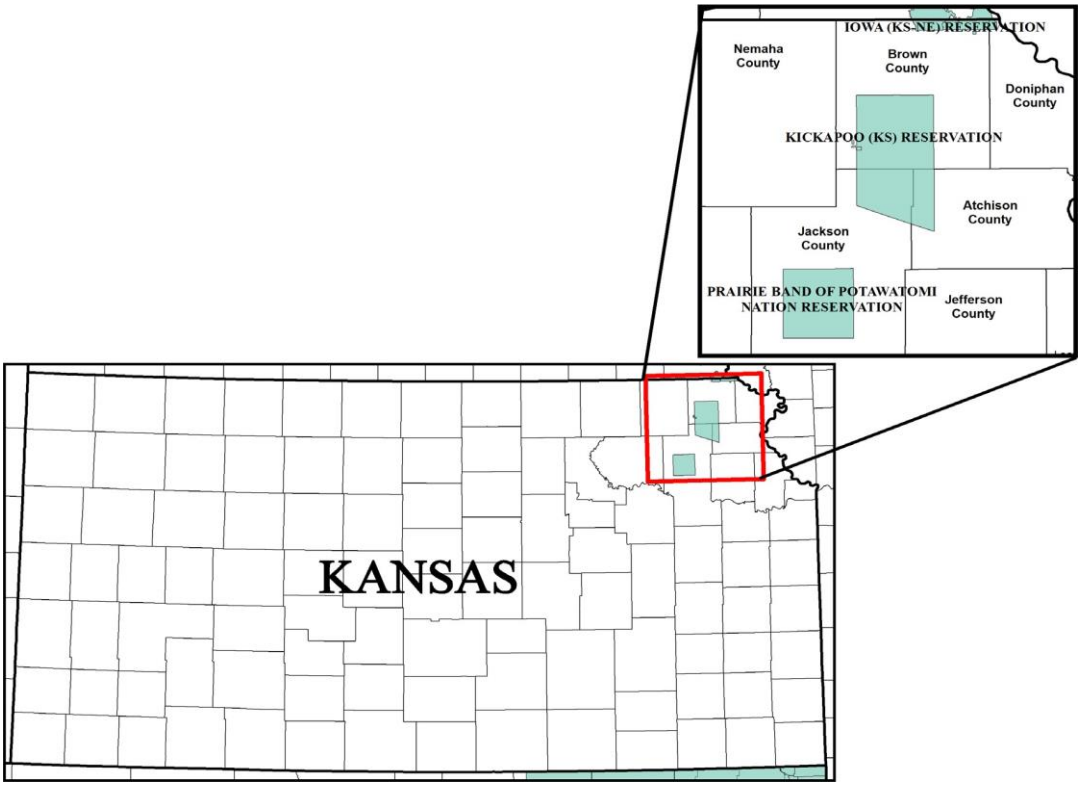
Note: AIANNH is American Indian/Alaska Native/Native Hawaiian areas.

Source: U.S. Census Bureau.

避免揭露系統（DAS）將州內所有美國原住民（American Indian）、阿拉斯加原住民（Alaska Native）、夏威夷原住民（Native Hawaiian）地區（AIANNH Areas）整合後進行處理，以提供州內該區域的總人口數，可以減少由於隱私保護導致的人口數據失準風險。

以堪薩斯州為例（圖 3-9），將該州的三個美國原住民區域合併到「州」級別處理，這樣的群組方式不僅保證這些區域人口統計的準確性，還可以避免特定區域出現系統性人口低估的情況。

圖 3-9　TDA 堪薩斯原住民地區範例



Source: U.S. Census Bureau.

（五）轉換成微數據（Conversion to Microdata）

TDA 最後一個步驟是生成整個國家隱私保護的微數據紀錄，這些個別的數據紀錄涵蓋每個地理層級，基於每一層級的噪音測量結果，同時也遵循人口不變量及額外限制條件。這些微數據紀錄會傳送至製表系統，用於產生重新劃分選區數據。

三、TDA 結果補充說明

TDA 雖設下多種限制條件，惟「重新劃分選區數據檔案」中仍然可能出現不合理的情況。例如某個街廓僅一個有人居住的住宅單位，但卻顯示有十幾個人居住，或是某個街廓顯示有未滿 18 歲的孩童居住，但卻沒有成年人；又或者某個街廓顯示有人居住，但該區域的所有住宅單位使用狀況卻都顯示為空閒；或某個街區的已占用住宅單位數超過實際居住人數等。這些矛盾的統計結果，主因隱私保護過程中引入的噪音和數據調整所致，且特別容易出現在人口數極少的地理單位中。例如表 3-5 的數據顯示，有「家戶人口」，但卻無「有人居住的住宅單位」

的矛盾情形，在街廓（Blocks）層級占 4.8%，在地理範圍較大的街廓群組（Block groups）或人口普查區（Tracts）中的比例僅約 0.1%。

隨著數據被彙總至更大的地理區域，不合常理或不可能的結果會逐漸減少，而數據的準確性也會提高。對於許多使用情境（例如進行詳細的住宅或家庭人口分析），街廓數據可能過於雜亂，使用街廓群組、人口普查區或其他更大地理範圍的數據，將減少隱私保護措施引入噪音所造成的負面影響，可能會是較好的選擇。

表 3-5　使用 TDA 統計結果不合理的情形

	街廓		街廓群組		普查區		郡	
	數量	比率	數量	比率	數量	比率	數量	比率
有「家戶人口」，但無「有人居住的住宅單位」	392,921	4.80	223	0.09	90	0.11	0	0.00
無「家戶人口」，但有1個以上「有人居住的住宅單位」	91,415	1.10	30	0.01	17	0.02	0	0.00
該區域所有人皆小於18歲（不包含集體住所）	101,127	1.80	27	0.02	17	0.05	0	0.00

資料來源：美國普查局。

第四節 SafeTab 演算法

一、詳細人口與住宅特徵檔案（Detailed Demographic and Housing Characteristics File）A

主要目標是提供全國種族、族裔、美國原住民及阿拉斯加原住民部落等詳細數據，簡稱 Detailed DHC-A，其群體包括：

- 30 個西班牙裔或拉丁裔群體，例如墨西哥人、薩爾瓦多人等。
- 270 個種族群體，例如日本人、夏威夷原住民、愛爾蘭人、黎巴嫩人、海地人、巴西人等。
- 1,187 個美國原住民及阿拉斯加原住民部落，例如納瓦霍保留地（Navajo Nation）、阿基亞克原住民社區（Akiak Native Community）等。

由於上述群體的人數相對較少，在保持機密性的同時發布統計數據成為一項挑戰。為了回應資料使用者的需求，Detailed DHC-A 的避免揭露框架專注於「在各種地理級別上，以符合機密保護為前提，儘可能提供準確的人口統計數據」。

二、SafeTab 運作原理

Detailed DHC-A 採用了自動調整設計（Adaptive Design），根據預先設定門檻值和地理層級的組合來決定發布的年齡分類方式（表 3-6）。這種設計使美國普查局能夠提供人口較多的種族、族裔更詳細統計數據，同時達到隱私保護。

表 3-6　各類表格之人口數門檻

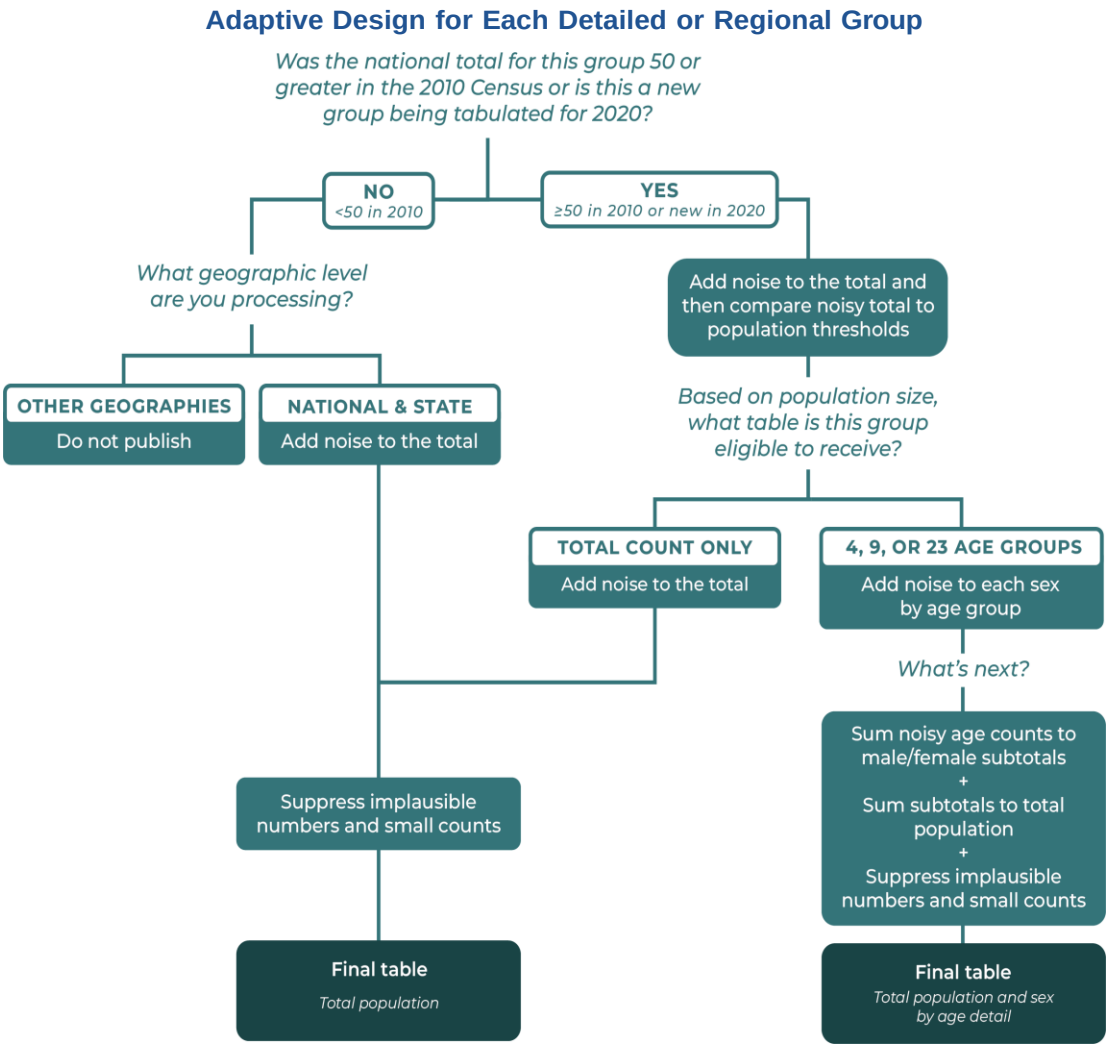
表格類型	詳細群體		區域群體	
	Nation and state	Substate and AIANNH	Nation and state	Substate
總數	0–499人	22–999人	0–4,999人	94–4,999人
性別按年齡分-4類	500–999人	1,000–4,999人	5,000–19,999人	5,000–19,999人
性別按年齡分-9類	1,000–6,999人	5,000–19,999人	20,000–149,999人	20,000–149,999人
性別按年齡分-23類	7,000+人	20,000+人	150,000+人	150,000+人

Note: AIANNH is American Indian/Alaska Native/Native Hawaiian areas. Substate includes county, place, and census tract.

資料來源：美國普查局。

SafeTab 演算法是一系列數學程序，以實現自動調整設計的構想（圖 3-10）。首先，SafeTab 會檢查某個群體是否屬於「2010 年人口普查中全國人口少於 50 的某一特定群體」。如果是，則僅會發布全國和州層級添加噪音的總人口數，不會產生性別或年齡別數據；如果屬於「在 2010 年人口普查中全國人口為 50 以上的群體，或 2020 年普查蒐集到的新群體」，先計算該群體加入噪音後的總人口數，再依表 3-6 所定之門檻決定統計表的詳細程度，此過程將針對每個地理區域、種族或族裔群體的組合反覆進行。

圖 3-10　自動調整設計



Source: Population Reference Bureau.

由於任何發布的統計數據（包括總人口數）都代表一部分隱私揭露風險，因此決定發布多少細節是基於加入噪音後的總數，而不是原始的總數。SafeTab 會向原始的總數加入噪音，並使用一個從未發布的新噪音來決定應生成何種年齡分

類。以美國的新加坡人總人口為例（表 3-7），假設人口總數為 5,350 人，添加噪音後總數是 5,355 人，數值介於 1,000 到 6,999 之間，依據表 3-6 在國家層級新加坡人的單一族群或任何交叉表將分為 9 個年齡類別；接著對性別及年齡類別的數值進行統計後添加噪音，即可得到發布資料。

表 3-7　美國的新加坡人加入噪音範例

性別和年齡	第1步驟：決定發布年齡詳細層級			第2步驟：加入噪音		
	假設人口數	加入噪音檢查總數	噪音後總數細節決策	假設人口數	加入噪音	發布資料
總數（步驟 1）	**5,350**	**5**	**5,355**	**X**	**X**	**X**
總數（步驟 2）	**X**	**X**	**X**	**5,350**	**22**	**5,372**
男性	**X**	**X**	**X**	**2,178**	**9**	**2,187**
未滿5歲	X	X	X	146	0	146
5～17歲	X	X	X	318	1	319
18～24歲	X	X	X	316	6	322
25～34歲	X	X	X	472	2	474
35～44歲	X	X	X	410	–3	407
45～54歲	X	X	X	326	0	326
55～64歲	X	X	X	141	1	142
65～74歲	X	X	X	31	1	32
75歲以上	X	X	X	18	1	19
女性	**X**	**X**	**X**	**3,172**	**13**	**3,185**
未滿5歲	X	X	X	147	–3	144
5～17歲	X	X	X	315	2	317
18～24歲	X	X	X	413	5	418
25～34歲	X	X	X	567	4	571
35～44歲	X	X	X	730	2	732
45～54歲	X	X	X	647	–3	644
55～64歲	X	X	X	264	2	266
65～74歲	X	X	X	67	3	70
75歲以上	X	X	X	22	1	23

X Not applicable.

Source: U.S. Census Bureau, Detailed Demographic and Housing Characteristics File A （Detailed DHC-A） Proof of Concept.

對於有發布「性別按年齡分」資料的群體，SafeTab 會將每個年齡組別添加噪音之資料相加，得到男性及女性總人口數，兩性資料再相加產生公布的總人口數。由於 SafeTab 會在不同地理區域和人口群體中獨立重複這個噪音加入過程，數據可能會在不同表格間出現不一致的情形。因此，美國普查局鼓勵數據使用者使用已發布的詳細群體和區域群體數據，而不要自行加總 Detailed DHC-A 數據，以免導致較不精確的結果。

三、誤差範圍衡量

在分析避免揭露系統對這些數據的影響時，所使用的關鍵指標是「誤差範圍（Margins of Error, MOE）」。對於 Detailed DHC-A 來說，誤差範圍並不是代表抽樣誤差，而是表示差分隱私所產生的噪音預期範圍。具體而言，誤差範圍（MOE）說明了添加噪音的數據與實際數據的接近程度，約 95% 的信心水準，預期「添加噪音的數據」與「實際數據」之間的差異會落在誤差範圍（MOE）以內。

換句話說，如果我們模擬重複執行 100 次添加噪音程序，實際數據大約有 95 次會落在添加噪音的數據和 MOE 之間。例如：添加噪音的數值為 20，MOE 為 3，我們有 95%的信心實際值會介於 17 和 23 之間。MOE 越小，發布的數據就越準確。不過，給定相同的 MOE，對於大規模群體的影響相對較小，而對於小群體則影響較大。

第五節 PHSafe 演算法

一、補充人口與住宅特徵（Supplemental Demographic and Housing Characteristics, S-DHC）檔案

S-DHC 表格按不同特徵（如年齡、家庭類型及住宅所有權屬）分類提供居住於家戶中的人數和平均家戶規模數據。

S-DHC 表格包括：

PH1.平均家戶規模－按年齡分（Average Household Size by Age）

PH2.家戶人口之家戶型態（Household Type for the Population in Households）

PH3.未滿18歲人口－按關係和家戶型態（Population Under 18 Years by Relationship and Household Type）

PH4.家庭人口數－按年齡分（Population in families by Age）

PH5.平均家庭規模－按年齡分（Average Family Size by Age）

PH6.有未滿18歲子女的家庭類型和年齡（Family Type and Age for Own Children Under 18 Years）

PH7.已占用住宅單位中的總人口數－按住宅所有權屬分（Total Population in Occupied Housing Units by Tenure）

PH8.已占用住宅單位的平均家戶規模－按住宅所有權屬分（Average Household Size of Occupied Housing Units by Tenure）

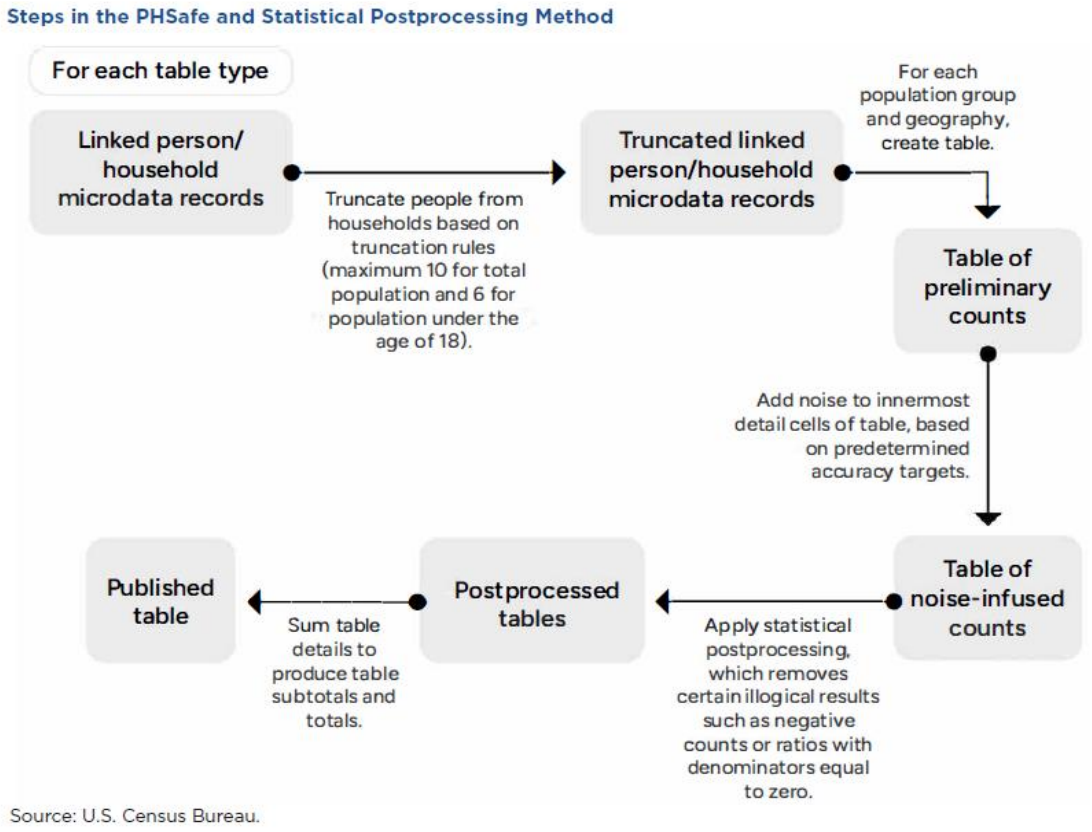
S-DHC 將家庭和其中成員的資訊結合在一起，這需要合併「個人檔案」和「住宅單位檔案」的資訊。「個人檔案」包含個人的特徵，例如年齡、種族、族裔、與戶長的關係；「住宅單位檔案」則包含家庭的特徵，如家戶型態（家庭/非家庭）、住宅所有權屬（擁有者/租戶）。當個人檔案和住宅單位檔案合併後，便能提供家戶內所有成員的資訊，包括家戶成員之間的關係（例如，可以顯示居住在已婚夫妻家戶中的人數）。將個人數據和住宅數據結合在一起的揭露風險比單獨發布其中一項資料要高，這是因為住宅內的相互關係將難以隱藏個人數據對其他成員的影響，進而增加保護這些數據的難度。

S-DHC 表格適用於全國、50 個州、哥倫比亞特區以及波多黎各，受限於保密性、數據品質準確性，僅能提供特定的地理區域，並不適用「州」以下的地理區域。

二、PHSafe 運作原理

PHSafe 是一個用於保護 S-DHC 的演算法，首先使用 2020 年機密普查編輯檔案（CEF）中的個人和家戶個別數據來運作，並依截斷規則限制家戶人口數，再產生統計表，對計算結果「加入噪音」，最後進行「後處理」，來確保資料符合邏輯，例如消除添加噪音造成的負值（圖 3-11）。

圖 3-11　PHSafe 和後處理流程



（一）家戶規模截斷

為保護家戶特徵的機密性，PHSafe 增加一個未曾應用於其他 2020 年人口普查數據產品的避免揭露步驟，亦即當列入清單中的家戶成員數超過某一個門檻值時，系統將隨機移除戶內成員，直到該家戶符合設定值。門檻值如下：

- 門檻為 10 人之表格：「家戶中人口」（表 PH2）、「家庭中人口」（表 PH4）和「住宅單位中人口」（表 PH7）。
- 門檻為 6 人之表格：「家戶中未滿 18 歲人口」（表 PH3） 和「未滿 18 歲子女」（表 PH6）。

截斷過程根據資料表針對家戶「總人口數」或「未滿 18 歲人口數」採取不同的步驟。首先每個家戶成員隨機分配到一個索引號，並依據該索引號進行排序。針對統計標的為「家戶總人口」的資料表，第 10 位以後的成員會被移除；而對於統計標的為「未滿 18 歲人口」的資料表，則先排除年滿 18 歲的成員，然後未滿 18 歲者按照索引號從低到高排序，移除第 6 位以後的成員。

表 3-8 展示每戶 10 人門檻的截斷過程，家戶 A 有 3 個成員，低於門檻，因此不做任何更動；家庭 B 有 12 個成員，超過 10 人的門檻，因此隨機移除 2 個成員，使其符合 10 人的限制。這種截斷方式能有效保護數據隱私，同時維持統計數據的合理性。

表 3-8　家戶總人口資料表截斷範例

初始家戶	家戶人口編號	處理	截斷後家戶人口編號
家戶A	1	No change.	1
家戶A	2	No change.	2
家戶A	3	No change.	3
家戶B	1	No change.	1
家戶B	2	No change.	2
家戶B	3	Selected at random for removal.	Not included.
家戶B	4	No change.	3
家戶B	5	No change.	4
家戶B	6	No change.	5
家戶B	7	Selected at random for removal.	Not included.
家戶B	8	No change.	6
家戶B	9	No change.	7
家戶B	10	No change.	8
家戶B	11	No change.	9
家戶B	12	No change.	10

資料來源：美國普查局。

全國因為家戶總人口超過 10 人門檻使資料遭截斷者有 126,263 戶（占總戶數 0.1%）。截斷對不同族群的影響程度不同，例如戶長是西班牙裔或拉丁裔家庭，成員數量超過 10 人者計 53,264 戶（占該族裔之 0.3%）；不同地區的截斷影響也存在差異，例如維吉尼亞州因超過門檻而被截斷者有 2,202 戶（占該州 0.1%），而在夏威夷有 3,543 戶（占該州 0.7%）。數據使用者可以比較 S-DHC 的截斷數據與 DHC 的未截斷結果，從而估算出被截斷的人數。

（二）添加噪音

PHSafe 在經過截斷處理的資料上進行特徵別初步統計，然後添加噪音，產製「噪音測量（Noisy measurements）」的初步結果。所添加的噪音量由「隱私損失預算」決定，這個預算設定係為了確保噪音測量的精確度達到預定目標（即「誤差範圍」），並且基於截斷後的人口數據進行，而非原始的人口數據，噪音測量在至少 90%的情況下能夠符合這些預設目標。

在添加噪音的過程中，僅會獨立地添加到最詳細層級的單元格，稱為「最內層單元格（Innermost cells）」，這樣的處理方式確保每個細節都能獲得個別的隱私保護（表 3-9）。在這一步驟中，不同於 TDA 做法，各表格中的小計、總計並不會直接添加噪音，將保留到後處理階段再進行。

表 3-9　僅在最內層的單元格加入噪音

Table element	Detail level
Total:	Total:
Householder, spouse, unmarried partner, or nonrelative	Innermost cell （noise added here）
Own child:	Subtotal:
In married couple family	Innermost cell （noise added here）
In cohabiting couple family	Innermost cell （noise added here）
In male householder, no spouse or partner present family	Innermost cell （noise added here）
In female householder, no spouse or partner present family	Innermost cell （noise added here）
Other relatives:	Subtotal:
Grandchild	Innermost cell （noise added here）
Other relatives	Innermost cell （noise added here）

Source: U.S. Census Bureau.

（三）後處理（Post-processing）

為了減少不合理結果、提高準確性及提供保護隱私相關的不確定性測量，添加噪音的結果將經過一系列「後處理」，以確保最終數據結果在保護個人隱私的前提下，能夠維持一定程度的準確性和合理性。

後處理步驟要求滿足以下標準：

- 數值不得為負數。
- 平均家戶規模至少為 1，因為每個家戶至少必須有 1 個人。
- 平均家庭規模至少為 2，因為每個家庭至少必須有 2 個人。
- 公布的統計數字顯示每個家戶不超過 10 人（未滿 18 歲者則不超過 6 人）。

後處理模型係使用貝氏方法（Bayesian approach），並結合添加噪音之統計結果和限制條件，產生一組新的估算，獨立應用於全國和州層級的表格；在每個表格內，後處理應用於最內層單元格，將這些單元格彙總，產生相應的小計、總計。因為「後處理」修正一些不合理的數據，經過後處理的結果通常比添加噪音後的初步結果更精確，並提供「可信區間（Credible Interval）」指標，用以衡量 S-DHC 統計數據的準確性。

第六節 差分隱私應用

在數據中添加噪音係為了提高機密保護力，但同時也會降低資料準確性，差分隱私的優點就是能夠以量化方式權衡其利弊。美國揭露避免系統（Disclosure Avoidance System, DAS）以差分隱私為基礎，開發 TDA、SafeTab、PHSafe 三種資料保護框架，除了前幾節介紹的人口普查產品外，亦廣泛應用於其他各項產品（表 3-10），以全面保護受訪者隱私。

表 3-10　2020 年美國人口普查產品所使用的重要避免揭露技術

2020 Census Data Products	Components of the 2020 Census Disclosure Avoidance System
Group I Products ● P.L. 94-171 重新劃分選區數據摘要檔案（P.L. 94-171 Redistricting Data Summary File） ● 人口概況（Demographic Profiles） ● 詳細人口與住宅特徵檔案（Demographic and Housing Characteristics File, DHC）	TopDown Algorithm （TDA） 產生受隱私保護的微數據，作為十年製表系統輸入資料。
Group II Products ● 詳細人口與住宅特徵檔案 A（Detailed Demographic and Housing Characteristics File A, Detailed DHC-A） ● 詳細人口與住宅特徵檔案 B（Detailed Demographic and Housing Characteristics File B, Detailed DHC-B） ● 補充人口與住宅特徵檔案（Supplemental Demographic and Housing Characteristics File, S-DHC）	SafeTab、PHSafe 直接產生受隱私權保護的表格。
Group III Products ● 公開使用微數據（Public Use Microdata） ● 研究型統計產品（Research-based Statistical Products） ● 研究人員存取（Researcher Access） ● 2020 普查資料後續使用（Out-year uses of 2020 Census data）	TDA、SafeTab、PHSafe 等其他隱私保護方法。

資料來源：美國普查局。

隨著電腦計算能力的提升，大幅改善統計資料處理的效率，使政府更能強化數據治理的應用。為了滿足各界對資料的殷切需求，公開的統計資料越來越多，以往認為只要不揭露特殊少數人口群體的統計資料，就能有效防止個別資料被辨識，傳統方法有資料遮蔽、資料交換等；然而面對重建資料庫與重新識別的攻擊，許多人口群體亦成為新的攻擊目標，傳統方法的保護效果卻很有限。美國普查局為確保資料具有一定程度的保護力，對於 2020 年人口普查提供的各種產品，開發 TDA、SafeTab 與 PHSafe 等演算法，以因應不同產品的保護需求。每一種演算法都有其特點，TDA 採用地區層級與不變量框架，使範圍越大的地區有越精準的資料；SafeTab 導入自動調整設計，避免發布過度精細的資料；PHSafe 採用截斷技巧，降低住宅單位與個人檔案串聯所增加的風險。差分隱私技術的使用並非一成不變，唯有深入瞭解資料特性與使用情境，才能兼顧資料的隱私保護及其可用性。

第四章　心得與建議

本次赴美國研習重點係探討「地址主檔（MAF）的建置與維護機制」、「差分隱私技術」等議題，深入瞭解地址主檔定期更新方式及地理資訊系統（GIS）的整合運用，大幅減少實地清查地址工作，有效減省人力及成本，並確保抽樣底冊的正確性與完整性，差分隱私技術的創新應用則聚焦於統計資料公布如何在保護個人隱私與數據應用間取得平衡。研習期間承蒙美國普查局、費城地區辦公室長官及專家親自指導，對我國相關工作之規劃提供重要啟發，或可作為未來精進方向，茲將本次研習心得與建議臚列如下：

一、維護地址母體檔的完整性與正確性，以提升調查資料品質

美國普查局在 2020 年人口普查重新改進地址清查流程，實現動態、持續的更新機制，確保地址母體檔的完整性。地址母體的基礎係結合郵政服務數據、地方政府提供的地址更新資訊，以及實地訪查等多元方式，完備抽樣母體，再整合地理資訊系統資料，另獲取地理編碼、經緯度等地理特徵，為實地地址判定或辦理調查提供相關輔助資訊。我國除了擁有完善的地址登記資料（如戶籍地址、村里門牌地址與稅籍地址），還有各類普抽查資料，隨地址資料愈趨完備、整合技術愈趨成熟，或可參考美國辦理方式，逐步建置我國地址母體檔，作為各項調查之抽樣底冊及資料推計依據。

二、善用影像資訊及自動化判定，降低實地清查的人力與成本

為改善 MAF 的覆蓋率及正確性，美國普查局在前一次普查的基礎上啟動地理空間支援系統（GSS）計畫和辦公室內地址清查（IOAC）的程序，與部落、州和地方政府合作，獲取最新地址和道路資料，相關的動態更新程序可以確保地址母體檔正確性。其中辦公室內地址清查程序係透過 5 個核心組成部分的協同運作，相較上次普查需要實地清查的數量大減 35%，並節約近 4.2 億美元。我國除了現有的公務登記資料，如能取得相關影像資訊（如衛星、街景影像），開發比對系統，未來也可以作為支援地址母體檔的更新機制，為日後的調查、普查工作建立更可靠基礎。

三、審慎評估避免資料揭露技術之可行性，降低隱私被重新識別之風險

2020 年美國人口普查中，普查局面臨「如何在發布數據的同時保護個人隱私」的挑戰，因而引入了差分隱私技術。差分隱私技術在商業上的應用已非常廣泛，如 Apple、Google、Microsoft、Uber 等企業為蒐集使用者的偏好和行為，優

化使用者體驗或推薦個人化商品，相關資訊皆使用加入噪音技術。我國人口普查係採「公務登記與調查整合式普查」方式辦理，僅抽取 15%樣本普查區（約 120 萬戶）進行區內全查，發布之地理層級至「鄉鎮市區」，與美國全面普查之方式不同。未來可參考美國做法，研究差分隱私技術在不同資料類型和地理層級上的適用性，惟考量差分隱私的技術門檻相對較高，國內相關研究仍有限，如能強化官學合作，更可兼顧數據開放與隱私保護。

四、加強隱私保護數據說明，協助使用者理解資料使用限制

美國普查局 2020 年人口普查發布的統計結果已經過差分隱私技術處理，為了避免個別資料被識別而在數據中加入噪音，即使相關不合理的結果經過「後處理」，仍有部分不一致情形；普查局在數據說明文件中詳細解釋這些噪音的影響，並提供噪音影響的範圍，協助使用者理解應用相關數據的潛在偏差與不一致性。透過不同差分隱私演算法所產生的數據產品，適用情境亦不盡相同，有些產品將資料彙總後誤差可能較小，有些則相反。我國未來若有應用差分隱私技術，亦應參考美國隱私保護數據說明，協助使用者理解資料之使用限制。