論述 》統計 · 調查

美國人口普查避免個別資料揭露方法之變革

隨著資料探勘及資訊技術的發展，傳統避免個別資料揭露方法之有效性受到挑戰與質疑，如何在強化資料保護與兼顧使用者需求之間取得平衡是越來越困難的課題，美國普查局爰對 2020 年人口普查結果之避免揭露系統進行現代化改造，本文介紹其變革歷程，提供各界參考。

陳艷秋（行政院主計總處國勢普查處專門委員）

壹、前言

聯合國官方統計基本原則強調「統計機構爲統計彙編所蒐集的個體資料，不論涉及自然人還是法人，都應嚴格保密，而且只用於統計目的」。我國及先進國家統計法亦有相關規定，對調查取得之個別資料均有使用限制，並避免發布有洩漏個別資料疑慮的統計資訊，惟隨著資料探勘、大數據分析及資訊技術的發展，傳統避免個別資料揭露方法之有效性受到挑戰與質疑，如何在強化資料保護與兼顧使用者需求之間取得平衡是越來越困難的課題。美國普查局於 2018 年研究發現，過去人口普查（以下簡稱普查）避免揭露（DA,disclosure avoidance）[1] 的方式已無法因應現有的解密技術，爰決定運用新型態之方法來避免 2020 年普查個別資料被揭露，本文介紹其變革歷程，提供各界參考。

貳、2010 年以前普查避免揭露方法

美國早期的普查對於避免揭露著重於防止直接辨識個別資料，主要採刪除辨識碼方式處理，1970 年普查開始採全表抑制方式，取消部分小地區統計表，以因應資料間接洩漏的威脅，1990 年普查導入新的保護方式，對高識別風險資料重新編碼及視爲缺漏資料予以設算，以減少全表抑制情形並將發布資料之地理層級延伸至街廓（block）；2000 年針對日益增長的資安威脅增加了更多的保護措施（下頁表 1）

一、重新編碼和四捨五入（Recoding and Rounding）

公共使用微數據樣本（Public Use Microdata Samples，PUMS）中類別變數之每個類別必須至少代表全國10,000個加權人口或家庭，未達門檻之類別須與其他類別合併；連續變數則可能採重新編碼或將資料四捨五入至設定位數，以減少資料被辨識的風險。

二、超過上下限重新編碼（Topcoding and Bottom-Coding）

通常用來消除PUMS中連續變數的特異值，例如年齡和金額等。超過上限重新編碼係將所有值排序後前0.5%或非零值前3%截斷（以對筆數影響最少者爲準），並以被截斷資料之平均值或中位數取代，被截斷資料至少須包含3筆，否則須降低截斷點至滿足該門檻；低於下限值重新編碼之執行原則類似。

三、地理人口數門檻（Geographic Population Thresholds）

PUMS所有地理級別之加權人口數不得少於100,000人。

四、資料交換（Data Swapping）

將具有高洩漏風險家庭之部分特徵資料與所設定屬性相同的其他家庭進行交換，通常與較大地理級別內不同小地區之另一家庭交換，例如，跨區但在同一郡內。

五、部分合成資料（Partially Synthetic Data）

資料交換應用於團體宿舍（group quarters，GQ，包括住在療養院、監獄、大學宿舍和軍營等家庭以外人口）效果不佳，且可能產生無意義的統計資料，因此普查局採部分合成資料來保護GQ資料。運用線性化模型對原始資料進行建

表1　1960至2010美國普查PUMS應用之避免揭露技術

Decennial censuses	Topcodes and Recodes	Blank and impute	Swapping	Category size thresholds	Noise infusion	Partially synthetic data
1960						
1970						
1980	×					
1990	×	×				
2000	×		×	×	×	
2010						
Households	×		×	×	×	
Group quarters	×			×		×

資料來源：Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples。

模，僅就有揭露風險的變數以模型所產生的資料取代，而其所在地理位置及 GQ 類型須維持不變。

六、添加噪音（Noise Infusion）

2010 年普查僅少部分資料使用添加噪音，主要用於隱藏個人或家庭極端特徵資料，例如，同時生七個孩子的人、15 歲成爲執業醫師的人或人數眾多的家庭，均存在較高個資揭露風險，因此資料處理過程可能會更改這些不尋常資料。

參、2020 年普查避免揭露系統

一、變革緣起

普查保護個別資料機制是否足夠，一直是各界關注的議題，2018 年普查局研究人員以 2010 年普查發布結果進行資料庫模擬重建，研究人員重建了普查所有 3.09 億人的地理位置（即普查街廓 census block）、性別、年齡、種族及西班牙裔 / 拉丁裔，重建結果有 1.44 億人或 46%的人口前揭 5 個變數都與普查回復相同，另 7,600 萬人除了誤差 1 歲外，其餘特徵資料亦相同，進一步將重建資料檔與其他商業資料檔連結，最終可被正確辨識人口約 5,200 萬人或占美國總人口之 17%。顯示傳統避免資料揭露方法已無法防禦新型態資料庫重建技術，且隨著大數據應用的蓬勃發展，私部門亦可能擁有大量個人資料，藉此獲取普查蒐集之受訪者特徵資料風險大增。因此，普查局資料管理執行政策委員會（Data Stewardship Executive Policy Committee，DSEP）決定對 2020 年普查避免揭露系統（Disclosure Avoidance System，DAS）進行現代化改造。

二、主要架構

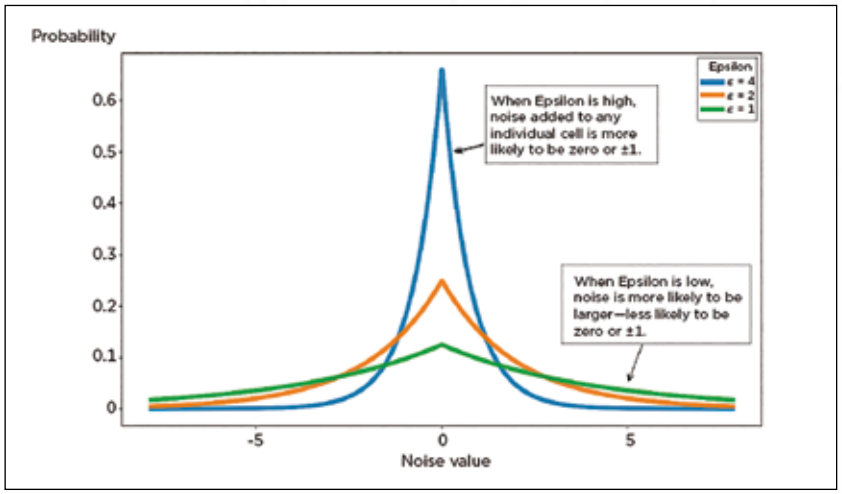
普查局應用「差分隱私」（differential privacy，DP）作爲 2020 年普查避免揭露系統之主要架構，差分隱私是微軟研究院於 2006 年提出的隱私保護模型，主要透過演算法對資料庫資料進行隨機、細微修改（或稱爲添加噪音），以避免資料庫產生之統計結果可反推任何個別資料，但又可維持統計分析之有效性，其原理類似對組成螢幕圖像之數百萬個像素小點進行細微更改，以避免個別小點被正確識別，但仍可保留圖像所要呈現意象。2020 年普查資料是普查局第一個採差分隱私保護的大型數據資料，相較於過去僅對高風險資料進行保護措施，新方法假設資料庫中所有資料都可能被攻擊，因此隨機對資料添加噪音，當添加的噪音量越大時，個別資料被辨識的風險越低，但同時降低資料確度。普查局在差分隱私的基礎上，建構更爲複雜的數學模型，藉由隱私損失預算（privacy-loss budget，PLB）參數設定來精準調控對特徵變數添加之噪音量，以權衡資料保密與確度之取捨，噪音分布由 0 向正負值擴散，所設定 PLB 值越高，噪音值越集中於 0 附近，則產生的資料確度越高，保密性越低（下頁圖 1）。由於差分隱私係於設定之隱私損失預算及噪音分布下，對各特徵資料隨機添加噪音，其原

始碼（source code）和各特徵變數所設定的隱私參數可公開，故資料使用者可據以評估其對資料之可能影響。

三、近程發展

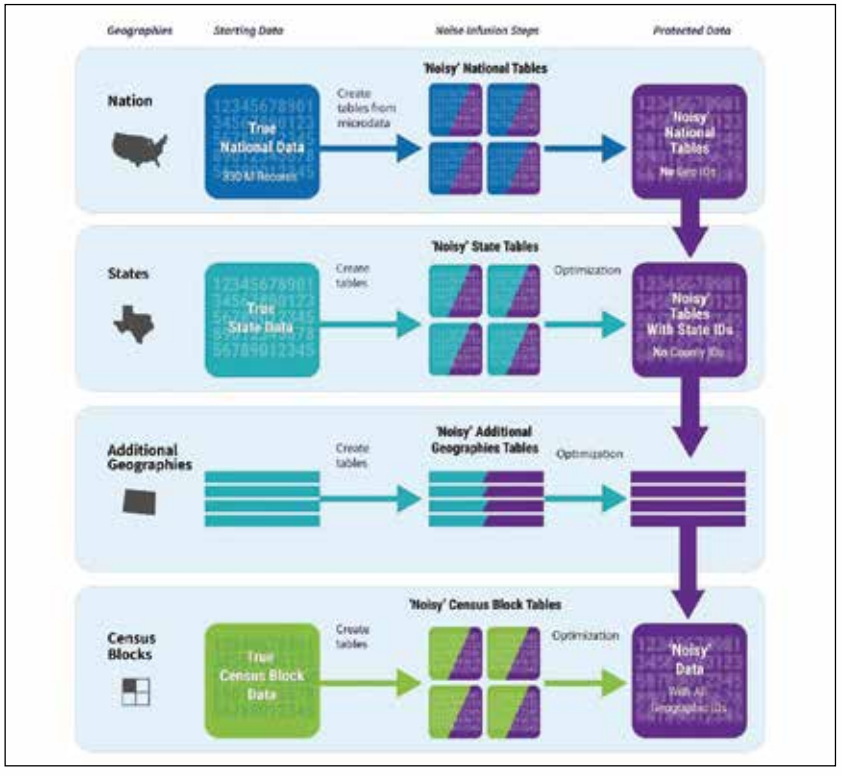
美國憲法規定必須依實際普查結果分配各州眾議院代表席次，爰除各州總人口數、每個普查街廓之住宅數與各類型團體宿舍數依實際調查結果發布外，其餘所有 2020 年普查數據產品均受新版 DAS 保護，目前已發布項目爲提供各州劃分選區之重新劃分資料（Redistricting Data，P.L.94-171），其處理流程採由上向下演算法（TopDown Algorithm，TDA）（圖 2），先進行全國性特徵資料差分隱私，再依普查局對普查資料所設定條件進行後續處理；全國性資料確定後，接著進行州資料之差分隱私及處理，且各州相關變數合計數必須與全國總數相等，以此向下推算各地理層級，最終產生預定發布統計結果之微數據（microdata records），據以編製相關統計結果。由於添加

圖 1　隱私損失預算參數（epsilon）設定與噪音量之關係



資料來源：Disclosure Avoidance for the 2020 Census:An Introduction。

圖 2　2020 年普查重新劃分資料 P.L.94-171 建構差分隱私之過程



資料來源：Disclosure Avoidance for the 2020 Census:An Introduction。

之噪音量與群體大小無關，且小群體被辨識風險相對較高，因此，就地理級別而言，最底層之街廓資料確度受影響程度相對較大，隨地理級別提高及人口數增加，準確度亦越來越高，另街廓資料亦較容易出現住宅使用情形與居住人口數不合、街廓內僅居住未滿 18 歲人口等不合理情況（表 2），普查局於重新劃分資料發布後，亦公開說明使用資料應注意事項及建議，俾利資料使用者評估使用限制。

四、與資料使用者溝通

如何讓資料使用者了解及接受新方法是普查局的另一項

表 2　郡級以下地理級別資料不一致或不合理情形

Inconsistency	Blocks affected		Block groups affected		Tracts affected		Counties affected	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Zero occupied housing units but more than zero household population	392,921	4.80	223	0.09	90	0.11	0	0.00
Zero household population but more than zero occupied housing units	91,415	1.10	30	0.01	17	0.02	0	0.00
Everyone in area under age 18 (excludes areas with group quarters population)	101,127	1.80	27	0.02	17	0.05	0	0.00

資料來源：Disclosure Avoidance for the 2020 Census:An Introduction。

表 3　2010 年普查街廓人口數差異分析－新舊方法比較

Blocks by size	Number of blocks	Mean absolute error (number of people)[1]	Error：middle 90 percent (counts of people)[1]	
			Minus	Plus
All blocks with housing units or GQs	6,398,202	4.89	-11	+10
Blocks with total population between 0-249	6,221,561	4.61	-10	+10
Blocks with total population between 250-749	156,251	13.50	-34	+7
Blocks with total population between 750-1,249	15,294	23.37	-53	+3
Blocks with total population between 1,250-1,749	3,515	28.16	-64	+3
Blocks with total population between 1,750-1,949	524	31.08	-69	+3
Blocks with total population between 1,950-2,049	197	30.00	-73	+2
Blocks with total population between 2,050-2,249	265	29.71	-78	+2
Blocks with total population between 2,250-2,749	323	30.34	-74	+4
Blocks with total population between 2,750-3,249	142	28.61	-81	+3
Blocks with total population at or above 3,250	130	22.32	-80	+3

註：1. Mean absolute error 係 2010 年普查公布之街廓統計結果與運用新方法產生結果之絕對差異平均；middle 90 percent 係指街廓差異值由小至大排序後中間 90%之上下限。
2. 本表僅比較有住宅單元或 GQ 人口之街廓，且不包括波多黎各。
資料來源：Understanding Disclosure Avoidance- Related Variability in the 2020 Census Redistricting Data。

挑戰，美國普查局資料管理執行政策委員會於確定 DAS 方法、如何實施和參數設定前，除徵詢人口及統計相關委員會、專家學者、使用資料之政府機關及權益攸關團體外，並於 2019 年 10 月釋出 2010 普查結果運用 DAS 產生之第一個測試資料（epsilon ＝ 6.0，其中 4.0 用於個人資料、2.0 用於住宅資料），邀請使用者提供回饋意見，依據建議修正後，共釋出 5 次 2010 年普查測試資料。最終版增加並重新分配之隱私損失預算，主要爲提高地理級別與部分種族及西班牙裔 / 拉丁裔資料之準確性，惟對提高街廓資料準確度之建議則有所保留，主因係其隱私風險通常最大。資料管理執行政策委員於 2021 年 6 月 8 日確定 2020 年普查重新劃分資料之 PLB 參數（epsilon ＝ 19.61，其中 17.14 用於個人資料、2.47 用於住宅資料），並於結果發布後再提供確定版 PLB 參數所產生之 2010 年普查重新劃分資料及街廓、郡等資料之新舊方法差異分析（上頁表 3），讓各界進一步了解新方法對資料確度之影響。

肆、結語

保障受訪者隱私不被揭露與滿足使用者對資料品質及內涵需求均爲政府統計調查職責，傳統避免揭露方法主要針對小部分可能遭到識別的資料加強保護措施，隨著資訊技術進步，計算能力顯著提高及演算法的開發，爲資訊安全攻擊開闢了一個令人擔憂的渠道，電腦科學家 Kobbi Nissim 和 Irit Dinur 於 2003 年研究顯示，統計資料庫允許任意查詢不可能不洩露隱私訊息，且僅需很少次的隨機查詢就可能洩漏資料庫之紀錄，因此，當發布太多詳盡的統計資料時，將導致洩漏全部底層資料的風險提高。美國普查局原規劃在 2022 年發布之人口概況及人口與住宅特徵統計結果，爲了建構更周延的資料保護機制及兼顧資料應用價值，將發布日期延至 2023 年 5 月，相關普查結果之發布時程亦將延後。我國近兩次人口及住宅普查係採公務登記及調查整合式普查，對外發布資料之最小地理級別爲鄉鎮市區，個別資料被揭露風險相對較低，惟隨著各界對普查資料之多元化需求，未來提供資料或建置普查資料庫時，如何在滿足使用者需求及確保個別資料安全之間取得平衡，仍是必須持續精進的課題。

註釋

1. 美國普查局將避免揭露（DA,disclosure avoidance）定義爲用於保護受訪者個人信息機密性之過程。

參考文獻

1. 美國普查局網站，A History of Census Privacy Protections。
2. 美國普查局網站，Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples。
3. 美國普查局網站，Disclosure Avoidance for the 2020 Census:An Introduction。
4. 美國普查局網站，Understanding Disclosure Avoidance- Related Variability in the 2020 Census Redistricting Data。
5. 美國普查局網站，2020 Census Disclosure Avoidance System-DEVELOPMENT & RELEASE TIMELINE。❖