內政大數據模擬資料總說明

一、緣由

內政部(以下稱本部)辦理內政大數據連結應用專案計畫，串接跨域資料，運用模型建置、機器學習等技術，進行議題現狀分析，提供更宏觀之政策擬定視角，提升政策擬定品質，精確萃取資訊，掌握最需要幫助的族群，主動遞送個人化服務，將有限資源投放在刀口上，達到精準輔助決策之目標。

為使上述串接資料之使用效益擴及民間，活絡資料的使用度，蒐集民間創意，與政策結合提升民眾福祉，並在個人資料無外洩之疑慮下，達到資訊開放的最大程度，爰以「模擬資料」方式讓各界共享資料整合成果。

二、模擬資料特性

模擬資料，係參照母體資料之整體結構(structure)或態樣(pattern)，而模擬(simulate)出來的非真實性資料集，因此在模擬資料集內，個別資料均是虛擬的，並不存在於母體資料中。但因為模擬時參照母體資料的整體結構或態樣，因此將模擬資料進行平均、占比等聚合(aggregate)運算時，仍能得到與母體資料相同或相近的結果。另模擬資料之欄位格式、代碼等均與母體資料相同，相較於統計資料，使用者不僅能夠更自由地進行各種運算外，模擬資料也可以作為在取得母體資料前之程式撰寫測試練習，俟申請取得母體資料後，程式即能直接套用。

綜合上述，模擬資料具有以下幾點特性：

1.能夠反映母體資料之整體結構或態樣。

2.個別資料均是虛擬產生，無個資洩漏風險。

3.欄位格式、代碼等與母體資料相同，相較於統計資料，可更自由地進行各類交叉運算。

三、模擬資料產製步驟

針對不同的資料集，模擬方法可能略有差異，但原則上產製過程包含以下幾個步驟：

1.計算母體資料之整體結構

在各資料集內，不同欄位可能代表著不同物件的屬性，例如性別、年齡等是「人」的屬性，但建物屋齡、建材是「建物」的屬性，因此需先將不同物件屬性的欄位進行分類後，分開模擬再進行合併。另原直接識別欄位(如身分證號等)則直接產製虛擬代碼，用以標註單一人口或建物。

接著，將連續型欄位進行分組，例如年齡可分為 10 齡組；類別型欄位中分類過細者，則進行合併，例如建物主要建材分類多達 20 多種，適度合併為 5-6 類。

依據上述資料分組與合併結果，分別計算各屬性欄位之母體的聯合分布。

2.依母體結構產製個別虛擬資料

(1)不同屬性欄位分開進行模擬：

依據上述母體整體的聯合分布，分別對各屬性欄位進行模擬。

(2)進行不同屬性欄位之合併：

選取各不同屬性中之關鍵欄位，計算其母體之條件機率，例如性別、年齡等是「人」的屬性，家庭組織型態、戶內人數是「戶」的屬性，計算各特定家庭組織型態下之性別、年齡分布，再依此分布進行兩類欄位之結合。

3.將模擬後之資料進行擬真處理

在步驟 1 中，先將資料重新分組與合併，完成模擬後，再依據母體真實情形或假設均勻分布，將分組資料還原為連續型資料，例如模擬年齡組為 20-29 歲，假設均勻分布，隨機產生出 21 歲或 27 歲等。

4.進行代表性驗證

最後將模擬資料與母體資料進行檢定，檢定其結構是否一致。

四、應用策略

綜合上述，模擬資料在無個資洩漏風險之下，保存母體資料集最多訊息的方法。針對模擬資料之特性及限制，提供應用(使用) 策略如下：

1. 個別資料均是虛擬的，並不存在於真實世界中。
2. 將模擬資料進行聚合(aggregate)運算時，能得到與母體資料相同或相近的結果。
3. 可以作為在取得母體資料前之程式測試。

4. 每份模擬資料均附詳細產製過程及驗證結果。
5. 當母體欄位數過多時，囿於計算過於複雜，可能只會有數個重要欄位會被要求通過驗證，剩餘的欄位其聯合分布未必與母體一致。
6. 在母體中用於串接的 key 值欄位(通常為直接識別欄位)，在模擬資料中係以流水號替代，故無法進行跨表之串接。
7. 模擬資料無法正確反映稀少態樣的正確比例，如使用者欲對稀少態樣進行分析者，結果將失真。

五、釋疑

Q1.為何要產製模擬資料?有何優勢?

A1:

內政大數據整合跨領域資料，包括人口、建物、長照、用電用水、新住民等，並且逐年擴大整合範圍。資料整合範圍愈大，應用效益愈高，惟相對個人資料保護工作愈顯重要。在個人資料保護前提下，欲使資料達到最高的使用效益，以模擬資料為最佳途徑，因為模擬資料中個別資料均是虛擬的，並不存在於母體資料中，但由於模擬時參照母體資料的整體結構或態樣產製，因此將模擬資料進行平均、占比等聚合(aggregate)運算時，仍能得到與母體資料相同或相近的結果。其揭露的訊息量比傳統統計資料更多，卻甚至不會有稀少態樣而產生的間接識別情形。

綜上模擬資料具有以下幾點優勢：

1.能夠反映母體資料之整體結構或態樣。

2.個別資料均是虛擬產生，無個資洩漏風險。

3.欄位格式、代碼等與母體資料相同，相較於統計資料，可更自由地進行各類交叉運算。

Q2.開放模擬資料是否存在個人資料外洩之疑慮?是否有間接識別情形?

A2:

模擬資料是參照母體資料各欄位整體的聯合分布(Joint probability distribution)，模擬出來的非真實性資料集，因此母體的個別資料並不存在於模擬資料裡，故無個人資料外洩之疑慮。

另在傳統統計表中，為避免間接識別，通常會將統計值未滿 3 者進行遮罩或概化，惟在模擬資料產製過程中，模擬之筆數依需求為母

體資料之 5%~10%，因此在母體資料中屬稀少態樣者，例如在某特定區域內 80 歲以上女性、離婚且國小畢業者只有 1 人，在母體資料中占比太低，模擬時很可能會被忽略，即使剛好模擬出這種情形，因模擬資料與母體資料的筆數不同，使用者無法得知此稀少態樣原來應有的筆數是多少，因此也不會有間接識別之情形。

Q3.模擬資料是非真實的資料，為何卻能代表母體?

A3:

承 A2，模擬資料是參照母體資料各欄位整體的聯合分布(Joint probability distribution)而模擬出來的，因此模擬資料之整體結構或態樣會與母體相同，在計算平均數或特定類型之占比時(如各地區 50-59 歲男女比例等)，能得到與母體資料相同或相近的結果，故模擬資料雖非真實資料，在整體上仍能代表母體。

Q4.使用模擬資料有甚麼優點?相較於統計資料或抽樣資料，有何差異?

A4:

如「模擬資料之特性」所述，模擬資料具有「能代表母體資料之整體結構或態樣」、「無個資洩漏風險」及「可更自由地進行各類交叉運算」之優點。

由 A1 可知，模擬資料無個人資料洩漏的風險，保留的資訊及運算靈活度卻較次級資料或統計資料高，對於無法授權使用母體資料之數據分析者，透過模擬資料可進行更多樣化的數據分析。

「抽樣資料」係為減少個資洩漏風險，從母體資料中隨機抽取 5%~10%之筆數，以代表母體整體結構，且同樣具有運算的靈活度，惟抽樣資料中個別資料為真實資料，不似模擬資料均為虛擬，因此抽樣資料之個資洩漏風險較高。

Q5.模擬資料在使用上有甚麼限制?

A5:

模擬資料雖以母體資料各欄位整體的聯合分布產製，惟欄位愈多，聯合分布之態樣也愈多，使得模擬計算所需時間成本愈高，例如「鄉鎮市區」共 368 種、「年齡組別」共 10 組、「教育程度」分 10 種，僅此 3 個欄位，聯合分布就有 36,800 種組合，在模擬 10 個欄位以上時，組合常超過百萬種。因此，一個母體資料集如欄位過多，通常會主觀選擇若干重要欄位(例如 10 個)，以其聯合分布模擬，剩餘欄

位則只以其單一分布進行模擬，再結合起來。此結果在進行檢定時，所選取之 10 個重要欄位不論個別欄位分開檢定或聯合檢定都可以通過，惟剩餘欄位則只會通過個別欄位檢定。意即，當欄位數過多時，囿於計算過於複雜，我們只會要求數個重要欄位通過聯合檢定，剩餘的欄位其聯合分布未必與母體一致。

另模擬資料之個別資料均為虛擬，因此在母體中用於串接的 key 值欄位(通常為直接識別欄位)，在模擬資料中係以流水號替代，此流水號僅具有分辨是否為同一單元之用途，不具識別功能(例如，「戶號」在模擬資料中不能用以識別特定戶，僅能用以分辨哪些人是同一戶)。因此，兩個模擬資料之間無法透過關鍵欄位進行串接。

最後，承 A2 所述，模擬資料無法正確反映稀少態樣的正確比例，如使用者欲對稀少態樣進行分析者，結果將失真。

Q6.因為模擬資料並非由母體資料直接抽樣，欄位間是否會產生邏輯錯誤?

A6:

因模擬的過程中，是依照母體資料之整體結構或態樣而產製，理論上母體資料中不會出現的態樣，模擬資料就不會出現，例如母體資料中沒有「未滿 10 歲、離婚」的人口，則模擬資料也不會出現。

模擬資料若有多種代表不同物件屬性的欄位時，欄位間之互相結合可能會發生一些比例失真，例如性別、年齡等是「人」的屬性，但家庭組織型態、戶內人數是「戶」的屬性，這 2 類欄位分開模擬後，再依兩類欄位的條件機率進行連結(例如特定家庭組織型態下之性別、年齡分布)，連結時只能依據部份關鍵欄位，無法將所有欄位納入考量，因此兩類欄位之結合，部分比例會失真。