內政大數據模擬資料
產製、驗證及應用成果

中華民國 110 年 10 月 14 日

目次

壹、 緣起

內政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，進行議題分析，提供更宏觀之政策擬定視角，以精準輔助決策，提升施政品質。資料領域包括人口、建物、長照、用電用水、新住民等，並逐年擴大整合範圍，提高應用效益，但要如何兼顧個資保護，為政府部門發展大數據資料應用所共同面臨的課題。現行行政院主計總處提供外界使用之農林漁牧業普查外釋資料抽樣檔，及衛生福利部建置之教學模擬資料檔，均採抽樣結合去識別化方式，但部分人士認為此等檔案仍保有部分真實紀錄，可能藉由結合其他資料間接識別特定個人，爰本部統計處研究團隊思考以何種方式提供外界資料較為妥適，以確保個資安全。經評估後，採用產製模擬資料方式供各界應用，因其為虛擬樣本，無個人資料外洩之虞，並建立完整驗證程序，及以適合度檢定搭配其他差異比較指標(如平均絕對誤差 MAE)進行驗證，確保與母體資料分配具高相似性。

以下內文將就本部統計處產製109年「屋主資料」、「人+建物+地理資訊資料」及108年「銀髮安居資料」等3項模擬資料之流程與驗證結果，及其應用情形作完整介紹。

貳、 模擬資料產製與驗證

一、屋主資料

(一) 母體資料簡介

- ❒ 資料集名稱：HOUSE_OWN_SUBSET_SUMY_RAWDATA
- ❒ 資料時點：109年8月底
- ❒ 資料筆數：10,253,917筆

(二) 模擬資料欄位及說明

- ❒ 資料集名稱：HOUSE_OWN_SUBSET_SUMY_RAWDATA _SIMU-LATION.csv
- ❒ 資料時點：109年8月底
- ❒ 資料筆數：1,025,392筆

資料說明：依據109年「屋主資料」之整體結構，模擬約10%之樣本資料，資料欄位說明如下表所示。

表 1「屋主模擬資料」欄位說明表

資料欄位	型態	資料欄位名稱	說明
person_sn	INT	建物所有權人序號	
gender_cd	VARCHAR(1)	性別	1: 男 2: 女 3: 不詳或非自然人
marriage_cd	VARCHAR(1)	婚姻狀況	1: 未婚 2: 有偶 3: 離婚 4: 喪偶 5: 婚姻關係消滅 6: 同性婚姻 7: 終止同性婚姻 8: 同性結婚喪偶
education_cd	VARCHAR(1)	教育程度註記	1: 國中以下 2: 高中職 3: 專科(含二專、三專、五專) 4: 大學 5: 碩博士 6: 不詳及 NULL
age	INT	年齡	以資料年度為基準
first_child_birthday	DATE	第一個小孩出生日期	格式為: YYYY-MM-DD
first_own_house_yr	INT	初次持有房屋年份	
own_yr	INT	持有年數	以資料年度為基準
b_area	NUMERIC	持有面積	該屋主持有該建號之面積(平方公尺)
building_sn	INT	建物序號	
county_cd	VARCHAR(1)	建物所在縣市代碼	
countyntownship_cd	VARCHAR(3)	建物所在鄉鎮市區代碼	

資料欄位	型態	資料欄位名稱	說明
purpose_group_cd	VARCHAR(1)	主要用途分類名稱	A: 住家用 F: 住商用 G: 住工用 L: 國民住宅 Q: 其他
materials_group_cd	VARCHAR(1)	主要建材分類	A: 鋼骨或金鋼筋混凝土 B: 磚造 C: 金屬造 D: 其他
floor_group_cd	VARCHAR(2)	建物總層數組別	0: 地下室 1: 1~5 樓 2: 6~10 樓 3: 11~15 樓 4: 16~20 樓 5: 21 樓以上 99: 不明
register_reason_group_cd	VARCHAR(1)	登記原因分類	A: 第一次登記 B: 買賣 C: 贈與 D: 繼承 E: 拍賣 F: 其他
b_age	INT	屋齡	以資料年度為基準

(三) 模擬資料產製流程

1. 步驟 1：將 17 個模擬資料欄位進行分類

(1) 屋主人口模擬資料

包含性別、婚姻狀況、教育程度、年齡、第一個小孩出生日期、初次持有房屋年份、持有年數、持有面積等 8 個欄位，在相同屋主時均應相同。

(2) 建物模擬資料

包含建物所在縣市、建物所在鄉鎮市區、主要用途分類、主要建材分類、建物總層數組別、登記原因分類、屋齡等7個欄位，在相同建號序號時其資料均應相同。

(3) 虛擬欄位

包含建物所有權人序號、建物序號等2個欄位，與原始資料之身分證號或建號無任何關係，僅用以標註單一屋主或建物，其中相同屋主之建物所有權人序號相同、相同建物之建物序號相同。

2. 步驟2：資料預處理

將婚姻狀況、教育程度進行類別合併，合併結果如「屋主模擬資料」欄位說明表所示。此外，也將年齡、第一個小孩年齡(以第一個小孩出生日期計算資料年度之年齡)、初次持有房屋年份與資料年度之差距(以 2020 減去初次持有房屋年份計算之)、持有年數、持有面積、屋齡轉換為離散型變數，將年齡以20歲切分為5類；第一個小孩年齡以10歲切分6類；初次持有房屋年份與資料年度之差距以10年切分6類；持有年數以10年切分為6類；持有面積以100平方公尺切分為5類；屋齡以10年切分為5類，以利後續產製模擬資料使用。

3. 步驟3：產製屋主人口模擬資料

(1) 取得屋主人口模擬資料階層順序

依「性別→年齡→建物所在縣市→建物所在鄉鎮市區」階層順序，透過 Goodman Kruskal's Gamma 統計檢定方法找出後續階層順序的變數，得到階層結果為「性別→年齡→建物所在縣市→建物所在鄉鎮市區→婚姻狀況→第一個小孩年齡→初次持有房屋年份與資料年度之差距→持有年數→教育程度→持有面積」，將利用這個階層順序來產製屋主「人口特性模擬資料」。

(2) 取得原始資料分布

先產製屋主之「性別、年齡」組合之人數，再依上述之階層順序取得原始資料分布，後續將依此資料分布進行模擬資料產製。

(3) 產製屋主人口模擬資料

第一步依上述取得原始資料之階層分布作為模擬參數，產製「屋主人口特性模擬資料」。第二步依原始資料之屋主持有建物數分布，建立持有建物的「屋主結構模擬資料」，再以「性別、年齡」欄位，將「屋主人口特性模擬資料」串接「屋主結構模擬資料」，得到「屋主人口模擬資料」。

4. 步驟 4：產製建物模擬資料

(1) 取得建物模擬資料階層順序

依「建物所在縣市→建物所在鄉鎮市區」階層順序，透過統計檢定找出後續階層順序的變數，得到階層結果為「建物所在縣市→建物所在鄉鎮市區→主要建材分類→建物總層數組別→屋齡→主要用途分類→登記原因分類」，將利用這個階層順序來產製建物模擬資料。

(2) 取得原始資料分布

先產製建物之「建物所在縣市、建物所在鄉鎮市區」組合之建物數，再依上述之階層順序取得原始資料分布，後續將依此資料分布進行模擬資料產製。

(3) 產製建物模擬資料

第一步依上述取得原始資料之階層分布作為模擬參數，進行「建物特性模擬資料」產製。第二步依原始資料之建物被持有數分布，建立被持有建物的「建物結構模擬資料」，再以「建物所在縣市、建物所在鄉鎮市區」欄位，將「建物特性模擬資料」串接「建物結構模擬資料」，得到「建物模擬資料」。

5. 步驟 5：建立屋主人口模擬資料與建物模擬資料對應關係

(1) 建立對應關係

透過「縣市」進行屋主人口與建物模擬資料整合，依每個屋主持有建物所在縣市，從建物模擬資料中分派至屋主人口模擬資料。然而，有 14,921 筆建物模擬資料不足以分派，則依建物模擬資料之建物所在縣市修正屋主人口模擬資料之建物所在

縣市，並將該筆建物模擬資料分派給該筆屋主人口模擬資料。例如，某一屋主人口模擬資料應對應3筆建物模擬資料，分別在臺北市、臺北市、新北市，但建物模擬資料只剩下臺北市、新北市、桃園市的資料可供分派，則將屋主的一筆臺北市建物資料修改為桃園市，並將這3筆建物模擬資料分派該屋主。

(2) 轉換年齡、第一個小孩年齡、初次持有房屋年份與資料年度之差距、持有年數、持有面積、屋齡等為連續型變數

依年齡組別、持有年數組別、持有面積組別、屋齡組別之類別，在該類別區間中，依母體分布模擬其連續型數值，例如：屋齡組別為「0~9年」，則在0~9這個區間，以母體分布模擬一個屋齡數值，如「0~9年」→「6年」。年齡、第一個小孩年齡、初次持有房屋年份與資料年度之差距、持有年數(係指該筆房屋)、持有面積亦依相同方式模擬其連續型數值。

(3) 資料邏輯校正

因屋主人口模擬資料與建物模擬資料中，「持有年數」不應大於「屋主年齡」或「屋齡」，故在屋主模擬資料中，納入邏輯校正之流程。當發生「持有年數」大於「屋主年齡」或「屋齡」時，強制將「持有年數」調整為同時小於「屋主年齡」及「屋齡」之不小於0的隨機整數，且每個整數被挑選之機率相同。

(四) 模擬資料驗證

1.驗證欄位及方法說明

經研究團隊討論結果，挑選「性別」、「年齡組別」、「建物所在鄉鎮市區」、「婚姻狀況」、「教育程度」、「持有面積」、「屋齡組別」、「主要建材分類」、「建物樓層組別」、「用途分類」作為模擬資料驗證欄位(以下簡稱「驗證欄位」)。

經研究團隊深入研究，分別採用「適合度檢定」、「無母數相關係數」及「平均絕對誤差」共3項方式進行驗證。因模擬資料之母體為1,025多萬筆(屋主資料)，考量「適合度檢定」在大樣本檢定時較為敏感，導致容易拒絕假設，無法通過驗證，且其應用層面通常為檢定抽樣樣本間或與母體間之機率分配；且模擬資料過程中存在邏輯校正之議題，如屋主持有房屋年數不可超過屋主年齡及屋

齡，此時須以人工介入方式調整模擬資料；人工介入時即違反該變數模擬時之隨機性，亦可能造成模擬資料無法完全符合隨機變數之母體分布，以致傳統的統計檢定方法難以應用於本專案模擬資料之驗證。為此，研究團隊嘗試其他驗證方式：「無母數之相關係數」及「平均絕對誤差」，以確保模擬資料符合母體分布。3 項驗證之說明如下：

❒ 適合度檢定(Goodness Of Fit Test)

在抽樣分析中，可藉由卡方適合度檢定來檢定樣本是否服從理論分布，即

H_0: 樣本分布與母體分布一致 $v.s.$ H_1: 樣本分布與母體分布不一致。

❒ 無母數相關係數(Spearman's Rank Correlation Coefficient)

因模擬資料皆以類別變數進行驗證，故採用無母數之方法計算母體與模擬資料分布之相關係數。計算方式如下：

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)}$$

其中，n 為類別數、x_i 為樣本中第 i 個類別占比之排序、y_i 為母體中第 i 個類別占比之排序。以屋主資料之年齡欄位為例，其中占比係以資料筆數計算之，相關係數計算方式如下表所示。

表 2「屋主資料」年齡組別_母體與模擬資料之無母數相關係數計算說明

年齡組別	母體		模擬資料		排序差 $(x_i - y_i)$	r_s
	占比 (p_i)	占比排序 (x_i)	占比 $(\hat{p}_i)$	占比排序 (y_i)		
1	0.428%	6	0.426%	6	0	1
2	12.800%	3	12.810%	3	0	
3	44.937%	1	45.007%	1	0	
4	31.467%	2	31.396%	2	0	
5	3.138%	5	3.113%	5	0	
6	7.231%	4	7.249%	4	0	

❒ 平均絕對誤差(Mean Abslute Error, MAE)

將樣本中各類別之占比與母體相減，取絕對值後再求平均值，即$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{p}_i - p_i|$，其中，$n$為類別數、$\hat{p}_i$為樣本中第$i$個類別之占比、$p_i$為母體中第$i$個類別之占比。以屋主資料之年齡欄位為例，其中占比係以資料筆數計算之，MAE 之計算方式如下表所示。

表 3「屋主資料」年齡組別_母體與模擬資料之平均絕對誤差計算說明

年齡組別	母體占比(p_i)	模擬資料占比($\hat{p}_i$)	占比差絕對值($\lvert\hat{p}_i - p_i\rvert$)	MAE (單位：百分點)
1	0.428%	0.426%	0.002 百分點	0.033
2	12.800%	12.810%	0.009 百分點	
3	44.937%	45.007%	0.071 百分點	
4	31.467%	31.396%	0.071 百分點	
5	3.138%	3.113%	0.025 百分點	
6	7.231%	7.249%	0.018 百分點	

以下分別以「資料筆數」、「人數」及「建物數」進行驗證。

2. 以資料筆數驗證

依前述各驗證欄位檢定結果如下表所示。

表 4「屋主資料」單一欄位之資料筆數驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
gendar_cd	性別	**0.6441**	1.00	0.0255
age_disc	年齡組別	**0.3626**	1.00	0.0327
countyntownship_cd	建物所在鄉鎮市區代碼	**0.2419**	0.9995	0.0031

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
marriage_cd	婚姻狀況	**0.4982**	1.00	0.0231
education_cd	教育程度	**0.4630**	1.00	0.0266
b_area_disc	持有面積	<0.01	1.00	0.1920
b_age_disc	屋齡組別	<0.01	1.00	0.0671
materials_group_cd	主要建材分類	**0.1896**	1.00	0.0260
floor_group_cd	建物樓層組別	<0.01	1.00	0.0304
purpose_group_cd	用途分類	<0.01	1.00	0.0319

註：黑體標註以α=0.05 或 5%為基準評估通過檢定之驗證量值

由上表可見絕大部分欄位之檢定結果皆通過，未通過之欄位之相關性極高，且絕對誤差亦小於 0.5 個百分點，可說明模擬產製之資料與母體資料非常相似。

此外，與屋主特性相關模擬資料欄位，依模擬階層順序組合驗證，結果如下表所示。

表 5 屋主特性欄位組合驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
gendar_cd	性別	<0.01	0.8029	0.0001
age_disc	年齡組別			
countyntownship_cd	建物所在鄉鎮市區代碼			
marriage_cd	婚姻狀況			
education_cd	教育程度			
b_area_disc	持有面積			

註：黑體標註以α=0.05 或 5%為基準評估通過檢定之驗證量值

與房屋特性相關之模擬資料欄位，依模擬階層順序組合驗證，結果如下表所示。

表 6 房屋特性欄位組合驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
countyntownship_cd	建物所在鄉鎮市區代碼	<0.01	0.9285	0.0004
b_age_disc	屋齡組別			
materials_group_cd	主要建材分類			
floor_group_cd	建物樓層組別			
purpose_group_cd	用途分類			

註：黑體標註以α=0.05 或 5%為基準評估通過檢定之驗證量值

在 2 項組合驗證結果中，皆未通過適合度檢定。考量模擬資料與母體資料量皆高達百萬及千萬，受限於卡方檢定在樣本數高時容易拒絕虛無假設，研究團隊另從相關性及 MAE 交互參照評估模擬資料之驗證結果，顯示模擬資料與母體之分布相似程度非常高。

3. 以人數驗證

依屋主特性之驗證欄位檢定結果如下表所示。

表 7「屋主資料」單一欄位之人數驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
gendar_cd	性別	<0.01	1.00	0.409
age_disc	年齡組別	<0.01	1.00	0.637
marriage_cd	婚姻狀況	<0.01	1.00	0.467
education_cd	教育程度	<0.01	0.94	1.067

註：黑體標註以α=0.05 或 5%為基準評估通過檢定之驗證量值

上表屋主特性相關欄位雖皆未通過適合度檢定，但相關係數皆高於 0.9。此外，在 MAE 的評估上，除教育程度欄位較高外，其他欄位皆小於 1 個百分點。在人數驗證中，結果相較於以資料筆數驗證差，主因為產製屋主特性資料時，除人物特質外，尚有「持有面積」欄位，每個屋主持有面積不同，且同一人持有不同房屋之面積亦不完全相同，即以人數進行驗證較不符合資料特性，以致結果較差。

4. 以建物驗證

依建物特性之驗證欄位檢定結果如下表所示。

表 8「屋主資料」單一欄位之建物數驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
countyntownship_cd	建物所在鄉鎮市區代碼	0.01	0.9995	0.003
b_age_disc	屋齡組別	<0.01	0.94	0.293
materials_group_cd	主要建材分類	<0.01	1.00	0.273
floor_group_cd	建物樓層組別	<0.01	1.00	0.326
purpose_group_cd	用途分類	<0.01	1.00	1.8

註：黑體標註以α=0.05 或 5%為基準評估通過檢定之驗證量值

由上表可見各欄位雖皆未通過適合度檢定，但相關係數皆高於 0.9，且除用途分類欄位之 MAE 為 1.8 個百分點，其餘各欄位之 MAE 皆小於 1 個百分點，顯示模擬資料與母體資料之分布相似性頗高。

二、人+建物+地理資訊資料

(一) 母體資料簡介

- ❒ 資料集名稱：MOI_PEOPLE_HOUSE_DTL
- ❒ 資料時點：109 年 8 月底
- ❒ 資料筆數：23,574,082 筆

(二) 模擬資料欄位及說明

- ❒ 資料集名稱：MOI_PEOPLE_HOUSE_DTL_SIMULATION
- ❒ 資料時點：109 年 8 月底
- ❒ 資料筆數：2,357,408 筆
- ❒ 資料說明：依據 109 年「人+建物+地理資訊資料」之整體結構，模擬約 10%之樣本資料，資料欄位說明如下表所示。

表 9「人+建物+地理資訊模擬資料」欄位說明表

資料欄位	型態	資料欄位名稱	說明
person_sn	INT	統號序號	
household_sn	INT	戶號序號	
coun_cd	VARCHAR(5)	戶籍縣市	
addr_city_cd	VARCHAR(8)	戶籍鄉鎮市區	
addr_village_cd	VARCHAR(11)	戶籍村里	詳參電子檔(人房地-村里代碼表.txt)
gender_cd	VARCHAR(1)	性別	1: 男 2: 女
birthday_date	DATE	出生日期	格式為: YYYY-MM-DD
birth_place_cd	VARCHAR(5)	出生地	
education_cd	VARCHAR(1)	教育程度註記	1: 國中以下 2: 高中職 3: 專科(含二專、三專、五

資料欄位	型態	資料欄位名稱	說明
			專) 4: 大學 5: 碩博士 6: 不詳及 NULL
marriage_cd	VARCHAR(1)	婚姻狀況	1: 未婚 2: 有偶 3: 離婚 4: 喪偶 5: 婚姻關係消滅 6: 同性婚姻 7: 終止同性婚姻 8: 同性結婚喪偶
living_type_cd	VARCHAR(1)	原住民身分	0: 無 1: 原住民
household_type_cd	VARCHAR(1)	戶口組織型態_依家庭收支調查	1: 單人戶 2: 夫婦戶 3: 三代戶 4: 祖孫戶 5: 核心戶 6: 單親戶 7: 其他戶
new_disability_category	VARCHAR(1)	身心障礙類別	0: 非身心障礙者; 1: 身心障礙者
low_type_cd	VARCHAR(1)	中低收入戶列冊款別	L: 低收入戶 M: 中低收入戶 N: 非低收、中低收入戶
move_cnt	INT	遷徙次數	
movein_date	DATE	最近遷入日期	格式為: YYYY-MM-DD
first_child_birthday	DATE	第一個小孩出生日期	格式為: YYYY-MM-DD

資料欄位	型態	資料欄位名稱	說明
first_own_house_year	INT	初次持有房屋年份	
child_cnt	INT	出生子女數	含非同戶籍子女(含已過世)
is_living_same_county	VARCHAR(1)	是否與子女戶籍同縣市	0: 否 1: 是
having_house_type_cd	VARCHAR(2)	有無殼分類	A: 標準有殼(本人有殼) B: 標準無殼(本人無殼) C1: 與屋主關係為成年子女/其他人士(住宅為本戶所有) C2: 與屋主關係為成年子女/其他人士(住宅不為本戶所有) D1: 與屋主關係為配偶父母未成年子女(住宅為本戶所有) D2: 與屋主關係為配偶父母未成年子女(住宅不為本戶所有) E: 持有房屋不在戶籍同縣市(住宅不為本戶所有) F: 持有房屋在戶籍同縣市(住宅不為本戶所有) 99: 無法分類
floor_group_cd	VARCHAR(2)	建物總層數組別	0: 地下樓層 1: 1~5 樓 2: 6~10 樓 3: 11~15 樓 4: 16~20 樓 5: 21 樓以上 99: 不明
purpose_group_cd	VARCHAR(1)	主要用途分類名稱	A: 住家用 F: 住商用

資料欄位	型態	資料欄位名稱	說明
			G: 住工用 L: 國民住宅 Q: 其他
materials_group_cd	VARCHAR(1)	主要建材分類	A: 鋼骨或金鋼筋混凝土 B: 磚造 C: 金屬造 D: 其他
b_area	NUMERIC	建物面積(平方公尺)	戶籍所在建物總面積
b_age	INT	屋齡	以資料年份為基準
is_in_active_fault	VARCHAR(1)	斷層帶註記	0: 建物非落在斷層帶 1: 建物落在斷層帶 2: NULL
is_in_lique_faction	VARCHAR(1)	土壤液化註記	0: 建物非落在土壤液化區 1: 建物落在低潛勢土壤液化區 2: 建物落在中潛勢土壤液化區 3: 建物落在高潛勢土壤液化區 4: NULL
is_in_dip_slope	VARCHAR(1)	順逆向坡註記	0: 建物非落在順逆向坡 1: 建物落在順逆向坡 2: NULL

(三) 模擬資料產製流程

1. 步驟 1：將 29 個模擬資料欄位進行分類

(1) 人口模擬資料

包含性別、出生日期、出生地、教育程度註記、婚姻狀況、原住民身分、身心障礙類別、遷徙次數、最近遷入日期、第一個小孩出生日期、初次持有房屋年份、出生子女數、是否

與子女戶籍同縣市、有無殼分類等 14 個欄位，其中出生日期將會透過年齡來進行模擬產製。

(2) 戶口與建物模擬資料

包含戶籍縣市、戶籍鄉鎮市區、戶籍村里、戶口組織型態_依家庭收支調查、中低收入戶列冊款別、建物總層數組別、主要用途分類、主要建材分類、建物面積、屋齡、斷層帶註記、土壤液化註記、順逆向坡註記等 13 個欄位，在相同戶號序號時其資料均應相同。

(3) 虛擬欄位

包含統號序號、戶號序號等 2 個欄位，與原始資料之身分證號或戶號無任何關係，僅用以標註單一人口或戶口，其中相同戶口之戶號序號相同。

2. 步驟 2：資料預處理

將婚姻狀況、教育程度、原住民身分、身心障礙類別、出生地進行類別合併，合併結果如「人+建物+地理資訊模擬資料」欄位說明表所示。此外，也將年齡、遷徙次數、最近遷入時之年齡(以最近遷入日期計算之)、擁有第一個小孩時之年齡(以第一個小孩出生日期計算之)、擁有第一間房時之年齡(以初次持有房屋年份計算之)、出生子女數、建物面積、屋齡轉換為離散型變數，將年齡以 20 歲切分為 5 類；遷徙次數超過 6 次為第 7 類；最近遷入時之年齡以 10 年切分 6 類；擁有第一個小孩時之年齡以 10 歲切分 6 類；擁有第一間房時之年齡以 10 年切分 6 類；出生子女數超過 4 個為第 5 類；建物面積以 100 平方公尺切分為 5 類；屋齡以 10 年切分為 5 類，以利後續產製模擬資料使用。

3. 步驟 3：產製戶長、建物及戶內人口模擬資料

(1) 取得戶長及建物模擬資料階層順序

依「戶口組織型態→戶籍縣市→戶籍鄉鎮市區→戶籍村里→原住民身分」階層順序，透過統計檢定找出後續階層順序的變數，得到階層結果為「是否與子女戶籍同縣市→順逆向坡註記→擁有第一個小孩時之年齡→婚姻狀況→出生子女數→出生地→最近遷入時之年齡→斷層帶註記→土壤液化註記→中低收

入戶→年齡→性別→身心障礙類別→有無殼分類」，將利用這個階層順序來產製戶長及建物模擬資料。此外，部分同一戶口之建物相關欄位(建物面積、屋齡、主要用途分類、主要建材分類、建物總層數組別)，以及單一人口相關欄位(遷徙次數、擁有第一間房時之年齡、教育程度)，則分別以各自原始資料分布進行模擬。

(2) 取得原始資料分布

依上述之階層順序取得原始資料分布，以及戶內人口數分布，後續將依此資料分布進行模擬資料產製。

(3) 產製戶內人口模擬資料

依上述取得之階層原始資料分布以及戶內人口數分布作為模擬參數，產製戶口與建物模擬資料，此時該模擬資料包含戶長(1,175,698 筆資料)與非戶長(1,181,710 筆資料)兩個部分，共有 2,357,408 筆。

4. 步驟 4：產製非戶長人口模擬資料

(1) 取得非戶長人口模擬資料階層順序

依「戶口組織型態→戶籍縣市→戶籍鄉鎮市區→戶籍村里→原住民身分→性別→年齡」階層順序，透過統計檢定找出後續階層順序的變數，得到階層結果為「是否與子女戶籍同縣市→擁有第一個小孩時之年齡→出生子女數→婚姻狀況→出生地→遷徙次數→最近遷入時之年齡→教育程度→擁有第一間房時之年齡→有無殼分類」，將利用這個階層順序來產製非戶長人口模擬資料。此外，身心障礙類別欄位，則以其原始資料分布進行模擬。

(2) 取得原始資料分布

依上述之階層順序取得原始資料分布，後續將依此資料分布進行模擬資料產製。

(3) 產製非戶長人口模擬資料

依上述取得之階層原始資料分布作為模擬參數，以及非戶長人口模擬資料筆數，進行非戶長人口特性模擬資料產製。

5. 步驟 5：建立戶口建物模擬資料與人口模擬資料對應關係

(1) 建立對應關係

透過戶口組織型態及戶籍所在地進行戶口建物與人口模擬資料整合，依不同戶口組織型態與戶籍所在地為條件，將相對應該戶口特性之人口資料分派於該戶。例如：夫妻戶的條件下，該戶中有一位 40 至 60 歲的男性戶長，則會從人口模擬資料中分派一位與該戶長年齡組別相近的已婚女性。資料合併採用逐漸放寬條件之方式進行，故無對應不到關係之資料。

(2) 將離散型變數轉換為連續型變數

依年齡組別、遷徙次數組別、最近遷入時之年齡組別、擁有第一個小孩時之年齡組別、擁有第一間房時之年齡組別、出生子女數組別、持有面積組別、屋齡組別之類別，在該類別區間中，依母體分布模擬其連續型數值，例如：屋齡組別為「0~9 年」，則在 0~9 這個區間，以母體分布模擬一個屋齡數值，如「0~9 年」→「6 年」。其餘離散型欄位亦依相同方式模擬其連續型數值。此外，出生日期則透過模擬得到之年齡為基礎出生年份，再由出生年份隨機模擬出生日期，如年齡為 30 歲，則出生年份為 1990 年，則隨機模擬之出生日期為 1990-03-31。

(3) 資料邏輯校正

在人房地模擬資料中，「戶口組織型態」與戶內人口息息相關，然而產製過程中，無法完全符合「戶口組織型態」之規則，故在產製模擬資料流程上，加入邏輯校正機制，在「戶口組織型態」及「戶長特性」下，強制調整「戶內非戶長之人口特性」，包含「性別」、「年齡」、「婚姻狀況」、「出生子女數」及「是否與子女戶籍同縣市」，例如某戶戶長需要對應一位父親或母親，而非戶長資料中已無符合其父母條件之人口，則隨機挑選並將其年齡強制調整為適合戶長父母之年齡層，以符合資料邏輯。此外，「年齡」亦與「擁有第一個小孩時之年齡」、「最近遷入時之年齡」及「擁有第一間房時之年齡」相關，同樣基於「年齡」，強制調整其他欄位之資料。

(四) 模擬資料驗證

經研究團隊討論結果，挑選「戶口組織型態」、「戶籍鄉鎮市區」、「原住民身分」、「性別」、「年齡組別」、「婚姻狀況」、「建物總層數組別」、「教育程度」、「用途分類」、「建材分類」作為模擬資料驗證欄位(簡稱驗證欄位)。驗證方式與屋主資料相同，請參考屋主資料驗證章結(一、(四)、1)。

以下分別以「資料筆數」及「戶數」進行驗證。

1. 以資料筆數驗證

依前述各驗證欄位檢定結果如下表所示。

表 10「人+建物+地理資訊模擬資料」單一欄位之資料筆數(人數)驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
household_type_cd	戶口組織型態	<0.01	1.00	0.049
addr_city_cd	戶籍鄉鎮市區	1	0.99	0.0003
living_type_cd	原住民身分	<0.01	1.00	0.063
gender_cd	性別	<0.01	1.00	0.279
age_disc	年齡組別	<0.01	1.00	2.369
marriage_cd	婚姻狀況	<0.01	1.00	2.530
floor_group_cd	建物總層數組別	<0.01	1.00	0.130
education_cd	教育程度	<0.01	0.943	0.880
purpose_group_cd	用途分類	<0.01	1.00	0.099
materials_group_cd	建材分類	<0.01	1.00	0.119

註：黑體標註表示因資料邏輯校正，而改變分布之欄位。

由上表可見，因「人+建物+地理資訊模擬資料」筆數較大，以致檢定結果多未通過，然未通過檢定欄位之相關性極高。而以 MAE 來看，「年齡組別」及「婚姻狀況」欄位與母體分布有些偏差，係因其於模擬資料產製過程中有進行邏輯校正所致。

此外，將「人+建物+地理資訊」模擬資料欄位，依模擬階層順序組合驗證，結果如下表所示。

表 11「人+建物+地理資訊模擬資料」組合驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
household_type_cd	戶口組織型態	<0.01	0.645	0.00003
addr_city_cd	戶籍鄉鎮市區			
living_type_cd	原住民身分			
gender_cd	性別			
age_disc	年齡組別			
marriage_cd	婚姻狀況			
floor_group_cd	建物總層數組別			
education_cd	教育程度			
purpose_group_cd	用途分類			
materials_group_cd	建材分類			

「人＋建物＋地理資訊」資料之模擬涉及人口、戶口、建物等資料分別產製後再行組合，模擬流程複雜，相較於較單純的「屋主資料」之組合欄位驗證結果，「人＋建物＋地理資訊」之相關係數驗證結果稍差。主因「性別」、「年齡組別」及「婚姻狀況」欄位在模擬流程中，為確保產出結果之合理性而進行邏輯校正，人工介入時即違反該變數模擬時之隨機性外，組合產製之次數增加，亦可能造成模擬資料無法完全符合隨機變數之母體分布。但依 MAE 之結果來看，顯示本次模擬資料與母體資料仍非常相似。

2. 以戶數驗證

依戶相關欄位之驗證檢定結果如下表所示。

表 12「人+建物+地理資訊」單一欄位之戶數驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
household_type_cd	戶口組織型態	<0.01	0.9643	5.383
addr_city_cd	戶籍鄉鎮市區	<0.01	0.9989	0.005
floor_group_cd	建物總層數組別	<0.01	1.00	0.052
purpose_group_cd	用途分類	0.3958	1.00	0.023
materials_group_cd	建材分類	0.0101	1.00	0.040

在戶相關欄位之驗證結果中，僅用途分類通過檢定，但從相關係數及 MAE 之驗證結果來評估其他欄位，除戶口組織型態落差較大外，其餘欄位與母體之相關性高且絕對誤差小於 0.1 個百分點。由於人房地模擬資料係以人出發進行模擬，故以戶數進行驗證之結果僅供參考。

三、銀髮安居資料

(一) 母體資料簡介

- ❒ 資料集名稱：angels_108_raw
- ❒ 資料時點：108 年 8 月底
- ❒ 資料筆數：3,537,144 筆

(二) 模擬資料欄位及說明

- ❒ 資料集名稱：ANGELS_SIMULATION
- ❒ 資料時點：108 年 8 月底
- ❒ 資料筆數：353,714 筆
- ❒ 資料說明：依據 108 年「銀髮安居資料」之整體結構，模擬約 10% 之樣本資料，資料欄位說明如下表所示。

表 13「銀髮安居資料」欄位說明表

資料欄位	型態	資料欄位名稱	說明
person_sn	INT	統號序號	
is_use_long_term_care	VARCHAR(1)	有無使用長照	0: 無 1: 有
age	INT	年齡	以資料年份為基準
disability_lv	VARCHAR(1)	身心障礙程度	0: 無身心障礙; 1-5: 有身心障礙(數值愈低代表愈嚴重)
family_type	VARCHAR(1)	家庭型態	1: 獨居 2: 老老照顧 3: 其他
child_cnt	INT	子女數	
is_living_same_county	VARCHAR(1)	子女是否住在同縣市	0: 沒有子女住在同縣市 1: 有子女住在同縣市
low_type_cd	VARCHAR(2)	低收或中低收入戶	0-4: 低收入戶(數值愈低代表愈嚴重)

資料欄位	型態	資料欄位名稱	說明
			5: 中低收入戶 99: 非低收中低收入戶
having_house_type	VARCHAR(1)	有無殼類別	1: 標準有殼 2: 有住宅(本縣市)但非所在戶籍 3: 有住宅(外縣市)但非所在戶籍 4: 戶籍地所有者之配偶、父母及未成年子女 5: 戶籍地所有者之成年子女及其他親友 6: 無法判斷 7: 標準無殼
build_age	INT	屋齡	以資料年份為基準
is_apartment	VARCHAR(2)	是否為無電梯公寓	以戶籍地是否為 5 樓建築，且住在非 1 樓住戶做判斷。 0: 不符合上述條件 2-5: 符合上述條件(數值愈高表示居住樓層愈高) 99: NULL 或無法判斷
bus	NUMERIC	與交通站牌距離	數值愈高表示距離交通站牌愈遠，-1 則為無法判斷。 1: 500 公尺以上才有公車站牌 0.8: 400-500 公尺才有公車站牌 0.6: 300-400 公尺才有公車站牌 0.4: 200-300 公尺才有公車站牌 0.2: 100-200 公尺才有公車站牌 0: 100 公尺內有公車站牌

資料欄位	型態	資料欄位名稱	說明
store	NUMERIC	與零售商距離	數值愈高表示距離交通站牌愈遠，-1 則為無法判斷。 1: 500 公尺以上才有便利超商 0.8: 400-500 公尺才有便利超商 0.6: 300-400 公尺才有便利超商 0.4: 200-300 公尺才有便利超商 0.2: 100-200 公尺才有便利超商 0: 100 公尺內有便利超商
hospital	NUMERIC	與醫院診所距離	數值愈高表示距離交通站牌愈遠，-1 則為無法判斷。 1: 1000 公尺以上才有醫院或診所 0.8: 800-1000 公尺才有醫院或診所 0.6: 600-800 公尺才有醫院或診所 0.4: 400-600 公尺才有醫院或診所 0.2: 200-400 公尺才有醫院或診所 0: 200 公尺內有醫院或診所
lique	VARCHAR(1)	土壤液化區	0: 建物非落在土壤液化區 1: 建物落在低潛勢土壤液化區 2: 建物落在中潛勢土壤液化區 3: 建物落在高潛勢土壤液

資料欄位	型態	資料欄位名稱	說明
			化區 4: NULL

(三) 模擬資料產製流程

1. 步驟 1：資料預處理

將年齡、屋齡轉換為離散型變數，將年齡以 10 歲切分為 5 類；屋齡以 20 年切分為 5 類，以利後續產製模擬資料使用。

2. 步驟 2：產製銀髮安居模擬資料

(1) 取得銀髮安居模擬資料階層順序

依「家庭型態→有無使用長照」為初始欄位，透過統計檢定找出後續階層順序的變數，得到階層結果為「子女是否住在同縣市→低收或中低收入戶→子女數→年齡組別→與醫院診所距離→與交通站牌距離→與零售商距離→有無殼類別→是否為無電梯公寓→土壤液化區→屋齡組別→身心障礙程度」，將利用這個階層順序來產製銀髮安居模擬資料。

(2) 取得原始資料分布，並產製模擬資料

依上述之階層順序取得原始資料分布，作為模擬參數，以產製銀髮安居模擬資料。

3. 步驟 3：將離散型變數轉換為連續型變數

依年齡組別、屋齡組別之類別，在該類別區間中，依母體分布模擬其連續型數值，例如：屋齡組別為「0~9 年」，則在 0~9 這個區間，以母體分布模擬一個屋齡數值，如「0~9 年」→「6 年」。其餘離散型欄位亦依相同方式模擬其連續型數值。

(四) 模擬資料驗證

經研究團隊討論結果，挑選「家庭型態」、「有無使用長照」、「年齡組別」、「有無殼類別」、「身心障礙程度」作為模擬資料驗證欄位(簡稱驗證欄位)。驗證方式與屋主資料相同，以「資料筆數」驗證欄位之檢定結果如下表所示。

表 14「銀髮安居模擬資料」單一欄位之資料筆數(人數)驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
family_type	家庭型態	0.684	1	0.040
is_use_long_term_care	有無使用長照	0.578	1	0.022
age_disc	年齡組別	0.229	1	0.056
having_house_type	有無殼類別	0.432	1	0.035
disability_lv	身心障礙程度	0.420	1	0.023

由上表可見，銀髮安居模擬資料之 5 項驗證欄位皆通過檢定，且 MAE 最高為 0.056 個百分點，可說明模擬資料之驗證欄位符合母體分布。

此外，將銀髮安居模擬資料欄位，依模擬階層順序組合驗證，結果如下表所示。

表 15「銀髮安居模擬資料」欄位組合驗證結果表

資料欄位	資料欄位名稱	適合度檢定 P 值	Spearman's 等級相關係數	MAE (單位：百分點)
family_type	家庭型態	0.197	0.977	0.00002
is_use_long_term_care	有無使用長照			
age_disc	年齡組別			
having_house_type	有無殼類別			
disability_lv	身心障礙程度			

組合欄位之驗證，同樣通過檢定，且其相關係數高於 0.9，MAE 亦相當小，顯示本次模擬資料與母體資料非常相似。

參、 模擬資料與真實資料比較

由於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)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。因此本部建教合作學校中信金融管理學院蔡明春教授之研究團隊，即分別利用本部大數據模擬資料與真實資料，進行縣市別與鄉鎮市區別之高齡化與獨居化統計，並進一步利用集群分析瞭解各地區社會再生性與經濟再生性之差異，以下為利用模擬資料與真實資料分析結果之比較。

一、高齡化比較分析

(一)六都

模擬資料	真實資料	比較說明
臺北市	臺北市	六都高齡化順序完全相同。
高雄市	高雄市	
臺南市	臺南市	
新北市	新北市	
臺中市	臺中市	
桃園市	桃園市	

(二)非六都

模擬資料	真實資料	比較說明
嘉義縣	嘉義縣	非六都高齡化順序從第四名起順序有所不同。
雲林縣	雲林縣	
南投縣	南投縣	
臺東縣	屏東縣	
屏東縣	臺東縣	
花蓮縣	花蓮縣	
澎湖縣	宜蘭縣	
宜蘭縣	苗栗縣	
基隆市	澎湖縣	
苗栗縣	基隆市	

金門縣	彰化縣	
彰化縣	嘉義市	
嘉義市	金門縣	
連江縣	新竹縣	
新竹市	新竹市	
新竹縣	連江縣	

(三)鄉鎮市區

模擬資料	真實資料	比較說明
新北市平溪區	新北市平溪區	鄉鎮市區從第三名起高齡化順序有所不同。
高雄市田寮區	高雄市田寮區	
苗栗縣西湖鄉	臺南市左鎮區	
苗栗縣獅潭鄉	苗栗縣獅潭鄉	
臺南市左鎮區	臺南市龍崎區	
新北市雙溪區	新北市雙溪區	
嘉義縣鹿草鄉	新竹縣峨眉鄉	
嘉義縣義竹鄉	嘉義縣鹿草鄉	
雲林縣水林鄉	嘉義縣六腳鄉	
新北市坪林區	嘉義縣義竹鄉	
花蓮縣豐濱鄉	臺南市大內區	
雲林縣四湖鄉	高雄市美濃區	
高雄市美濃區	新北市坪林區	
高雄市杉林區	臺南市後壁區	
雲林縣元長鄉	雲林縣水林鄉	
嘉義縣六腳鄉	彰化縣大城鄉	
臺東縣長濱鄉	花蓮縣鳳林鎮	
臺南市將軍區	臺南市白河區	
臺南市玉井區	雲林縣元長鄉	
花蓮縣富里鄉	苗栗縣西湖鄉	

二、高齡獨居化比較分析

(一)六都

模擬資料	真實資料	比較說明
臺北市	臺北市	六都高齡獨居化順序完全相同。
新北市	新北市	
高雄市	高雄市	

臺南市	臺南市	
桃園市	桃園市	
臺中市	臺中市	

(二)非六都

模擬資料	真實資料	比較說明
基隆市	基隆市	非六都高齡獨居化順序從2~12 順序多數不相同。
嘉義市	臺東縣	
花蓮縣	花蓮縣	
臺東縣	宜蘭縣	
宜蘭縣	嘉義市	
新竹市	南投縣	
澎湖縣	嘉義縣	
嘉義縣	澎湖縣	
南投縣	雲林縣	
雲林縣	新竹市	
新竹縣	苗栗縣	
苗栗縣	新竹縣	
屏東縣	屏東縣	
彰化縣	彰化縣	
金門縣	金門縣	
連江縣	連江縣	

(三)鄉鎮市區

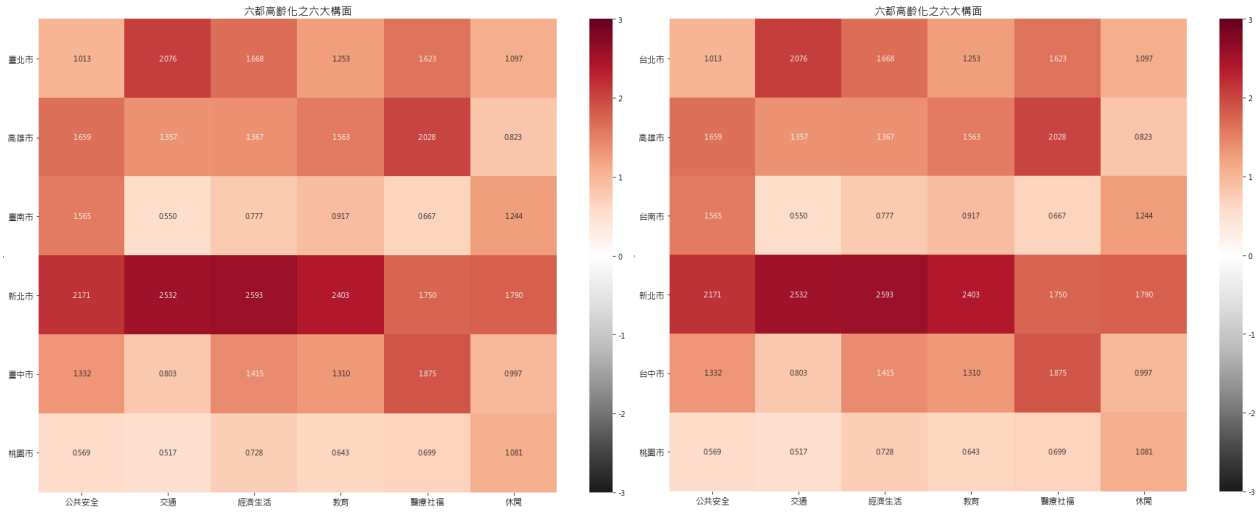
模擬資料	真實資料	比較說明
新北市平溪區	新北市平溪區	鄉鎮市區除第一名外高齡獨居化順序均不相同。
新北市雙溪區	新北市石碇區	
高雄市田寮區	臺中市中區	
新北市石碇區	臺中市和平區	
臺南市東山區	高雄市六龜區	
高雄市六龜區	高雄市田寮區	
高雄市鹽埕區	新北市三芝區	
高雄市前金區	嘉義縣大埔鄉	
南投縣鹿谷鄉	高雄市鹽埕區	

臺東縣東河鄉	新北市雙溪區	
嘉義縣大埔鄉	高雄市前金區	
新北市坪林區	臺東縣太麻里鄉	
屏東縣車城鄉	花蓮縣瑞穗鄉	
臺南市後壁區	基隆市仁愛區	
嘉義縣義竹鄉	臺北市中山區	
雲林縣水林鄉	臺北市萬華區	
嘉義縣鹿草鄉	臺南市東山區	
嘉義縣六腳鄉	南投縣鹿谷鄉	
臺南市大內區	新竹縣五峰鄉	
花蓮縣光復鄉	基隆市中正區	

三、不同高齡化地區單一時點資源排名

以下就兩階段集群分析之結果比較模擬資料與真實資料分析結果。

(一)六都高齡化分析結果比較

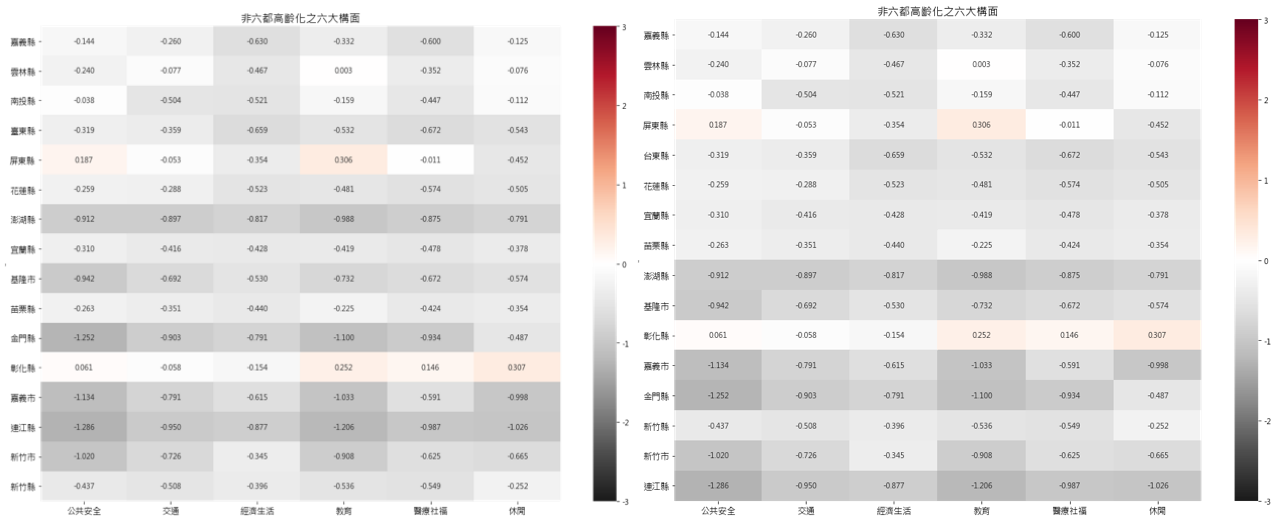


(a)模擬資料　　(b)真實資料

圖 1 不同高齡化程度六都於六大資源之比較

由上面兩圖可看出真實資料及模擬資料六都高齡化程度排名相同，因此六大構面分析也相同。

(二)非六都高齡化分析結果比較

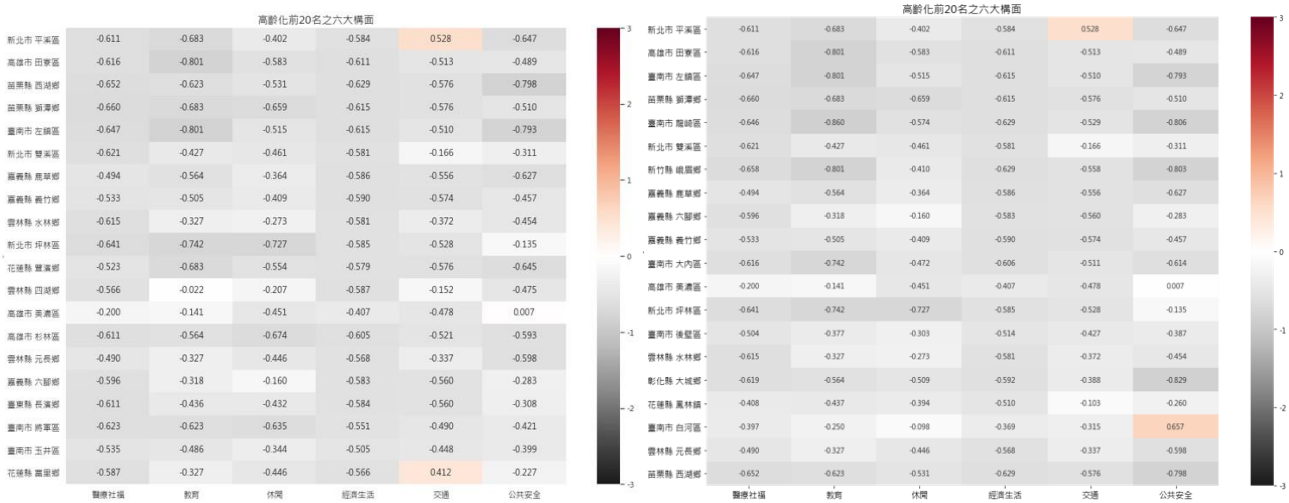


(a)模擬資料　　(b)真實資料

圖 2 不同高齡化程度非六都於六大資源之比較

由上面兩圖可看出，除嘉義縣、雲林縣、南投縣、花蓮縣及新竹市高齡化程度排名真實資料與模擬資料相同外，其餘排名皆不相同，**但皆可看出高齡化程度與六大資源構面具有正向的相關性**。

(三)鄉鎮市區高齡化分析結果比較



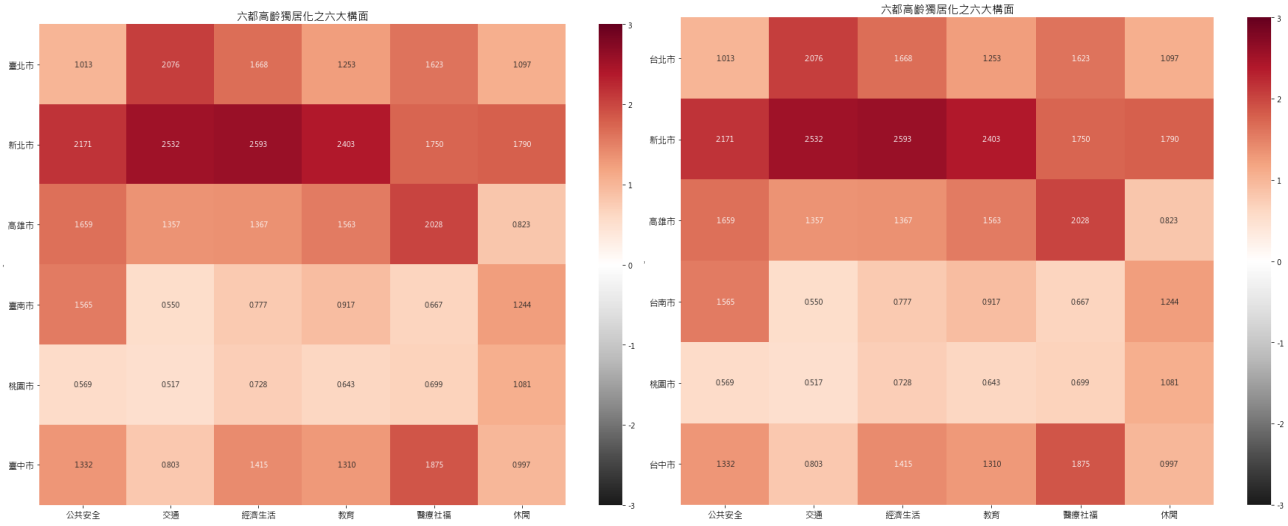
(a)模擬資料　　(b)真實資料

圖 3 不同高齡化程度鄉鎮市區於六大資源之比較分析

由上面兩圖可看出，僅有新北市平溪區、高雄市田寮區、苗栗縣獅潭鄉及新北市雙溪區高齡化程度排名真實資料與模擬資料相同外，其餘排名皆不相同，但兩張圖皆可看出高齡化程度較**高地區六大資源構面表現相對較差**(顏色多為灰色或深灰色)。

四、不同高齡獨居化地區單一時點資源排名

(一)六都高齡獨居化分析結果比較

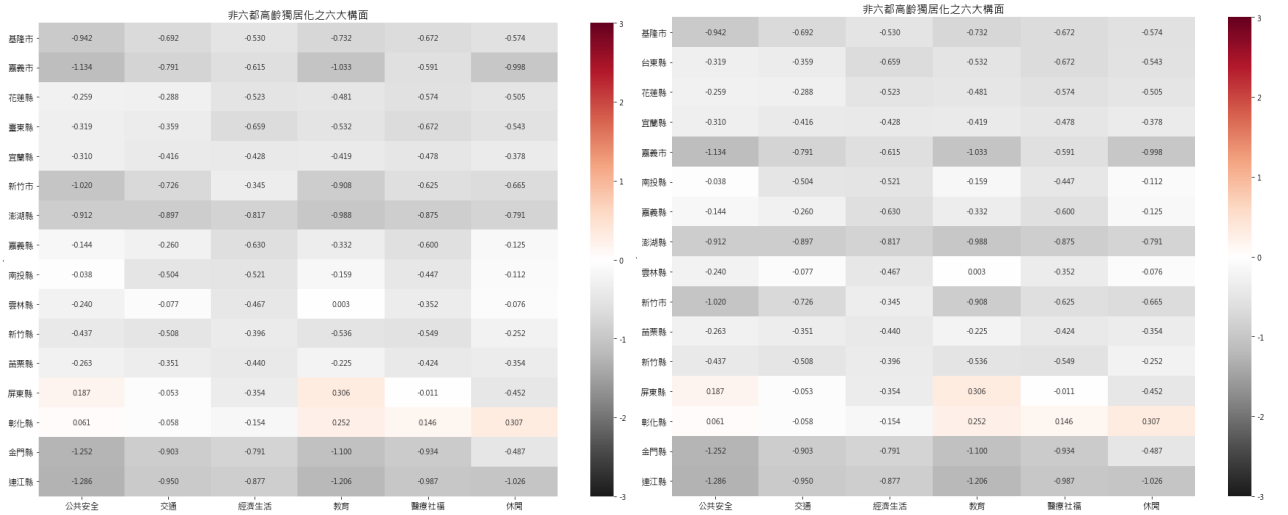


(a)模擬資料　　(b)真實資料

圖 4 不同高齡獨居化程度六都於六大資源之比較分析

由上面兩圖可看出真實資料及模擬資料六都高齡獨居化程度排名相同，因此六大構面分析也相同。

(二)非六都高齡獨居化分析結果比較



(a)模擬資料　　(b)真實資料

圖 5 不同高齡獨居化程度非六都縣市於六大資源之比較分析

由上面兩圖可看出，除基隆市、花蓮縣、屏東縣、彰化縣、金門縣及連江縣高齡獨居化程度排名真實資料與模擬資料相同外，其

餘排名皆不相同，但兩張圖**排除離島(澎湖、金門及連江縣)外，皆可看出高齡化程度與六大資源構面具有負向的相關性**。

(三)鄉鎮市區高齡獨居化分析結果比較

(a)模擬資料　　(b)真實資料

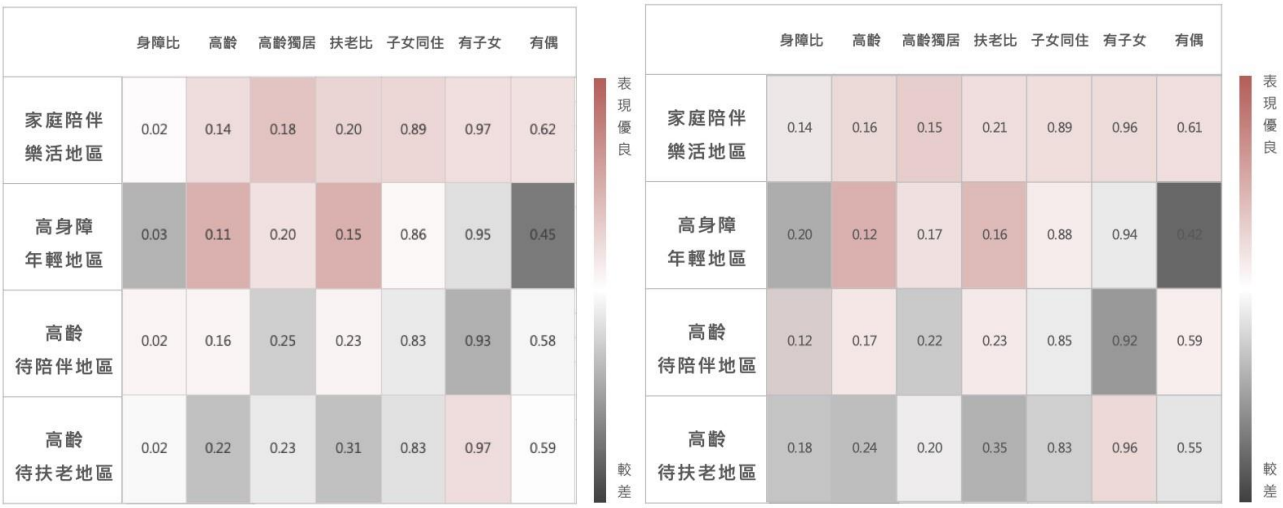
圖 6 不同高齡獨居化程度鄉鎮市區於六大資源之比較分析

由上面兩圖可看出，除新北市平溪區及新北市石碇區高齡獨居化程度排名真實資料與模擬資料相同外，其餘排名皆不相同，但兩張圖之圖形顏色的分散程度皆**難以看出高齡獨居化與六大資源構面具有相關性**。

五、地方創生之集群分析

以下就兩階段集群分析之結果比較模擬資料與真實資料分析結果。

(一)第一階段分群結果比較



(a)模擬資料　　(b)真實資料

圖 7 第一階段集群分析之比較

由上圖可得知，大部分的指標在集群表現差異不大，僅身障比在模擬資料中遠高於真實資料。

(二)第二階段分群結果比較

以下將比較二階段集群分析之結果，將分述如下：

1.「家庭陪伴樂活地區」分群結果比較

第二階命名	醫療社福	教育	休閒	經濟生活	交通	公共安全	代表性鄉鎮市區	
家庭樂活陪伴機能充足	3.15	2.80	1.99	3.44	2.71	2.33	新北 板橋	桃園 中壢
家庭樂活陪伴全機能良好	0.97	0.92	0.78	0.87	0.51	0.86	苗栗 竹南	宜蘭 羅東
家庭樂活陪伴全機能平庸	0.00	0.04	-0.03	-0.03	-0.11	0.00	桃園 觀音	彰化 鹿港
家庭樂活陪伴全機能不足	-0.48	-0.55	-0.32	-0.47	-0.46	-0.78	臺中 神岡	新竹 芎林

圖 8 家庭樂活陪伴地區第二階分群結果表格(模擬資料)

第二階段命名	醫療社福	教育	休閒	經濟生活	交通	公共安全	代表性鄉鎮區	
家庭陪伴樂活地區全機能充足	3.25	2.88	2.02	3.46	2.59	2.57	桃園中壢	桃園桃園
家庭陪伴樂活地區全機能良好	1.02	0.96	0.83	0.92	0.54	1.00	新竹竹北	桃園大園
家庭陪伴樂活地區全機能平庸	0.02	0.07	0.08	0.00	-0.08	0.05	台南善化	高雄仁武
家庭陪伴樂活地區全機能不足	-0.47	-0.51	-0.31	-0.46	-0.45	-0.67	連江北竿	金門金寧

圖 9 家庭樂活陪伴地區第二階分群結果表格(真實資料)

由上面兩個表格可看出，基本上分群結果類似，即模擬資料與真實資料在各群表現差異不大。

2. 「高身障年輕地區」分群結果比較

第二階命名	醫療社福	教育	休閒	經濟生活	交通	公共安全	代表性鄉鎮市區	
高身障年輕機能缺乏	-0.58	-0.66	-0.47	-0.61	-0.44	-0.45	臺東 金鋒	高雄 茂林
高身障年輕公安健全機能不足	-0.47	-0.01	-0.19	-0.55	-0.43	1.65	宜蘭 大同	南投 仁愛

圖 10 高身障年輕地區第二階分群結果表格(模擬資料)

第二階段命名	醫療社福	教育	休閒	經濟生活	交通	公共安全	代表性鄉鎮區	
高身障年輕地區機能缺乏	-0.59	-0.68	-0.53	-0.61	-0.50	-0.51	屏東春日	花蓮新城
高身障年輕地區公安健全機能不足	-0.40	0.00	-0.22	-0.53	-0.39	0.89	桃園復興	花蓮秀林

圖 11 高身障年輕地區第二階分群結果表格(真實資料)

由上面兩個表格可看出，基本上分群結果類似，即模擬資料與真實資料各群表現差異不大。

3. 「高齡待陪伴地區」分群結果比較

第二階命名	醫療社福	教育	休閒	經濟生活	交通	公共安全	代表性鄉鎮市區	
高齡待陪伴全機能不足	-0.33	-0.34	-0.16	-0.26	-0.20	-0.21	桃園 復興	新北 烏來
高齡待陪伴全機能良好	1.49	1.28	0.83	1.46	1.50	1.20	高雄 左營	桃園 龜山
高齡待陪伴全機能充足	2.33	2.62	1.67	2.72	2.92	2.23	臺北 中正	桃園 桃園
高齡待陪伴經濟醫療良好	1.02	0.54	0.29	1.04	0.23	0.74	新北 三峽	高雄 鼓山

圖 12 高齡待陪伴地區第二階分群結果表格(模擬資料)

第二階段命名	醫療社福	教育	休閒	經濟生活	交通	公共安全	代表性鄉鎮區	
高齡待陪伴地區全機能不足	-0.34	-0.42	-0.23	-0.35	-0.20	-0.32	新北八里	基隆信義
高齡待陪伴地區全機能良好	1.43	1.29	0.87	1.44	1.46	1.35	新北淡水	新北汐止
高齡待陪伴地區全機能充足	2.19	2.56	1.62	2.64	3.01	2.17	新竹東區	新北新店
高齡待陪伴地區經濟醫療良好	0.93	0.56	0.23	0.92	0.24	0.71	新北林口	台南中西

圖 13 高齡待陪伴地區第二階分群結果表格(真實資料)

由上面兩個表格可看出，基本上分群結果類似，各群表現也差異不大，即模擬資料與真實資料各群表現差異不大。

4.「高齡待扶老地區」分群結果比較

第二階命名	醫療社福	教育	休閒	經濟生活	交通	公共安全	代表性鄉鎮市區	
高齡待扶老全機能匱乏	-0.58	-0.59	-0.47	-0.57	-0.45	-0.73	花蓮 光復	臺東 鹿野
高齡待扶老經濟醫療不足	-0.17	0.04	0.21	-0.29	-0.03	0.42	苗栗 通霄	苗栗 後龍
高齡待扶老全機能不足	-0.48	-0.32	-0.29	-0.50	-0.32	-0.16	南投 鹿谷	臺東 長濱

圖 14 高齡待扶老地區第二階分群結果表格(模擬資料)

第二階段命名	醫療社福	教育	休閒	經濟生活	交通	公共安全	代表性鄉鎮區	
高齡待扶老地區全機能缺乏	-0.55	-0.49	-0.40	-0.55	-0.40	-0.53	澎湖湖西	新北坪林
高齡待扶老地區全機能普通	-0.12	0.01	0.13	-0.21	0.00	0.24	嘉義朴子	嘉義水上

圖 15 高齡待扶老地區第二階分群結果表格(真實資料)

由上面兩個表格可看出，真實資料少去經濟醫療不足與全機能不足的集群，多了全機能普通的集群，高齡待扶老地區為整體分群中差異最大的部分。

肆、 供內政黑客松競賽應用

為提升110年內政黑客松競賽資料應用層級，本部統計處特別釋出5個內政大數據模擬資料集，供參賽團隊與社會經濟空間資料(SEGIS)加值應用。

關於SEGIS　推動成果　資料與服務　統計地圖

現在位置：首頁 > 最新消息 > 最新消息詳細資訊

歡迎揪團參加『2021資料創新應用競賽內政黑客松』 最高可獨得新臺幣35萬元獎金

發布日期：2021/2/24

★2021年資料創新應用競賽分成6組，歡迎來參加內政部所設立的「內政黑客松」，參賽團隊可運用內政部「國土資訊系統社會經濟資料服務平台」(SEGIS)之資料或服務，結合「內政大數據模擬資料」及「電信信令統計模擬資料」，提出解決或改善社會議題之創作，最高可獨得新臺幣35萬元獎金，本部並擇優秀作品納入2022年大數據專案相關計畫，說明如下：

1.內政部「國土資訊系統社會經濟資料服務平台」(https://segis.moi.gov.tw)提供8萬多項開放資料及服務(如工商、學校、醫療院所、社會福利等點位資料，以及國土利用調查、綜合所得申報等統計資料)，另內政大數據資料彙整國民自出生、就學、婚姻生育、家戶組成及建物登記，至養護、終老、低度用水用電等資料；電信信令資料包含平/假日夜間停留人口、日間活動人口及特定區域旅次等人口分布情形，皆可與SEGIS結 合作深化應用，發想解決民生議題。

2.本競賽除提供SEGIS資料產製之村里別、最小統計區別及銀髮安居資料集外，另特別釋出5個「模擬資料」集，供參賽團隊加值應用，其資料欄位格式說明參附件。 海選階段可依據欄位說明提出應用構想，海選後將提供晉級團隊完整欄位內容。

3.參賽作品須運用SEGIS之資料或服務，與本處提供之內政大數據、電信信令等模擬資料，透過挖掘、重組、混搭等方式，建立分析模型、Web或APP(iOS、Android)服務。

4.參賽團隊可申請統計地圖API，並於雲端進行統計地圖與圖表的展繪運用，申請網址及相關說明：https://segis api.moi.gov.tw。
參賽團隊可使用Amazon S3 (物件儲存服務) 與 Amazon SageMaker (機器學習服務)做為 AI 模型訓練平台與資料儲存空間。

5.海選晉級隊伍需於決選日前繳交「決選簡報檔(.PPTX)」及「2分鐘閃電秀短片(.MP4)」各一份。

●Amazon S3 基本說明：
https://aws.amazon.com/tw/s3/
●Amazon S3 線上學習資源：
https://aws.amazon.com/tw/s3/developerresources/
●Amazon SageMaker 基本說明：
https://aws.amazon.com/tw/sagemaker/
●Amazon SageMaker 線上學習資源
https://aws.amazon.com/tw/sagemaker/developerresources/

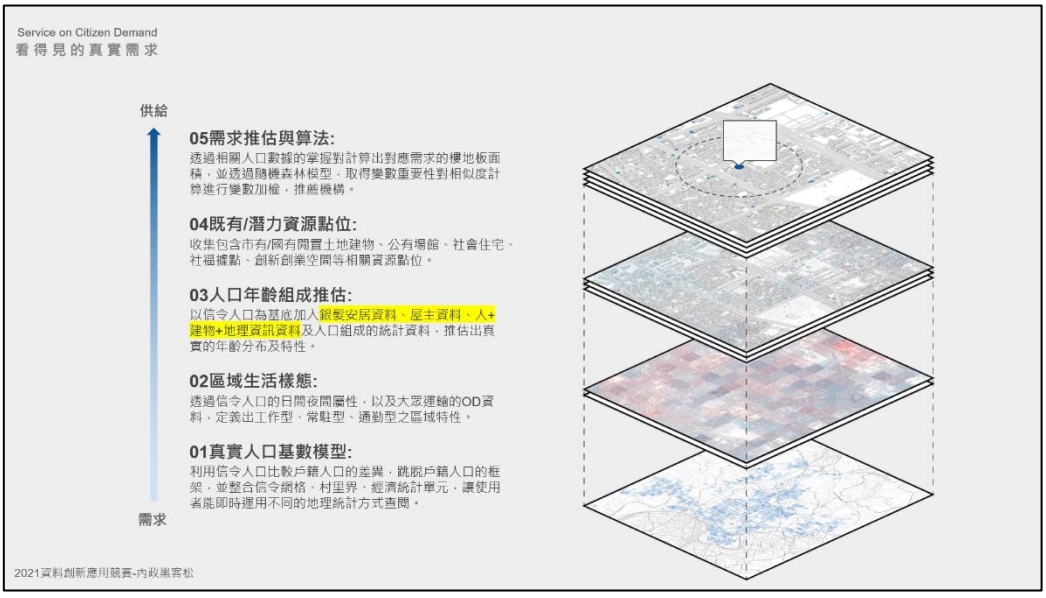
★報名期限至110年4月14日(三)17:00止，相關訊息請參閱：(報名網站)、(2020年競賽得獎作品)。

★本組設有高額獎勵及獎金，競賽獎勵除金獎(15萬元獎金)、銀獎(8萬元獎金)及銅獎(5萬元獎金)各1名，榮譽獎(1萬元獎金)2名外；另設立「電信資料應用特別獎」(10萬元獎金)及「統計寫手特別獎」(10萬元獎金)，可與其他獎項重複，最高可獨得新臺幣35萬元獎金。獲獎團隊另頒發價值1,000美元之AWS續優點數，並將擇具有發展潛力之作品，納入本部2022年大數據專案相關計畫執行。無論公司企業、學校師生、政府單位、學術團體通通免報名

圖 16 110 年內政黑客松競賽詳細資訊網頁

圖 16 110 年內政黑客松競賽詳細資訊網頁(續)

本次 84 個參賽團隊中，共計 22 隊使用模擬資料集，其中參賽團隊「看得見的真實需求-政策研擬平台」獲得銅獎、「安心選宅風水師」獲得榮譽獎，其決賽簡報分別摘錄如下：



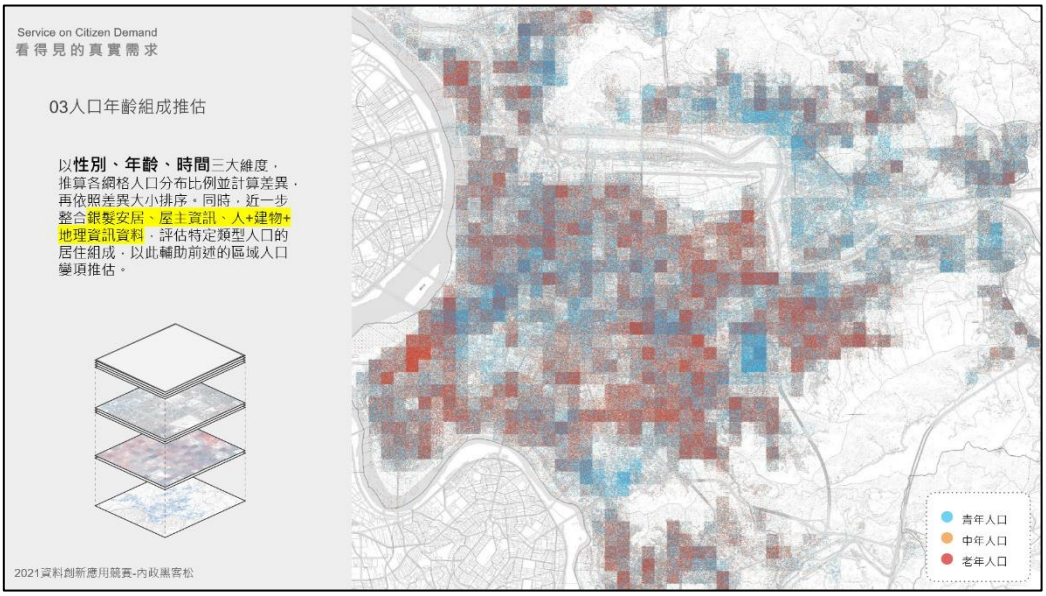


圖 17 參賽團隊「看得見的真實需求-政策研擬平台」決賽簡報摘錄

風險評估計算方式

以下每項條列式皆須符合

評分	地震	水災	坡災	火災	治安
高	• 位於斷層帶 / 高或中度土壤液化 • 房屋在民國 89 年以前建造 • 樓高三層以上 • 地下室三層以下	在 24 小時降雨 600 mm 狀態下 預估淹水達 2 公尺以上	• 位於順向坡 • 位於土石流潛勢區	• 該區平均屋齡大於30年 • 平均用電大於 60 度的戶數大於平均值 • 該區平均火災發生頻度大於全台平均值	• 該區刑案發生密度（發生數 / 實際居住人口數）大於該縣市的第三四分位數
中	• 非位於斷層帶但位於高或中度土壤液化區 • 房屋在民國 89 年以後建造 / 樓高三層以下 / 地下室三層以上 / 無資料	在 24 小時降雨 600 mm 狀態下 預估淹水達 0.3 - 2 公尺	• 位於順向坡 / 位於土石流潛勢區	• 該區平均屋齡大於30年 • 平均用電大於60度的戶數大於平均值 / 該區平均火災發生頻度大於全台平均值	• 該區刑案發生密度（發生數 / 實際居住人口數）大於該縣市的中位數
低	• 非位於斷層帶 • 位於低度土壤液化區	非位於淹水潛勢區	• 非位於順向坡 • 非位於土石流潛勢區	• 該區平均屋齡小於 30 年 • 且平均用電大於 60 度的戶數小於平均值 • 該區平均火災發生頻度小於全台平均值	• 該區刑案發生密度（發生數 / 實際居住人口數）未於該縣市的中位數

風險計算使用資料

採用資料來源包含經濟部、警政署等權威機關

地震
- 109 年人 + 建物 + 地理資訊模擬資料
- 經濟部中央地質調查所土壤液化潛勢系統
- ...等

淹水
- 2019 總統杯黑客松銀髮安居計畫 - S 環境安全需求指數
- ...等

坡災
- 110 年度 1726 條土石流影響範圍圖
- ...等

治安
- 109 年電信信令統計模擬資料
- 108 - 110 年警政署犯罪資料
- ...等

火災
- 108 - 110 年臺北市火災資料統計月報表
- 108 年建物 + 低度用電用水模擬資料
- 109 年電信信令統計模擬資料
- 109 年人 + 建物 + 地理資訊模擬資料
- ...等

圖 18 參賽團隊「安心選宅風水師」決賽簡報摘錄

伍、 結語

相較主計總處及衛福部以抽樣為基底建立之模擬資料，仍保有部分真實紀錄，可能藉結合其他資料比對出個人身分，本部創新建立之大數據模擬資料為近於母體分配之虛擬樣本，無個人資料外洩之虞，且其欄位格式、代碼與母體資料相同，能自由地進行各類交叉運算，可以最大程度開放外界應用。總結本次研究所帶來的效益如下：

一、建立模擬資料產製程序及驗證機制

由於內政大數據母體資料係串連不同資料庫而得，如「屋主資料」包括各建物屬性及其屋主人口屬性等多個欄位，因此母體筆數 10%的模擬資料要如何產製?並至少 10 個重要欄位符合母體欄位資料分配，其驗證方法為何?須定出一套程序及機制。以「屋主模擬資料」為例，係將17個欄位分類為「屋主人口模擬資料」及「建物模擬資料」，透過統計檢定找出產製階層順序，再建立兩者對應關係分派資料及進行邏輯校正，完成後再以適合度檢定、無母數相關係數及平均絕對誤差(MAE)驗證與母體資料分布相似程度，此方法可直接應用於後續產製之大數據模擬資料集，具高度應用性。

二、提升內政大數據資料處理時效

基於機敏性資料之安全管理，運用內政大數據實際資料須一星期前向本部申請，並排班至本部實體隔離環境使用資料，對本部大數據委外案研究團隊及建教合作學校(政治大學及位於南部之中信金融管理學院)來說有所不便，而模擬資料之欄位格式、代碼等與母體資料相同，相較於統計資料，更可自由地進行各項交叉運算，也可作為事前程式撰寫測試，因此可提供委外案研究團隊及建教合作學校先撰寫所需資料表格之程式，俟申請使用母體資料通過後，再至實體隔離環境將程式直接套用實際資料，可節省人力往返之路程與時間，並提升資料處理時效。