編碼：RES-107-02

行政院主計總處委託研究

農林漁牧業普查外釋資料抽樣檔建置之研究期末報告

受委託單位：國立臺北大學

計畫主持人：黃怡婷(國立臺北大學統計學系教授)

共同主持人：蘇南誠(國立臺北大學統計學系副教授)

研究助理　：張喻、林宥辰、陸薇安

行政院主計總處編印

印製日期：107 年 12 月

目次

表次

圖次

提要

對臺閩地區農林漁牧業每 5 年進行全面性普查讓政府單位能充分掌握農林漁牧業經營概況、家庭人口、勞動力狀況、作物栽培及畜禽飼養、森林作業情形、主要漁撈方式及主要養繁殖水產生物種類等資訊，歷次普查原始資料皆依「行政院主計總處提供普（抽）查資料管制作業要點」對外提供普查資料，以防止個資外洩，並保障受訪者權利，行政院主計總處為精進普查資料的使用效能，希冀參考國際作法增加建置外釋資料抽樣檔（微觀數據檔）供各界研究使用。故本計畫蒐集各國之作法，提出適合我國之微觀數據建置方式。

國外處理微觀數據建置外釋資料方式主要分成對原始資料進行去識別化與利用模擬產生資料兩種方式，本計畫採用原始資料去識別化方式，並參考國外處理變數方式，進行實證評估其適合性，以產生全檔去識別化資料檔。本計畫中採用三種去識別化方式，資料類別法、四捨五入法及產生衍生變數法，最後利用 Shlomo 等人（2010）所提的關聯評估指標來選擇合適的處理變數方式，並計算每筆觀測值的風險分數，以評估是否外釋高風險家戶資料。

針對已產生之全檔去識別化資料檔，再參考美國、英國、加拿大、日本及澳洲等國家建置抽樣檔的方法，依安全性與使用方式，決定資料外釋

抽樣檔比率，建置普查外釋資料抽樣檔，並經適合度檢定與關聯分析，以確保與普查母體資料一致性。

本次計畫實證結果顯示，在農牧戶、林業及獨資漁戶之全檔普查原始資料經去識別化處理後，唯一資料比率顯著減少，有效降低個別資料外釋風險，並同時保留關鍵變數之關聯性，可確保普查資料使用價值。再依不同業別進行研究分析，找出農牧戶、林戶及獨資漁戶最適之抽樣比率，前2者為1%之抽出率，後者為5%之抽出率，所抽樣產生之微觀資料檔經加權計算可推計母體主要變數總數，亦保留與全檔去識別化資料關鍵變數一致性，以供各界研究使用，並提升普查資料之應用價值。

第一章　緒論

第一節 研究動機與目的

政府辦理普查蒐集國家重要基礎統計資料，作為政策規劃及各界研究重要資料來源。惟普查資料包含個別資料，其釋出須在兼顧資料安全性及應用價值下進行，以發揮最大之效益，故許多國家普查資料，係將處理過的普查資料抽樣建置微觀資料（Microdata）或運用有興趣變數所建構的關聯來建置微觀模擬資料（Microsimulated Data）對外提供，除可兼顧產官學等相關單位進行政策規劃與學術研究需求，亦能確保普查個別資料不被識別。在微觀資料應用方面，如 Sundberg（2007）與 Klevmarken 及 Lindgren（2008）運用長期動態微觀模擬資料探討瑞典人口老化問題；運用美國人口普查微觀資料方面，有 Butrica 及 Iams（2000）探討退休離婚婦女的經濟福利，Favreault 及 Sammartino（2002）與 Sabelhaus 及 Topoleski（2007）討論社會改革對低收入戶與老年婦女的影響；Chen 等人（2012）利用微觀資料討論人口老化對英國房價的影響；Nadeau 等人（2013）以微觀數據建置加拿大體能活動之動態微觀資料模型。爰此，微觀資料於學術研究成果豐碩與多元，故本計畫將探討處理普查資料去識別化與抽樣方式，建置適合我國普查之外釋資料的流程，供為各界加值應

用。

本計畫參考世界各國建置微觀資料方式，整理各國建置微觀資料中所使用處理普查變數去識別化與抽樣方法，提出適合國內普查微觀資料之處理流程，且建置普查外釋資料抽樣檔，加值普查資料應用範疇與效益。

第二節 研究背景

行政院主計總處每隔 5 年舉辦一次農林漁牧業普查，以了解農林漁牧業經營現況，俾供農政機關研訂適切農業施政方針及各界參考運用。農林漁牧業普查對象包含三大業別：(1) 農牧業；(2) 林業；(3) 漁業，其中，農牧業又分農牧戶、農牧場、農事及畜牧服務業，漁業分獨資漁戶與非獨資漁戶。每次普查前，政府相關部門須辦理試驗調查及人員教育訓練，接著進行資料蒐集與建檔，普查資料係由普查員面訪蒐集而得，經指導員初審與審核員複審後，再利用光學字元辨識系統登錄資料，最後依普查表式產出統計結果表。

陳惠欣與周怡伶(2014)於農業調查研究特刊簡介我國農林漁牧業普查之推展與應用，包含將農業普查資料使用 GIS 呈現農業普查各業別資料、建置視覺化查詢系統，整合我國三大普查資料，希望呈現多元化普查數據，提升普查資料應用，提供總統府、行政院農業委員會、交通部、監

察院及相關學術單位應用。

雖然相關單位有建置資料查詢系統，增加資料的應用範圍，但系統所提供資料僅限簡單的統計結果及關聯分析。蒐集普查資料須花費大量人物力成本，若能提供產業或學界更加便捷及彈性使用資料，相信可以開發或找到更多有價值的資訊。

第二章　文獻回顧

第一節 人口普查微觀資料建置

一、　美國

美國每 10 年進行一次人口普查，2000 年人口普查問卷分長問卷與短問卷 2 種，短問卷僅包含 10 個題目，而長問卷不僅包含短問卷表格的題目，還包含家戶內所有成員的基本資料與住家資訊，而普查母體有 1/6 的家戶會收到長問卷（2010 年普查長問卷，已由社區調查問卷取代，尚未對外釋出微觀數據檔）。

美國普查局先使用郵寄方式寄送問卷，住戶的問卷回表率約為 74%，其餘未填答問卷的家戶，會由人口普查員面訪取得問卷資料。

長問卷的抽樣方式是根據人口普查區域的家戶單位數來決定，若區域中的家戶單位數未滿 800 戶，則長問卷的抽樣率為 1:2；家戶單位數為 800 戶至未滿 1,200 戶，則長問卷的抽樣率為 1:4；家戶單位數若為 1,200 戶至未滿 2,000 戶，則長問卷的抽樣率為 1:6；家戶單位數若大於或等於 2,000 戶，則長問卷的抽樣率為 1:8（見表 1）。

表 1 根據家戶單位數長問卷的抽樣率

家戶單位數	抽樣率
$x < 800$	1-in-2
$800 \leqq x < 1{,}200$	1-in-4
$1{,}200 \leqq x < 2{,}000$	1-in-6
$x \geq 2{,}000$	1-in-8

美國普查局為提供相關單位運用普查資料會提供 1 %與 5 %長問卷的微觀資料檔，但基於個人隱私保護資料，釋出檔案的變數資料會先經過處理，而資料處理方法包含資料置換（Data Swapping）、頂級編碼（Top Coding）、地理人口數門檻（Geographic Population Thresholds）、年齡擾動（Age Perturbation）及對一些類別型變數進行縮減內容。詳細變數資料處理執行方式如下：

(一) 資料置換（Data Swapping）是一種限制部分資料揭露的方法，用來保護次數性資料的隱密性，也就是有某種特性的人口數或人口的比率。資料置換的進行方式是編輯或進行資料樣本交換。置換原理原則會選取鄰近區域有相似屬性的個體，將個體資料進行互換，置換應用在個體資料，進而保護個體的數據。

(二) 頂級編碼（Top Coding）是一種限制部分資料揭露的方法，若蒐集到某變數的數值落在某一門檻值或是超過門檻值的所有個案，將該變數

數值分到特定一類。

(三)地理人口數門檻（Geographic Population Thresholds）是一種限制部分資料揭露的方法，當地理單位的人口數低於某一水準時，不能揭露個體或是家戶單位的資料。

(四)年齡擾動（Age Perturbation）是一種限制部分資料揭露的方法，用來調整家戶成員的年齡。為了保護年齡資料的隱密性，若家戶人口數較多時，需要使用擾動來修正家戶成員的年齡。

(五)類別型變數處理方式是一種限制部分資料揭露的方法，當任一類別發生的次數沒有達到某一特定最低門檻，此類別型變數的內容會被合併於其他類別。

資料置換與地理人口數門檻是針對觀察值，設定人口數門檻的安全性最高，但降低資料完整性；而資料置換相對安全性則差一些，惟提供的資料較完整，但資料置換的先決條件是觀察值需找到不同區域可配對的樣本。

而其他三類則是針對變數，頂級編碼會針對超過門檻值連續變項資料進行處理，雖有部分資料不揭露，惟可保有資料的完整性；年齡擾動的方式，僅針對家戶人口數多的成員，資料影響層面較小；類別型變數的處理，因原始資料已是分類型式，再將資料類別合併，資訊的損失會較多。

PUMS 抽樣檔有兩種設計，一種為從全國抽取 1％資料的抽樣檔，該

資料提供全國整體較全面與詳細特徵，包含最大量的社會、經濟和住宅資訊；沒有類別型變數的最低門檻限制，但有普查區域最低地理人口數門檻值 400,000 人。另一種為從州等級抽取 5％資料的抽樣檔，該資料提供以州為單位之相關資訊，惟詳細特徵資料較少；5％資料的抽樣檔中的每個州級與最小地理單位（Public-Use Microdata Area）須滿足最低地理人口數門檻值為 100,000 人，且類別型變數之每個類別人口數最低門檻設定為 10,000 人。為了保有隱私，美國這兩種微觀資料都有設定人口數門檻，如種族或是西班牙裔的人數少於 8,000 人的資料不提供外釋。

長問卷資料採分層隨機抽樣來建置抽樣檔，分層依據住戶居住房屋的屬性，房屋屬性分成三類，有住戶房屋、沒有住戶房屋與集體住戶。第一類有住戶房屋的住戶，採用 6 個分層變數，包含種族（71 類）、西班牙裔（5 類）、家庭型態（3 類）、家戶最年長成員的年齡（4 類）、住宅所有權（2 類）與抽樣比率（4 類），總共分為 34,080 層，第二類為沒有住戶房屋，採用 2 個分層變數，包含抽樣比率（4 類）與空屋狀態（3 類），總共分為 12 層。第三類為集體住戶分層，採用 4 個分層變數，包含種族（71 類），西班牙裔（5 類），居住地點屬性（2 類）與住戶年齡（4 類），總共分為 2,840 層，詳細分類方式見表 2。

表 2 依房屋屬性設定抽樣檔分層方式

房屋屬性	分層變數	類別數	類別
有住戶	種族	71	
	西班牙裔	5	非西班牙裔、墨西哥裔、 波多黎各裔、古巴裔、 其他西班牙裔
	家庭型態	3	家庭中有小於 18 歲的小孩、 家庭中沒有小於 18 歲的小孩、 非家庭（非親屬，例如：房客 或寄宿者、室友、未婚同居等)
	家戶最年長成員年齡	4	0-59 歲、60-74 歲、 75-89 歲、90 歲以上
	住宅所有權	2	自有住宅 承租
	抽樣比率	4	1:2 1:4 1:6 1:8
沒有住戶	空屋狀態	3	待售 待租 其他
	抽樣比率	4	1:2 1:4 1:6 1:8
集體住戶	種族	71	
	西班牙裔	5	非西班牙裔、墨西哥裔、 波多黎各裔、古巴裔、 其他西班牙裔
	年齡	4	0-59 歲、60-74 歲、 75-89 歲、90 歲以上
	居住地點屬性	2	住在機構或是軍隊、 住在非機構或是非軍隊

第一類與第二類採用住址檔為起始抽樣單位，方可了解有住人的住家單位比率，而第三類以人為抽樣單位，因此，抽取第一類與第二類的資料時，會先計算每州 1 %的家戶數，再依此隨機抽取長問卷第一類與第二類家戶資料，而抽取第三類資料，則會先計算每州 1 %的人數，再依此隨機抽取長問卷家戶資料第三類資料。5 %抽樣檔的操作方式與 1 %相似，僅需將流程中 1 %改成 5 %（見圖 1），而 1 %與 5 %抽樣檔提供不同資訊，因此不會重疊。

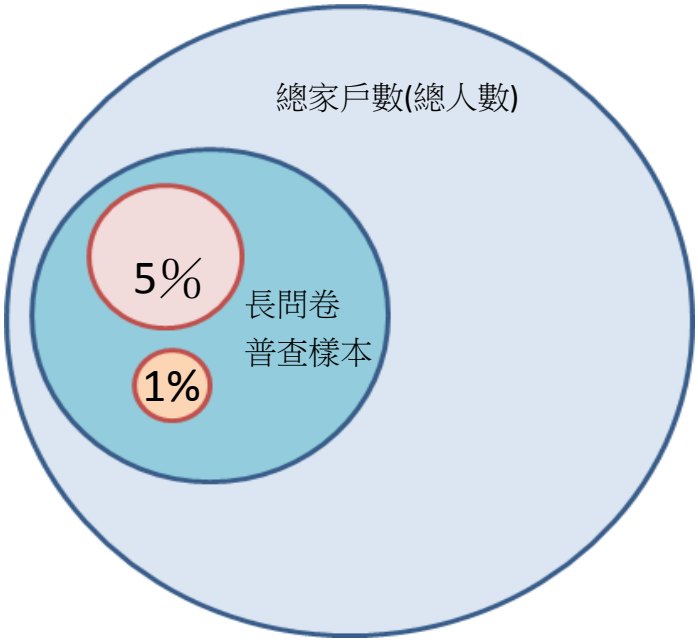

圖 1 長問卷抽樣檔取樣模式

長問卷原先抽樣分三類，每一類有不同的分層考量，分層的抽樣設定非常細，1％或 5％的微觀資料檔無法納入所有分層設定，再者，微觀資料檔分為人與家戶兩種釋出單位，並設定對應抽樣的樣本之權重，讓使用者得到可以回推母體的估計值。

由於第二類房屋屬性沒有住戶，若以人為抽樣單位，則抽樣資料僅針對房屋屬性為第一類與第三類的住戶；而若以家戶為抽樣單位，則不會納入第三類房屋屬性的資料。依照這兩種抽樣單位的權重計算方式如下：

(一)第一種以人為抽樣單位

第一類與第三類房屋屬性分別有 6 個與 4 個分層變數，其中種族與西班牙裔兩個變數設定一樣，第一類有家庭型態與家戶最年長成員年齡，而第三類則為年齡與居住地點屬性，最後，第一類另外包含住宅所有權與抽樣比率兩個變數，以下依照分層變數，將權重計算分成四個階段：

1. 初始階段

長問卷會依照普查區域的家戶數設定抽樣比例，每一個家戶的初始抽樣比例不同，因此每戶所對應的初始權重設定為該單位的抽樣率的倒數，例如：A 地區家戶數< 800 戶，每一抽樣單位初始權重為 2，即為該地區抽取率的倒數。

2. 第一階段

以人為單位的抽樣需考量家戶的總人口數，第一階段的權重計算針對家戶的人口數與家庭型態。家戶總人口數分為 2、3、4、5、6-7 與 8 人以上等 6 類，第一類的家庭型態分成 3 類，家庭中有小於 18 歲的小孩，家庭中沒有小於 18 歲的小孩與非家庭，兩個分層變數總共分成 18 類，另外，將第一類家戶僅住 1 人的額外分成 1 類，再加上第三類居住地點 2 類，住在集體住戶，住在庇護所的住戶，如沒有固定居所或是流浪漢，總共分成 21 類。

3. 第二階段

第一類的家戶有設定抽樣比率，分別為 1:2、1:4、1:6 與 1:8，但計算權重會將最後兩類合併成其它。

4. 第三階段

第一類的家戶有住宅所有權變數，第三階段依照這個變數分成自有住宅或承租兩類。

5. 第四階段

原本長問卷包含種族（71 類） 與西班牙裔（5 類）兩個種族相關的分層變數，共 355 層，但進行微觀資料抽樣時，將種族合併成 6 類，包含白人、亞洲人、美國印地安人或阿拉斯加原住民、黑

人或非裔美國人、夏威夷原住民或太平洋群島住民與其他種族，而西班牙裔僅分成西班牙裔與非西班牙裔 2 類，總共 12 類。

再者，以人為單位的微觀抽樣資料檔，抽樣額外加入家戶成員的年齡與性別等兩個基本變項，其中家戶成員的年齡分成 13 類，包含 0-4 歲、5-14 歲、15-17 歲、18-19 歲、20-24 歲、25-29 歲、30-34 歲、35-44 歲、45-49 歲、50-54 歲、55-64 歲、65-74 歲及 75 歲以上。因此，第四階段總共分成 312 層（6x2x13x2）。

初始權重等於觀察抽樣比率的倒數，依四個階段的分類將權重迭代。最後，由於抽樣單位為人，以上所計算的權重是實數，為了符合母體單位，權重需要設定成整數，例如某一抽樣單位所計算的最終權重為 7.25，則設定該權重為整數的方式是，隨機從這個抽樣單位選出 1/4 的樣本，設定權重為 8，而其餘的樣本則設定權重為 7。

(二)第二種以家戶為單位

由於第一類與第二類房屋屬性分層抽樣的設定差異很大，權重計算會分成第一類與第二類，有住戶的家戶與沒有住戶的家戶。第一類房屋屬性的微觀資料用 5 個分層變數，沒有考慮家戶最年長的成員年齡。

1. 有住戶的房屋

(1) 第一階段

第一階段的權重計算針對家戶的人口數與家庭型態。家戶總

人口數分為 2、3、4、5、6-7 與 8 人以上等 6 類，第一類的家庭型態分成 3 類，家庭中有小於 18 歲的小孩，家庭中沒有小於 18 歲的小孩與非家庭，兩個分層變數總共分成 18 類，另外，將第一類家戶僅住 1 人的額外分成一類，將有住戶的家戶分成 19 類。

(2) 第二階段

第一類的家戶有設定抽樣比率，分別為 1:2、1:4、1:6 與 1:8，但計算權重會將最後兩類合併成其它。

(3) 第三階段

進行微觀資料抽樣時，將種族合併成 6 類，包含白人、亞洲人、美國印地安人或阿拉斯加原住民、黑人或非裔美國人、夏威夷原住民或太平洋群島住民與其他種族，而西班牙裔僅分成西班牙裔與非西班牙裔兩類，總共 12 類。再加上住宅所有權包含自有住宅或承租 2 類，總共分成 24 類。

此計算過程與以人為單位的權重計算方法相同。

2. 沒有住戶的房屋

第二類沒有住戶的家戶抽樣分層變數為空屋狀態，分為待售或待租的家戶與其他 3 類，進行微觀資料抽樣時，沒有改變這個分層變數。

無住戶的家戶就無權重迭代的問題。

二、 英國

英國從 1801 年每隔 10 年進行一次人口普查，英國的普查由三個單位一起完成，英國國家統計局（Office of National Statistics；簡稱 ONS）、蘇格蘭國家紀錄局（National Record of Scotland；簡稱 NRS）、北愛爾蘭統計與研究局（Northern Ireland Statistics and Research Agency；簡稱 NISR），問卷內容會因單位不同，而有些微差異，但三個單位有設定共同題目，如家戶的房間數、住戶調整狀況（Accommodation Adaptation），志願服務工作（Voluntary Work）、語言等，三個地區普查相關資料見表 3。

表 3 英國三個地區普查相關資料

	England and Wales	Scotland	North Ireland
人口	25,000,000 家戶	40,000 家戶	15,000 家戶
執行方式	郵寄或上網	郵寄或上網	郵寄或上網
執行單位	ONS	NRS	NISR
問卷題目	14 題家戶 42 題個人資料	13 題家戶資料 35 題個人資料	14 題家戶 45 題個人資料
地理區	Local Authority District（Metropolitan District and London Boroughs）	Council Area	Local Government District
行政區	Wards/Civil Parishes/Communities	Civil Parishes	Communities

進行完普查，普查單位會產出微觀資料檔提供申請，由於微觀資料是匿名的資料，2011 年資料又稱其為匿名資料（Samples of Anonymised Records；簡稱 SARs）。為了保護個人資料，普查局採用統計揭露控制方法（Statistical Disclosure Control）建議來預防個資資料外洩，其執行主要有五種方式，使用資料置換（Data Swapping）、插補資料（Over-imputation）、細格干擾（Invariant ABS Cell Perturbation）等，以下簡述方法：

(一) 資料置換（Data Swapping）

資料置換是一種控制揭露部分資料的處理方法，假設有幾個重要的特徵變數，運用這些特徵變數，找出不同區域但成員或家戶在這些特徵變數數值相似，將資料進行置換。

進行資料置換前，須先找出有高揭露風險成員或家戶，因此會先計算成員與住戶資料的揭露風險分數，依照風險分數選取高揭露風險樣本，進行資料置換，置換的方式會選擇鄰近區域有相似特性的家戶，例如家戶成員數，鄰近區域定義有相同郵遞區號（Delivery Group），通常會包含數個普查單位（One or More Whole Local or Unitary Authority Districts）。

(二)插補資料（Over-imputation）

插補資料是一種控制揭露部分資料的處理方法，當部分普查資料出現極端值，會有個資揭露的風險，該部分的資料會先刪除，再進行插補

資料，但資料插補會影響變數間的關聯性，可能會造成最後變數關聯不一致的情況。

(三)細格干擾（Invariant ABS Cell Perturbation）

細格干擾是一種控制揭露部分資料的處理方法，該方法是澳洲國家統計局（Australian Bureau of Statistics；簡稱 ABS）採用 Fraser 及 Wooton（2006）所提的干擾演算法，引入隨機誤差項，使得每一個表格中的每一個細格數值有小幅度變動，達到保護資料的效果。但這個方法並沒有使用在 2011 年的外釋資料，主要原因為外釋資料所產出的資料關聯與原始母體的關聯不一致。

(四)設定區域門檻值

設定區域人口與戶數的最小門檻值，小於門檻值的區域僅提供有限制的摘要統計，或與其他區域資料進行合併，資料才釋出。區域整併方式會合併鄰近區域或是屬性相近區域。區域門檻值細節如下：

1. 標準表格：1,000 人與 400 戶以上。
2. 普查區域與主要統計資料：100 人與 40 戶以上。
3. 社區資料：50 人與 20 戶以上。

(五)小細格調整

小細格調整是一種控制揭露部分資料的處理方法，雖然外釋資料已經有進行資料置換，但有部分變數仍可能有細格人數過少的情況，個人

資訊仍可能有被辨識出的機會，因此，針對細格數小於門檻值，普查單位會再進一步做調整，調整方式：

1. 細格總和與子表格總和是利用調整後的表格計算。

2. 表格是獨立修正，相同的母體，不同表格有可能數據會不一樣。

3. 區域總表與個別區域的資料不一定會一致。

4. 經調整後，同一個區域的資料會一致。

資料置換與設定區域門檻與美國的處理方式一樣，而其他三種都是處理變數的方式，插補資料與頂級編碼類似，美國的處理方式是將超過門檻值的觀察值歸成一類，而英國則利用插補方式取代超過門檻值的觀察值，二者都有不錯的安全性，若無非常好的插補方式，則有可能會造成變數關聯不一致的情況。細格干擾與小細格調整會對細格數較少的資料進行擾動，此部分與插補法有相同缺點，最終有可能造成調整後的資料關聯，與原始資料的關聯不一致。

2001 年的英國的微觀資料檔採用所有處理方法，但由於不同地區對於小細格的調整方式有不同的看法，在釋出 2007 年的微觀資料前，英國國家統計局利用 2001 年的資料依照揭露風險、資料實用性與使用彈性比較各種資料處理方法，研究發現資料置換與資料插補對於資料關聯的改變較小，因此 2007 年以後的外釋資料僅採用資料置換與資料插補，但英國國家統計局最建議使用資料置換的方式，因爲資料在置換下變動最少，

而資料插補會扭曲原本變數間的關聯，而造成最終加權抽樣會與母體數不一致。

雖資料置換仍無法保護到自治區（Communal Establishment）或工作場所（Workplace），但 2011 年普查單位建議利用簽署合約方式或是設計其他置換方式彌補這部分的缺點。每一筆資料的揭露風險，英國使用以下兩種指標：

(一)經插補與置換後，真實屬性（True Attribute）揭露資料比率（$m1$）

(二)經置換後，表面屬性（Apparent Attribute）揭露資料比率（$m2$）

定義存疑率（Doubt）為 $1-(1-m1)\ (1-m2)$，該值會介於 0 到 1 之間。若資料均沒有進行置換與插補，則 $m1=\ \ m2=0$，存疑機率= 0，但若進行較多的置換或插補，則存疑率會提升。英國國家統計局會設定存疑率門檻，使得外釋資料足夠安全。

英國的普查由三個單位執行完成，微觀資料也是由三個單位各自釋出，因蘇格蘭與北愛爾蘭都是參考威爾斯與英格蘭，以下僅討論威爾斯與英國國家統計局所提供微觀資料的外釋情況。為了增加普查資料的使用度，英國國家統計局外釋微觀資料有三種檔案，第一種檔案為開放教育資料（Open Government Licence Teaching File），第二種檔案為需要簽保密協定的檔案（Safeguarded Files），第三種檔案為安全資料檔（Secure File）。以下詳細說明：

第一種檔案不需要申請，可以直接由網站下載（https://www.ons.gov.uk/census/2011census/2011censusdata/censusmicrodata/microdatateachingfile），資料檔案僅包含 1％的普查資料，且僅有主要關鍵變數（Key Variable），這些檔案可以從各行政區下載，總共有 569,741 筆資料，包含永久居民（98.5％）與短期住戶，每一筆資料包含 18 個變數，包含區域、住家的型態、住戶型態（Residence Type）、家庭成員（Family Composition）、基本成員（Population Base）、性別、年齡（8 類）、婚姻狀態（5 類）、學生、國籍、種族、信仰、經濟情況（9 類）、職業（9 類）、產業（12 類）、每週工作時數（4 類）與社經地位（4 類）。

第二種檔案為需要簽保密協定的檔案（https://www.ons.gov.uk/census/2011census/2011censusdata/censusmicrodata/safeguardedmicrodata），資料檔案包含 5％ 的普查資料，需簽保密協定，第二種檔案提供兩類個人層級的資料，第一類為 Region 層級的個人資料，第二類為 Grouped Local Authority Level 層級的個人資料，層級以 120,000 人進行分隔，但若層級區域的人口數低於門檻值，則會與鄰近的區域合併。第一類資料與第二類資料都包含 120 個變數，所有變數都是類別型式呈現，但有不同外釋變數處理方式，與第一類資料相比，第二類資料的地理層級較小，檔案提供的資訊較少，第二類資料部分變數會提供較少的分類數，例如年齡、出生地、種族、職業分類、家戶成員數、房屋

型態等。

最後，第三種檔案是安全資料檔，該檔案 2016 年才正式外釋，這份資料包含 10 %的普查資料，威爾斯與英格蘭會提供地理層級超過 5,000,000 人區域的資料，資料內含 247 個個人層級的變數（https://www.ons.gov.uk/census/2011census/2011censusdata/censusmicrodata/securemicrodata），蘇格蘭的資料會提供 180,000 人 220 個個人層級的變數，最後，北愛爾蘭的資料會提供 534,000 人 146 個個人層級的變數。最小的地理區域為 Local Authority Level，且自治區的資料也會納入。威爾斯與英格蘭的資料會提供超過 2,400,000 個家戶 245 個家戶層級的變數，蘇格蘭的資料會提供 74,000 個家戶 200 個家戶層級的變數，最後，北愛爾蘭的資料會提供 247,000 個家戶 220 個家戶層級的變數。該份資料放置在虛擬個體數據資料庫服務（Virtual Microdata Laboratory；簡稱 VML），經同意許可後，使用者可經由雲端帳戶登入進行資料分析，資料不可以攜出。

三、 加拿大

2011 年國家家戶調查（2011 National Household Survey；簡稱 NHS）採用抽樣方式進行，隨機抽取少於 30 %的家戶，約 4,500,000 家戶數。

NHS 分兩階段，第一階段先選擇 30 %家戶進行調查，經過一段時

間，第二階段再從沒有回覆的樣本中選取 1/3 的樣本進行調查。因爲 NHS 屬於自願填答的問卷，沒有回覆問卷的家戶會比一般強制填答的調查高，通常第一階段有回覆的家戶比率約為 68.6 %，再使用有回覆的資料進行加權，有回覆的家戶最後的權重會因抽樣設計與回覆率而有不同，其數值會介於 1 到 100 之間。給定一個家戶權重為這戶代表的戶數，而給定有回覆家戶的個人權重，則為這個人代表的人口數。

微觀資料檔（Public Use Microdata File；簡稱 PUMF）的樣本採用兩階段（Two Phases）方式從有回覆的受訪者抽樣資料，第一階段設定微觀資料檔的抽樣名冊，第二階段根據抽樣名冊進行抽樣。

第一階段會先將填答者的抽樣名冊（Sampling Frame）分成三部分，第一部分的名冊提供個人資料（Individual File Records），第二部分用來選擇分級檔名冊（Hierarchical File Records），亦即包含家戶與其成員，第三部分用來選擇名冊使得抽樣的資料可以進行國際比較。進行第一階段時，會先將有填答家戶依照住戶居住所屬省與領地（Province or Territory of Residence）及家戶常住人口數（Number of Usual Residents in the Household），按普查區（Census Division）、小普查區（Census Tract） 與傳播區域（Dissemination Area）排序，NHS 的資料再系統性分成三種不同的抽樣名冊。

第二階段為抽樣方式，依照第一階段得到的名冊，再設定抽樣比率產

生 PUMF 。加拿大提供兩個 PUMF 檔案，分別由第一部分與第二部分的名冊抽樣產生，而抽樣率分別為 2.7 %與 1 %。

兩個 PUMF 檔案都是採用 PPS （Probability-Proportional-to Size）系統抽樣選取樣本，其中樣本大小（size）是依照從第一階段所得到的權重來計算。為了讓 PUMF 樣本資料可以回推母體，選取的樣本會需要重新計算所對應的權重，其計算方式為權重除以第一階段的樣本分配率（Sampling Fraction），使得抽樣資料可以代表母體，第一部分與第二部分的樣本分配率分別為 1/（2.7 %） 與 1/（1 %）。

運用第一部分的名冊建立第一個 PUMF 微觀資料檔，統計局希望 PUMF 微觀資料檔是一個自我加權的樣本（Self-Weighted Sample），即抽樣檔是 NHS 第一部分母體的 2.7 %，但部分樣本第一階段所得到的權重可能大於 1/（2.7 %）。若依照系統抽樣，每一個家戶抽中的機會相同，就不可能產生自我加權的樣本。最佳的抽樣方式是先選出權重超過 32.4 的樣本，然後針對剩下的樣本，依 PPS 系統抽取。也就是，自我加權的樣本會從權重低於 32.4 的回覆資料中選取。系統抽樣會先將資料依照以下 5 個變數排序：

(一)住戶居住所屬省與領地（Province or Territory of Residence）

(二)城鄉指標（Urban－Rural Indicator）

(三)成員性別

(四)年齡，分成 0-15 歲，16-35 歲，36-65 歲，66 歲及以上。

(五)種族，分成 7 類。

樣本會從（0,32.4）取樣區間系統性選出，每一筆資料會抽中的機率與其對應由第一階段計算出權重有關。選取流程如下：先由區間（0,32.4）隨機選出一個數字，假設為（w_0），選取第一個隨機取樣樣本，假設其對應的權重為（w_1），若 $w_0+w_1 > 32.4$，則保留該樣本，反之，則刪除，然後往下選取樣本。若決定保留樣本，則樣本區間需先減去目前的累積總和，才可以進行選取下一個樣本。最後，第二階段的權重需要針對每一個家戶重新計算，權重計算方式為第二階段取樣機率的倒數。第一個 PUMF 檔案包含 887,012 筆資料。

運用第二部分的名冊建立第二個 PUMF 微觀資料檔，統計局希望 PUMF 微觀資料檔是一個自我加權的樣本（Self-Weighted Sample），即抽樣檔是 NHS 第二部分母體的 1%，但，若依照系統抽樣，每一各家戶抽中的機會相同，就不可能產生自我加權的樣本。最佳的抽樣方式是先選出權重較大的樣本，例如權重超過 90.3，然後運用剩下的樣本，依 PPS 系統抽取。也就是，自我加權的樣本會從權重低於 90.3 的回覆資料中選取。系統抽樣會先將資料依照以下 8 個變數排序：

(一)住戶居住所屬省與領地（Province or Territory of Residence）

(二)城鄉指標（Urban – Rural Indicator）

(三)家戶常住人口數（Number of Persons in the Household）

(四)普查家戶結構與住家型態（Census Families Structure and Type in the Household）

(五)有年長者的家戶（Elderly Person Presence Indicator in the Household）

(六)有少數民族的家戶（Visible Minority Presence Indicator in the Household）

(七)家戶人口種族相似指標 （Similar Ethnic Origins Indicator in the Household）

(八)家戶至少有一個人有工作（At Least One Person in the Labour Force Indicator in the Household）

樣本會從（0,90.3）取樣區間系統性選出，每一筆資料會抽中的機率與其對應由第一階段計算出權重有關。選取流程如下：先由區間（0,90.3）隨機選出一個數字，假設為（w_0），選取第一個隨機取樣樣本，假設其對應的權重為（w_1），若 $w_0+w_1 > 90.3$，則保留該樣本，反之，則刪除，然後往下選取樣本。若決定保留樣本，則樣本區間需先減去目前的累積總和，才可以進行選取下一個樣本。最後，第二階段的權重需要針對每一個家戶重新計算，權重計算方式為第二階段取樣機率的倒數。第二個 PUMF 檔案包含 132,192 筆家戶資料與 333,008 個成員，基於個資保護，有部分家戶成員的資料沒有提供。

為了保護個資，PUMF 資料調整方式如下：

(一)住家（Housing）

PUMF 整併出租住宅（Rented）與公共住宅（Band Housing）類別。並避免揭露個資，受訪者居住在公共住宅僅提供插補後的總房屋租金。

(二)區域

最小的地理單位為人口普查大都會地區（Census Metropolitan Area；簡稱 CMA），不會提供範圍小於 CMA 的資料，且資料僅包含 5 個最大的人口普查大都會地區與省（Province），其中育空（Yukon）、西北地方（Northwest Territories）與努納武特（Nunavut）整併成北加拿大區（Northern Canada）。

(三)類別型變數

類別型變數會採用選項整併的方式，整併方式會同時考量保護個資與提供有意義性的關聯分析，但針對少數民族的資料，有些類別資料會有一個「資料不提供」（Data Not Available）的類別。以下提供部分類別型變數分類方式：

婚姻狀態分成 6 組，包含未婚、已婚、普通法的伴侶關係、分居、離婚、喪偶。

出生地分成 6 組，包含加拿大、美國、歐洲、亞洲、其他、不提

供資料。

移民年份分成 6 組，包含 1981 年以前、1981-1990、1991-2000、2001-2005、2006-2011、不提供資料。

(四)連續型變數

連續型變數有幾種處理方式，第一種採用等距方式分組，第二種方式採四捨五入方式呈現，以 50、100、1,000 或 10,000 等為基底，第三種方式設定連續型變數的極端門檻值，超過門檻值以超過門檻值族群的平均值或中位數取代，或設定上（下）界。再者，為了降低填答者的負擔，2011 年 NHS 問卷增加一個題項「詢問受訪者是否同意使用稅收資料取代填答收入的資訊」，若回答不願意者，需填寫紙本資料或上網填答。2011 年 NHS 普查有 7 成的受訪者同意。以下提供部分連續型變數分類方式：

年齡分成 14 組，包含 0-9 歲、10-14 歲、15-19 歲、20-24 歲、25-29 歲、30-34 歲、35-39 歲、40-44 歲、45-49 歲、50-54 歲，55-59 歲、60-64 歲、65-74 歲、75 歲（含）及以上。

通勤距離分成 7 組，包含少於 5 公里、5-9.9 公里、10-14.9 公里、15-19.9 公里、20-24.9 公里、25-29.9 公里、大於 30 公里。

通勤時間分成 6 組，包含少於 15 分鐘、15-29 分鐘、30-44 分鐘、45-59 分鐘、60 分鐘（含）以上、不提供資料。

利用四捨五入方式來呈現所有與收入相關的資料，如：政府轉帳薪資

總額（Total Government Transfer Payments；變數名稱為 GTRFS）變數四捨五入以 100 為基底，總收入（Total Income，變數名稱為 TOTINC）、證券收入（Market Income；變數名稱為 MRKINC）、薪資（Employment Income；變數名稱為 EMPIN）、所得稅（Income Tax Paid；變數名稱為 INCTAX）、稅後收入（After-Tax Income；變數名稱為 TOTINC_AT）、可支配所得（Disposable Income for Market Basket Measure (MBM) for All Persons；變數名稱為 EFDIMBM）變數四捨五入以 1,000 元為基底。若變數數值超過 100,000 元，則四捨五入以 10,000 元為基底，但若變數數值大於 0 元，但四捨五入後變成 0 元，則數值設定為 1 元，而若變數數值小於 0 元，但四捨五入後變成 0 元，則數值設定為 -1 元。這樣的設定原則可以確保使用四捨五入後，資料仍舊保有原始資料來源中，正、負與零的關係。但因爲是針對每一個變數進行四捨五入的處理，處理後的資料會喪失部分收入的關聯。

為避免揭露個人隱私，刪除極端的收入族群。依區域與性別，除了可支配所得，其它與收入有關的變數若超過 99 百分位點的資料設定為頂端族群，頂端族群的收入利用該群資料的加權平均收入來取代。若是收入為負值，且低於某一個門檻值，則稱為底端族群，則門檻值設定方式依照居住地，若居住在大西洋區域則設為 -30,000 元，而其他地區則設為 -50,000 元。而可支配所得採用 98 百分位點的資料設定為頂端族群。頂端族群的

所得利用該群資料的加權平均所得來取代，若是可支配所得為負值，且低於某一個門檻值，則稱為底端族群，則門檻值設定方式依照居住地，若居住在大西洋區域則設為 -30,000 元，而其他地區則設為 -50,000 元。

考量樣本變異，人口數較多的普查區域的樣本變異會很小，但人口數較小的區域則抽樣誤差會較大，因此，若普查地區住戶數小於 40 戶或是人口數少於 250 人的區域不提供收入資料。

等距方式與四捨五入方式都是針對觀察值，限制可以外釋的條件，而其他處理變數的方式，加拿大的調整方式會增加一個類別，不提供資料，來處理數值超過門檻值，另外，連續型變數多一種四捨五入方式。

經由多倫多大學人文社會科學計算網站（網址：sda.chass.utoronto.ca）可以直接產生 PUMF 資料相關變數的列聯表與圖形，或是經由下載表單設定欲下載變數，該網站提供 5 種資料儲存格式。

四、 日本

日本從 1920 年開始，每 5 年進行一次人口普查，從 1973 年開始訂定地域方格作為最小普查資訊地區，在人口密集區，方格的大小為 500 公尺，而非人口密集區則設定為 1 公里。人口普查主要調查 13 項住戶的基本資料，包含姓名、性別、出生年月、與戶長的關係、婚姻狀態、國籍、居住期間、5 年前的住所、活動，設定公司與公司的屬性，職業、受僱情

況與工作或上學地點，而住家則調查 4 個變數，包含居住型態、住家有幾人、住家所有權、建築物的型態與居住樓層等。普查的方式是訪員逐戶提供網路帳號與密碼，由住戶自行上網填答。若住戶沒有上網填答，普查局會寄紙本問卷，之後再由訪員親自拜訪。

日本於 2007 年修訂「統計法」，提高對微觀數據資料的重視，日本從 2013 年開始釋出人口普查的微觀數據資料，微觀資料會在普查進行完成後 5 年釋出，目前普查局提供 2000 年與 2005 年普查的微觀數據資料，其釋出檔案的保護處理方式多為以下幾類：

(一)抽樣（Sampling）：僅公布抽樣率為 1 %的資料。

(二)頂（低）級編碼（Top / Bottom Coding）：若蒐集到某變數的數值落在某一門檻值或是超過 (低於) 門檻值，則將該變數數值分到特定一類。

(三)地理人口數門檻（Geographic Population Thresholds）：當地理單位的人口數低於某一水準，不能揭露個體或是家戶單位的資料，日本設定的門檻值為 500,000 人，當地理單位人口低於 500,000 人時，則不會公布該地理單位的資料。

(四)資料刪除（Data Deletion）：在遇到唯一一筆（Unique）的資料，刪除該筆資料。

第二節 農業普查相關資料

美國每 5 年進行一次農業普查，只要在普查年有生產或銷售（含預期銷售）農產品價值達 1,000 美金以上之農場就是普查的範圍，普查主要調查的內容包含農地使用情況、產權、雇員、生產方式、收入與支出等。園藝普查（Cencus of Horticultural Specialties）則從 1889 年開始執行，每間隔 10 年進行一次普查，但最近幾次則間隔年數不一致，最近一次在 2014 年，只要在普查年有種植花、苗圃或特殊作物之價值達 10,000 美金以上就是普查對象範圍，普查利用郵寄方式執行，主要調查的內容包含生產園藝作物、作物價值、農地種植面積、雇員、生產方式、收入與支出等。同農業普查，水產養殖業普查（Census of Aquaculture）在普查年有賣養殖水產達 1,000 美金以上就是普查對象範圍，該普查會蒐集漁獲量、方法、養殖的面積、水源、收入與銷售方式等。這些普查資料相關單位會產生年報；若針對某種特殊產品的種植面積、收入等資料，農業部亦提供互動式網頁讓使用者搜尋（網址：https://www.nass.usda.gov/），但該部分的資料沒有釋出微觀資料。

日本每 5 年進行一次農業普查，調查分成：農林業經營調查及農山村地域調查二類；前者是以生產農林業產品或接受委託進行農林業作業之「從事農林業生產活動人員」(如果是以法人組織方式進行農林業生產

活動的話，則以法人代表為對象)；後者則以日本全國的市町村或農業集落(已列屬市區化的所有農業集落除外)為對象。這些普查資料相關單位會產生年報（網址:http://www.maff.go.jp/），但該部分的資料沒有釋出微觀資料。

加拿大每 5 年進行一次農業普查，總共的問卷題目有 36 大題，183 個題目，包含基本資料、雇員、耕地面積、可耕種面積、種植作物包含穀物、蔬菜、水果等、苗圃或耶誕樹等、耕地種植範圍、夏季休耕、種植方式等。加拿大居民只要擁有農地與牧場或從事其他與農業相關的行業都需填答此問卷，但每一位填答大約只需回答 20 %的題目， 2016 年普查採三階段蒐集資料，第一階段提供網路帳號供受訪者填答，第二階段寄送提醒函，第三階段寄送紙本問卷。加拿大提供農業普查與人口普查連結的資料，該部分的資料以省為單位，以網頁互動方式取得表格資料（網址：https://www150.statcan.gc.ca/n1/en/type/data?MM=1），但沒有釋出微觀資料。

第三節 其他抽樣資料微觀資料建置

加拿大提供許多抽樣調查之外釋微觀資料檔，但所有資料均需申請且需付費，總共釋出有 121 份調查全檔資料，主題多元包含旅遊、健康、收入、財務、一般社會活動調查、家戶與環境、畢業生、勞工保險等（網

址：www150.statcan.gc.ca）。

雖然加拿大外釋多個全檔抽樣資料，但資料外釋前都有先經過處理，利用相關變數的資料合併，定義出新變數，將其稱為衍生變數（Derived variable）。

每 5 年會進行一般社會活動調查（General Social Survey；簡稱 GSS），早期每次調查樣本約 10,000 人，1999 年之後樣本數增加到 25,000 人，調查對象為 15 歲以上的加拿大居民，主要希望瞭解家庭狀況（Family）、民眾時間規劃（Timing）、社會認同（Social Indentity）、社會服務（Volunteer）、受害情況（Victimization）等，每一次調查會著重於不同的主題。該問卷使用電話簿當成抽樣名冊，採用 CATI 系統進行電訪，平均填答時間為 40 到 45 分鐘。但因成本考量，2010 年之後改採上網填答。2015 GSS 為第 29 次調查（Cycle 29），該調查主要重點係了解社會變遷趨勢，問卷題目涵蓋居民的生活條件、生活情況等 9 個面向。以下以 2015 GSS 所外釋的微觀資料為例說明衍生變數產生方式：

(一)受訪者的基本資料：性別、年齡、小孩等變數均使用家戶合成矩陣（Household Composition Matrix；參考 Akkerman，1980）來產生，例如年齡變項，年齡的家戶合成矩陣由兩種資訊組合，第一為將母體的年齡分組，然後計算出分組後每一組的母體數，與第二為戶長的年齡組別的母體人數，利用這兩個資訊來計算新的年齡變數。受訪者年齡，

分成 10 類，受訪者年齡與配偶的年齡差距，與小孩一同居住，此為二元變數，分成有與沒有。婚姻狀態由 5 個變數轉換而成，所有成員的資料都變成與第一成員的關係。

(二)工作變數：工作狀況利用 3 個變數轉換全職與兼差，並分成 5 類，工作 1-13 週、14-26 週、27-39 週、40-48 週與 49-52 週，總共為 10 類，工作地點轉成行政區。

(三)收入：包含工作薪資、酬勞、自營農戶收入、非自營農戶收入、利息、股利，其他收入與政府轉帳的支付款項等，外釋資料會有使用不同權重計算收入。另外有提供 10 個相等機率的百分位點。房貸與房租費用占總收入的占比，定義如下：

$$房貸費用占總收入比率 = \frac{房貸月支付款項}{總收入/12} \times 100$$

$$房租費用占總收入比率 = \frac{房租月支付款項}{總收入/12} \times 100$$

除了計算衍生變數，資料外釋時，還會再將變數類別化，區分為小於 30 %、30 % - 50 %與大於 50 %。

第四節 模擬數據

歐盟與澳洲採用模擬數據方式外釋資料，產生外釋模擬數據資料前會需要先設定有興趣的變數，並找出這些變數的關聯，再運用這些關聯模

擬資料。

一、 澳洲

澳洲人口普查使用澳洲區域分類（Australian Standard Geographical Classification；簡稱 ASGC），ASGC 將澳洲分成 4 個統計區域等級（Statistical Area Level 1-4；簡稱 SA1-SA4）與大都會區（Greater Capital City Statistical Area；簡稱 GCCSA），在這些區域下，ASGC 進一步設定最小地理區域（Basic Building Block），當成蒐集資料的基本區塊，稱為網格塊（Mesh Block），以網格塊為基準區域來蒐集人口普查資料，又稱人口普查區（Census Collection District；簡稱 CD）。

以 2006 年的澳洲人口普查資料為基準資料，模擬資料以人口普查區為單位，Namazi-Rad、Mokhtarian 與 Perez（2014）運用階層結構（Hierarchical Structure）方式，考慮區域（Geographic Area）、家庭（Household）與成員（Individual），抽取出 1% 樣本，為了使重抽資料有代表性，運用二種方式（1）重新建構合成法（Synthetic Reconstruction；簡稱 SR）;（2）爬坡法（Hill Climbing；簡稱 HC），來重新建置資料。

SR 法主要使用可決定演算法（Deterministic Algorithm）來重新建置母體，而 HC 法則使用隨機資料動態法來建置。SR 法以人口普查區為取樣單位，利用澳洲官方保密資料（Confidentialised Unit Record Files；簡

稱 CURF）當成基準資料（Seed Data），抽取 1 % 樣本，利用整合資料（Aggregate Data）產生人口普查區針對個人樣本有興趣的基本資料之邊際分配，再利用加權方式（Weighting Techniques）來生成資料。

HC 法運用最佳化方法來隨機生成資料，設定一個目標函數，如給定某一個特定區域某一個變數的邊際分配，例如年齡的分配，HC 法隨機生成資料，與目標函數比較，若與目標函數差異在設定的範圍外，需要進行資料置換或重新抽一個資料，直到找到最佳解（Optimum Solution）。但這個方法有可能找到區域最佳解（Local Optimum），重新設定隨機亂數的種子（Seed）可以確認找到最佳解。運用最佳組合演算法（Combinatorial Optimization Algorithm；簡稱 CO），需先設定目標區域所欲探討的變數與其對應的邊際分配，從普查資料隨機抽取一群人，樣本大小與特定母體一樣，計算出該樣本的樣本分配，運用適合度檢定來檢驗該樣本分配與欲探討變數之邊際分配，若檢驗不通過，從非整合資料重新選一個樣本資料（Record），再進行一次檢驗，重複前述步驟，直到檢驗結果符合預期人口分布，預期人口分布可以為年齡、性別與住家種類所建置出的交叉分配。

最後，再加人口出生、死亡、結婚與移民的人口變動模型，來建置人口普查模擬資料，並使用 2011 年的普查資料來驗證該人口普查模擬資料。

二、 歐盟

參考歐盟收入與生活狀況調查（European Union Income and Living Conditions；簡稱 EU-SILC），Alfons 等人（2011）討論複雜抽樣問卷調查的合成資料或微觀數據檔資料產生方式，他們指出模擬母體資料時需要注意以下條件：

(一)需要設定區域（Regions）、分層（Strata）的實際大小。

(二)應該正確地表示變數之間的邊際分配（Marginal Distributions）和交互作用。

(三)允許分組之間的非均質性（Heterogeneities），特別是區域方面。

(四)應避免從樣本中完全複製，因為這通常會導致較小的子群內單位變異性極小。

(五)必須確保數據保密性。

模擬重點說明如下：

(一)家戶結構的設置：家戶結構是分別由分層和家戶大小的每個組合進行模擬的。

(二)類別型變數的模擬：Münnichetal 等人（2003）與 Münnich 及 Schürle（2003）直接根據樣本中類別型變數頻率分布（Frequency Distribution）所估計出的條件分布（Conditional Distribution）來模

擬，但此方法需要一個相當大的樣本量，而且不是很靈活。除此之外，此方法不允許產生樣本中不存在的組合。為了克服此缺點，Alfons 等人（2011）提出使用多項式羅吉斯迴歸模型（Multinomial Logistic Regression Models）來估計條件分布的方法。運用問卷調查所得出的樣本來建構多項式羅吉斯迴歸模型，將所得到的估計參數用來推估類別型變數的母體發生機率，再用類別發生機率來模擬出變數的數值。該模擬方法可以用來處理細格為零的問題。

(三)連續型變數的模擬：Alfons 等人（2011）提出兩種不同的方法模擬連續型變數，除了一般連續型變數，兩種方法都能夠處理半連續型變數（Semi-Continuous Variables），即包含大量零的變數。以下列出二種方法：

1. 多項式模型隨機抽取結果的類別（Multinomial Model with Random Draws from Resulting Categories）：假設有興趣的反應變數為類別型，運用前述的模擬方式生成資料後，給定生成反應變數的類別，連續型解釋變數的值是依均勻分布（Uniform Distribution）隨機生成。這種方法背後的想法是將資料分成相對較小的子集，但如果區間過長，則使用均勻分布可能會太簡化。然而，這種方法的優點在於它可以選擇離散化斷點，使得經驗分布（Empirical Distribution）與模擬母體

變數的分配相近，需要考慮的是斷點數越多，計算時間越長。通常使用 10 % 的分位點來模擬資料，但若是極端值的區段則可以考慮用 1 %、5 % 與 99 % 的分位點來分隔區間。但若要分析的變數會包含極端值，如收入等，則最極端的類別可以使用極端值分配來模擬產生。

2. 包含隨機誤差項的（兩步驟）迴歸模型：第二種方法使用一般線性迴歸模型來生成連續型資料，但半連續型變數則使用兩步驟迴歸模型來生成。運用樣本資料來配適線性迴歸模型，利用所得的估計參數來生成對應母體的資料，生成資料時，在給定相同一組參數時，需加入隨機誤差項，方可使生成資料數值相異，生成隨機誤差項的方法有二種：

   (1) 由殘差隨機抽取。

   (2) 隨機由常態分布 N（μ , σ^2）產生，其中，由樣本中位數與絕對平均離差值（Mean Absolute Deviation；簡稱 MAD）來估計 μ 和 σ^2。

第一種方式比較依賴資料，而第二種方式是按照常態分布誤差的假設，對於這兩種情況，需要仔細選擇分位點。如果太小，非常大的隨機誤差項可能會導致分布的大量偏差；如果它太大，隨機誤差項可能不會有足夠的變化性（Variability）。

若需要處理收入變數，可考慮先進行對數轉換，再使用上述步驟來產生資料。

(四)數個連續型變數組成的成分（Components）模擬：Kraft（2009）指出模擬數個連續型變數組成的成分時，即便僅有少數連續型變數其對應的變數關聯還是很複雜。再者，若有連續型變數的成分會使資料有稀疏個數（Sparseness）的問題，例如收入變數可能僅有少數類別有資料，其餘的類別觀察值個數可能為零。為了處理這些問題，基於分數的條件樣本重抽法（Resampling Fraction），Alfons 等人（2011）發展出一個簡單而有效的方法，亦即在調整控制變數時，僅考慮非常少的高影響力的類別型變數。依據元素的占比，使用分數的條件樣本重抽法選取樣本，並設定母體對應的數值。利用重複抽選方式，具有避免模擬成分中有不切實際或不合理的組合的現象。同時，它不會使得模擬是完全複製，因為模擬個體的絕對值通常與調查數據中對應的個體不太相同。

最後，模擬資料方式可使用 simPopulation 的 R 套件，而資料診斷的圖形則可使用 vignette 套件（Alfons 等人，2010）。

第五節 其它微觀資料檔建置評估相關文獻

一、 唯一資料處理

普查檔之公布往往伴隨著個資法的問題，因此在公布前需要進行去識別化的動作，而上述去識別化的方法，小區域的資料很容易會遇到唯一資料(Unique)，最簡單做法是將資料刪除(Data Deletion)，但 Ito 及 Hoshino（2014）認為刪除資料會喪失原本母體的特性，參考 Shlomo、Tudor 及 Groom（2010）之建議，應該使用資料置換將唯一資料進行處理，方可使微觀數據資料更具代表性，並討論 3 種資料置換，分別為目標資料置換（Targeted Data Swapping）、隨機資料置換（Random Data Swapping）、合併目標與隨機資料置換（Combination of Targeted and Random Data Swapping），以下詳細介紹：

(一)目標資料置換

首先辨認高風險個體或家戶，依不同地理層級計算出每個人的風險分數，若此分數高於所設定的臨界值，即為高風險個體，而只要有一成員為高風險個體，則其所在家庭定義為高風險家戶。通常相關單位須決定資料交換比率，假設交換率為 $p\%$ 的家戶數。通常置換會發生在大地理區域中的區段(Block)。例如英國的威爾斯與英格蘭普查的地理層級依序為 Delivery Group Area、Local Authority（LA）、Wards、Output Area（OA），

則交換會是在相同的 Delivery Group Area，但不同 LA 的資料進行互換。

置換樣本數等於以沒有插補資料的總家庭數乘以 p %。置換樣本會依比率分配到 OA，而比率分配可以有兩種方式：

1. OA 層級沒有插補資料的家庭數的倒數
2. OA 層級高風險家戶比率

若 OA 層級的家戶數不超過 20 %，OA 的最後樣本數等於兩種比率分配方式的平均值。依照 PPS 設計選取隨機樣本，而 size 變數的值與風險分數成正比，值越大，代表抽中機率愈高，再依照先前設定的控制變項，例如，年齡組別、性別、種族等，找對應的配對。

除了依照個體設定風險分數，Young、Martin 與 Skinner（2009）建議依照距離，與個體設定風險分數一樣，先決定一組關鍵變數，分地理區域，計算每一個關鍵變數的分布，若在地理區域 g，有一個個體風險分數高於門檻值，則該個體在地理區域 g 設定為高風險個體，則家戶地域風險等級設定有風險個體的最高地理區域的風險等級。例如在 Wards 等級，有一個個體在某類關鍵變數唯一，而該戶有另一個個體在 LA 等級其某類關鍵變數唯一，則該戶的家戶地域風險等級為 LA。地域風險等級進行置換的方式如下：

1. 檢查每一戶的地域風險等級
2. 在相同控制變項的條件下，進行配對

配對需在相同地域風險等級層級進行，例如某家戶在 LA 等級，則配對的等級也需找在 LA 等級，但在相同 Delivery Group Area。

以下以英國 2001 年人口普查資料為例，考慮三個控制變數，分別為性別、種族、婚姻狀況，性別分男性與女性，種族分白人、拉丁裔、黑人、華人與原住民，婚姻狀況分已婚、單身與離婚。若成員 1 在 Wards 中有揭露風險、成員 2 在 OA 中有揭露風險，則此家庭在 Wards 中有揭露風險，該家庭有揭露風險的地理層級為 Wards。若家庭在 Wards 中有揭露風險，則與不同 Wards 但有性別、種族與婚姻狀況(成員 1：已婚白人男性、成員 2：已婚白人女性)及在相同 LA 的另一家庭置換。若家庭在 OA 中有揭露風險，則與不同 OA 但有相同性別、種族與婚姻狀況（成員 1：已婚黑人男性、成員 2：已婚黑人女性，成員 3：單身黑人男性）及相同 Wards 的另一家庭置換。

(二)隨機資料置換

隨機資料置換與目標資料置換一樣會需要設定交換率，但置換的方式改成隨機交換，亦即，在相同控制變數下，與配對樣本進行隨機交換。

(三)合併目標與隨機資料置換

合併目標與隨機資料置換方法是上述兩種置換方式合併使用，假設需要執行置換的家戶比例為 p，則 $1/2\,p\%$ 的家戶會採用目標資料置換方法，而另外 $1/2\,p\%$ 的家戶會使用隨機資料置換方法。

二、 資料揭露與指標

分地理區域，首先須先決定一組關鍵變數，計算每一個關鍵變數的分布。假設有 M 個關鍵變數，每一個關鍵變數有 k_m 類，$m = 1,2,\cdots,M$。在地理區域 g，假設 $N_{k_m}^{g}$ 表示第 m 個關鍵變數的第 k_m 類的人數。因此細格個數會與風險分數成反比，令位於地區 g 的第 i 位受測者 m 個關鍵變數的分類組合為 $k = (k_1, \ldots, k_M)'$， Shlomo 等人（2010）定義風險分數為

$$HR_k^g = \left(\sum_{m=1}^{M} \frac{1}{N_{k_m}^g} \right) / M$$

假設外釋資料的風險門檻值，當風險分數超過門檻值則設定為高風險個體，若家戶中有 1 人為高風險，則該家戶設定為高風險。

另外，Ito 及 Hoshino（2014）建議揭露風險與資料效用 2 種指標來評估揭露風險，其定義如下：

(一)揭露風險（Disclosure Risk；簡稱 DR）

揭露風險定義為交換前樣本的所有唯一資料與交換後的所有唯一資料的比率，定義如下：

$$DR = \frac{\sum_c I(T^O(c) = 1, T^P(c) = 1)}{\sum_c I(T^O(c) = 1)}$$

其中，$T^O(c)$ 為交換前細格值，$T^P(c)$為交換後細格值。DR 的值

愈大，揭露風險愈高，因此希望 DR 愈小愈好。

(二)資料效用風險（Data Utility：簡稱 DU）

資料效用風險定義為總細格數與交換前後所有唯一資料差值比率，定義如下：

$$DU = \frac{\sum_c |T^P(c) - T^O(c)|}{n_T}$$

其中，n_T 為總細格數。DU 值愈大，代表置換前後資料分配有差異，資料效用風險愈高，因此希望 DU 愈小愈好。

運用這兩種指標，Ito 及 Hoshino（2014）建立 RU 圖（Risk-Utility Confidentiality Map），該圖可以說明唯一資料處理的三種置換方式的優劣，以英國的資料為例，圖 2 的菱形代表使用目標資料置換的資料，正方形表示使用隨機資料置換的資料，而三角形表示使用合併目標與隨機資料置換的資料，圖形顯示目標資料置換的揭露風險比較小，隨機置換的資料效用風險小，而合併目標與隨機資料置換則介於兩者之間，因此合併目標與隨機資料置換較為合適。

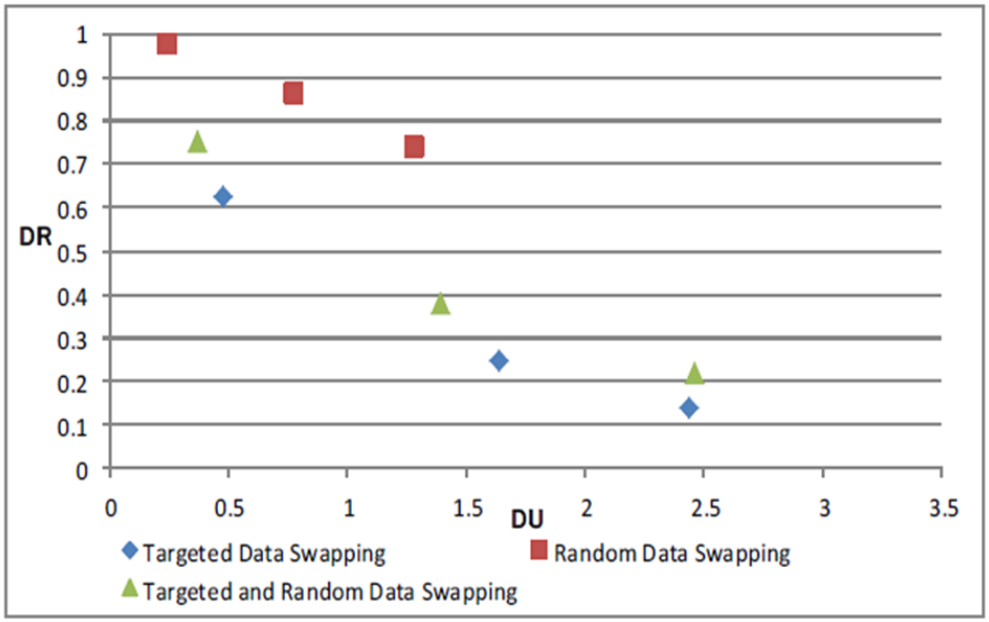

圖 2 三種置換方法的揭露風險及資料效用風險

(三)資料去識別方法評估指標

Shlomo 等人（2010）建議使用修正卡方值來衡量去識別變數資料關聯損失情況，令 χ^2 表示關聯分析的卡方檢定統計量，並定義 Cramèr V

$$CV = \sqrt{\frac{\chi^2/n}{\min(R-1, C-1)}}$$

其中，R 與 C 分別表示列變項與行變項的類別數，而 n 表示總樣本數。Shlomo 等人（2010）定義效用值（Utility Measure）為

$$RCV = 100 \times \frac{CV(T^P) - CV(T^O)}{CV(T^O)}$$

其中 $CV(T^P)$ 與 $CV(T^O)$ 分別代表利用置換樣本與原始樣本的 CV 值。若 RCV 值愈小，代表置換樣本的變數關聯沒有被改變太多，反之，則改變很多。

第三章 研究架構與方法

第一節 研究架構圖

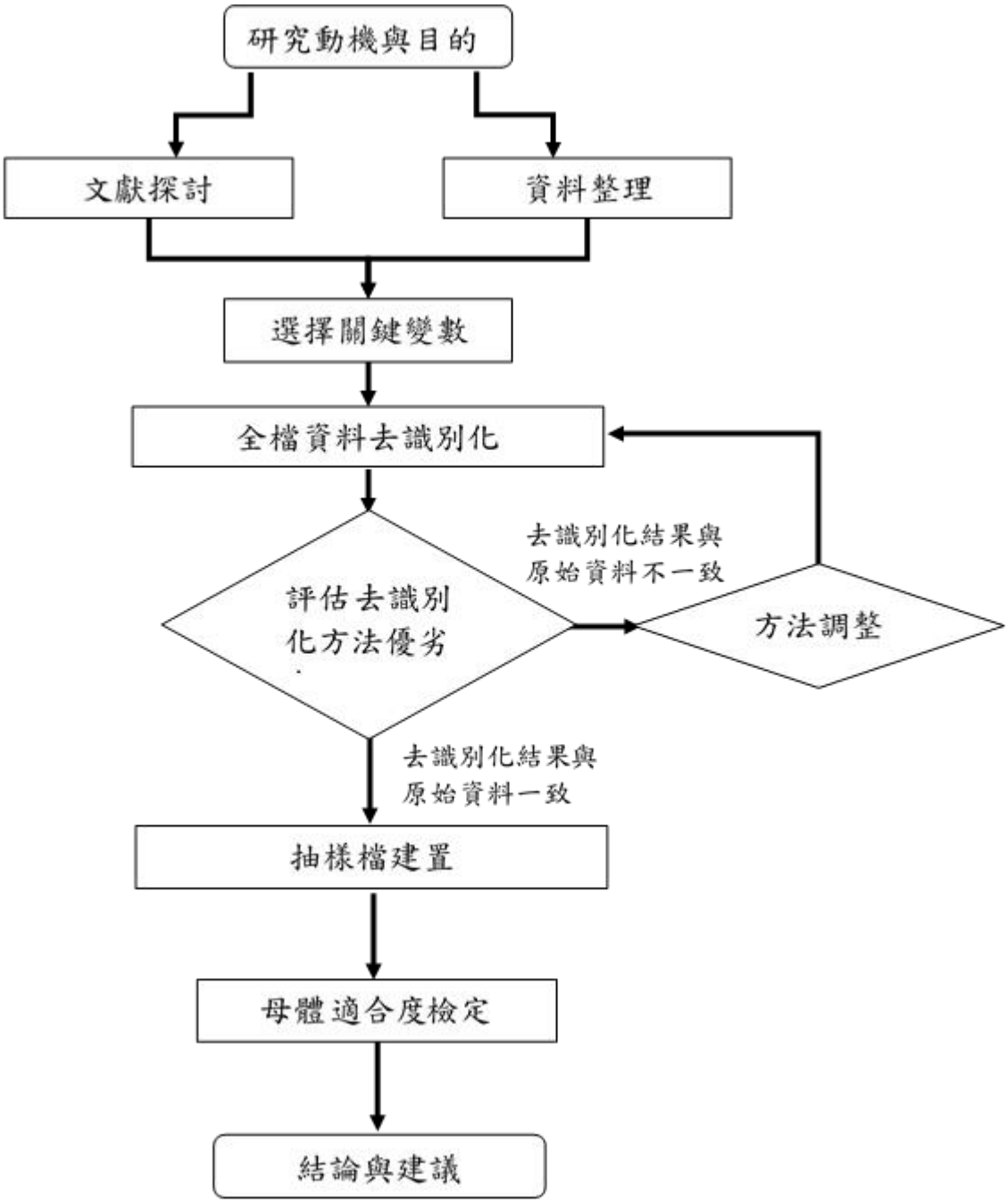

圖 3 研究架構圖

第二節 全檔資料去識別化方法

目前所蒐集有關農業普查的文獻資料尚未有國家有農業普查之外釋全檔資料，本計畫提供本國臺灣地區農林漁牧業普查外釋資料變數去識別化方法，家戶外釋風險分數，與去識別化方法評估，希冀增加農林漁牧業普查資料的實用性，並能兼顧家戶個資保護。

一、 變數處理方式

農林漁牧業普查資料的變項包含面積、人口、收入等許多連續型變數，與教育程度、主要經營種類等類別型變數，以下說明本計畫處理方式：

(一)類別型變數：

類別變項處理方式包含兩種，刪除數量過少之組別與合併組別。

(二)連續型變數：

1. 設定成類別型變數：

將連續型變數設定類別型變數的方式有等距（Equal interval）、等比率（Equal probability）或依照實務應用面建議。

例如人口普查資料年齡會利用等距方式，加拿大的家戶調查有問到距離與時間，設定方式也是採等距。

但若連續型變數的資料不對稱，採用等距方式分類，則

會造成分類結果過於集中於某些類別，如若資料右偏，則結果會集中在數值較小類別，而數值較大的類別，占比會偏少。但若仍要採用分類方式，則建議先進行變數變換，使轉換後變數資料較對稱，再使用上述分類方式進行分類。

2. 四捨五入法：

農林漁牧業普查資料的變項包含面積、人口、收入等許多有單位的連續型變數，可以採用不同基底，用四捨五入的方式處理這些變數。

3. 重新定義衍生變數（Derived variable）：

加拿大在抽樣調查方面外釋許多幾乎全檔的資料，除了少數變項，幾乎所有變項都是衍生變數，例如年齡、性別等基本變項，資料均使用家戶合成矩陣進行轉換。其餘的變項也會依照變數特性進行轉換。

除對資料進行去識別化，文獻中人口普查的資料還會進一步進行資料置換，將高風險的資料進行配對置換，配對的方式是利用鄰近區域，但由於農林漁牧業普查的區域單位範圍較大，林業與獨資漁戶的區域單位為縣市，總共有 20 縣市，而農牧戶的區域單位雖為鄉鎮，然而各區域異

質性高，要找到合適配對區域進行資料置換困難度高，不適合進行配對置換。因此，本計畫僅會將資料去識別化，而不會進行資料置換。

二、 資料去識別化後風險及關聯評估

(一)風險評估

分區域，首先須先決定一組關鍵變數，計算每一個關鍵變數的分布。假設有 M 個關鍵變數，每一個關鍵變數有 k_m 類，$m = 1,2,\cdots,M$。在區域 g，假設 $N_{k_m}^{g}$ 表示第 m 個關鍵變數的第 k_m 類的人數。因此細格個數會與風險分數成反比，令位於地區 g 的第 i 位受測者 m 個關鍵變數的分類組合為 $k = (k_1, \ldots, k_M)'$，參考 Shlomo 等人（2010）定義風險分數，定義為

$$HR_k^g = \left(\sum_{m=1}^{M} \frac{1}{N_{k_m}^{g}} \right) / M$$

以下簡稱第一種風險分數。

由於 Shlomo 等人（2010）利用分組人數來計算定義風險分數，若是分組的人數多，所得到的數值會非常集中在 0，為了讓數據可以有差異，本計畫建議採用分組比率取代分組人數，修正風險分數為

$$mHR_k^g = \left(\sum_{m=1}^{M} \frac{1}{p_{k_m}^{g}} \right) / M$$

其中 $p_{k_m}^{g}$ 表示第 m 個關鍵變數的第 k_m 類占比，以下簡稱第二種

風險分數。

假設風險門檻值，當風險分數超過門檻值則設定為高風險個體，個體為高風險，若家戶中有 1 人為高風險，則該家戶設定為高風險。

(二)關聯評估

本計畫採用多種去識別化方式來進行變數去識別化，但對資料進行去識別化一定會影響變數間的關聯，本計畫參考 Shlomo 等人（2010）使用效用值（Utility measure）來探討關聯喪失度，若效用值愈小，代表去識別化結果沒有影響原本資料變項間的關聯，代表去識別化方式愈合適。

Shlomo 等人（2010）原本所提的效用值利用檢定值來計算，但本計畫所討論的變項去識別化前後的資料屬性可能會從連續型式變成類別型式，關聯分析所採用的統計量檢定會不同，本計畫將公式微幅調整為

$$RCV^{a} = 100 \times \frac{pCV(T^{R}) - \mathrm{p}CV(T^{o})}{pCV(T^{o})}$$

其中 $pCV(T^{R})$與 $pCV(T^{o})$分別代表利用合併樣本與原始樣本運用檢定統計量所得到關聯性檢定的 p 值。若關鍵變數是連續型變數，運用 T 檢定值、 單因子變異數分析與 Kruskal-Wallis 無母數檢定計算得到 p 值，其中無母數檢定用於不對稱連續型變數資料，而前兩種用於對稱連續型變數資料。若關鍵變數是離散型式變數，則運用卡方檢定計算得到 p 值。

第三節 抽樣檔資料建置方法

各國大多數的微觀資料建置檔都是針對人口普查，為了避免個資揭露，應先將全檔的資料進行資料變數去識別化，但變數釋出的個數，則會由釋出的比率與釋出方式決定。依照釋出的最小地理單位，美國的釋出 1% 的外釋資料會有詳細的個人變項，而 5% 的外釋資料則會有詳細的區域變數，而英國釋出的方式，根據微觀資料等級區分，可開放給社會大眾免費使用之資料，變數個數及外釋資料數較少；需要簽署保密協定或只能遠端連線使用之資料，變數個數及外釋資料數較多。本計畫不會特別討論可釋出的變數個數，僅討論抽樣檔的建置方式。

抽樣檔的建置方式會按照業別採用區域抽樣，農牧戶與林業抽樣單位為縣市，而獨資漁戶抽樣單位為全國。農牧戶、林業與獨資漁戶各別提供 3 個、2 個與 3 個關鍵變數作為抽樣的分層變數（見表 4）。

分層的抽樣比例參考美國 PUMS 微觀資料，PUMS 的最小抽樣單位為人口普查小區（Census block），總共採用 6 個抽樣分層變數，一般來說若想了解一個變數的屬性，需要 30 筆觀察值，因此每個抽樣單位至少需包含 180 筆觀察值，若區域戶數少於 800 戶，組中點為 400 戶，抽樣率設定為 1:2，則可以抽出 200 戶，若介於 800 戶至 1,200 戶，組中點為 1,000 戶，抽樣率設定為 1:4，則可以抽出 250 戶，若介於 1,200

戶至 2,000 戶，組中點為 1,600 戶，抽樣率設定為 1:6，則可以抽出 267 戶，若超過 2,000 戶，抽樣率設定為 1:8，則至少抽出 250 戶，符合分層變數至少需要的觀察值筆數。

農牧戶與獨資漁戶資料有 3 個分層變數，而林業有 2 個分層變數，若以村里為抽樣單位，則每個抽樣單位至少需包含 90 筆與 60 筆觀察值，則本計畫設定方式應為區域人數少於 400 戶，組中點為 200 戶，抽樣率設定為 1:2，若介於 400 戶至 600 戶，組中點為 500 戶，抽樣率設定為 1:4，則可以抽出 125 戶，若介於 600 戶至 1,000 戶，組中點為 800 戶，抽樣率設定為 1:6，則可以抽出 134 戶，若超過 1,000 戶，抽樣率設定為 1:8，則至少抽出 125 戶。

由於本次研究資料能使用的最小抽樣單位為縣市，比美國人口普查小區（規模約為臺灣村里的大小）大，抽樣單位規模大，變異會較大，因此應抽取更多樣本。綜合考量變數個數與抽樣單位規模大小，故本次研究的抽樣比率設定方式，仍舊採用 PUMS 微觀資料設定區間。

表 4 抽樣檔建置方法

	農牧戶	林業	獨資漁戶
抽樣單位	縣市	縣市	全國
分層關鍵變數	主要經營種類 可耕作地面積 農牧業收入	主要經營種類 林業土地面積	主要經營種類 養繁殖總面積 漁業收入
抽樣比率	< 800 戶 → 1:2 800-1,199 戶 → 1:4 1,200-1,999 戶 → 1:6 >= 2,000 戶 → 1:8	< 800 戶 → 1:2 800-1,199 戶 → 1:4 1,200-1,999 戶 → 1:6 >= 2,000 戶 → 1:8	< 800 戶 → 1:2 800-1,199 戶 → 1:4 1,200-1,999 戶 → 1:6 >= 2,000 戶 → 1:8

參考美國 PUMS 抽樣檔設定方式，本計畫抽樣檔詳細建置流程如下：

(一)將各個抽樣區域依照分層變數排序，根據其各自的抽樣比率進行系統抽樣，以確保分層變數中的各個類別都能被抽取到。

(二)家戶初始權重為抽樣比率之倒數，為了不讓家戶配適過大或過小的權重，導致抽樣檔與全檔產生過多差異，因此需進行權重的迭代計算，使家戶可以得到合適的權重，迭代後權重不只用來推計原始總數，還要使分層變數能維持其關聯。

(三)從系統抽樣檔中，根據分層變數比率抽出總資料數 1% 的資料，即為 1% 的抽樣檔。抽出總資料數 5% 的資料，即為 5% 的抽樣檔。

最後，本計畫會利用卡方適合度檢定來檢視抽樣檔與全檔之關鍵變數分布是否一致，再使用 RCV^a 來檢查抽樣檔之分層變數與關鍵變數之關聯與全檔結果是否一致。

第四章 實證結果

第一節 全檔資料

一、 農牧戶

(一)普查項目

農牧戶普查表涵蓋資源分布與運用、家庭人口、勞動力特性、作物及畜禽生產情形及收入狀況等經營概況資訊，共十四問項，內容如下：

1. 全年農牧業經營情形。
2. 戶內人口數。
3. 年底可從事農作物栽培的可耕作地及人工鋪面面積。
4. 全年農作物種植情形。
5. 全年家畜家禽飼養情形。
6. 全年主要經營農牧業種類。
7. 全年自家初級農畜產品生產銷售分配情形。
8. 全年自家初級農畜產品加工情形。
9. 全年經營休閒農業類型及其面積。
10. 全年經營農業加工、休閒以外之相關事業。
11. 各月份從事自家農牧業工作外僱人力。

12. 戶內滿 15 歲以上人口特性與工作狀況，每位蒐集（1）稱謂、（2）性別、（3）出生年次、（4）教育程度、（5）農牧業身分、（6）從農年資、（7）全年從事自家農牧業工作日數、（8）全年主要工作狀況。

13. 全年戶內人口從事自家農牧業外工作情形。

14. 全年農業相關收入。

首先，為了確認資料外釋安全性，以可耕作地面積、農牧戶經營管理者性別、年齡、教育程度及 104 年從事自家農牧業工作日數、主要經營種類、農牧業收入、從事自家農牧業工作人數等 8 個關鍵變數計算唯一資料比率，其中得出唯一資料比率高達 91.9 %，因關鍵變數中包含連續型變數，會導致唯一資料比率過高，在只考慮類別型關鍵變數情況下，唯一資料比率為 3.39 %。

(二)類別型關鍵變數處理

由於未經處理的原始檔唯一資料比率過高，為了降低唯一資料的比率，對於部分連續型關鍵變數進行分組處理；部分類別型關鍵變數之分類家數太少者，則進行合併分類，其中除了未從事農牧業者占 7.39 %，其餘的選項詳細說明如下：

1. 經營管理者性別：男性占 73.24 %；女性占 19.36 %。

2. 經營管理者年齡：未滿 45 歲占 5.23 %；45-54 歲占 16.85 %；55-64

歲占 28.00％；65-74 歲占 22.87％；75 歲以上占 19.66％。

3. 經營管理者教育程度：不識字占 8.41％；小學及自修占 33.5％；國(初)中占 20.49％；高中(職)占 22.09％；大專及以上占 8.12％。

4. 經營管理者 104 年全年從事自家農牧業工作日數： 1-59 日占 46.22％；60-149 日占 30.7％；150 日以上占 15.69％。

5. 主要經營種類：稻作休耕占 6.51 ％；轉型休閒占 0.01 ％；稻作占 27.68％；雜糧占 7.61％；特用作物占 4.36％；蔬菜占 18.28％；果樹占 23.9％;食用菇蕈占 0.16％;花卉占 0.65％;其他農作物占 1.71％；牛占 0.1％；豬占 0.67％；其他家畜占 0.19％；雞占 0.54％；鴨占 0.15％；其他家禽占 0.04％；其他畜牧業占 0.04％。

6. 從事自家農牧業工作人數：從事自家農牧業工作人數最少 1 人，最多 13 人，因為從事自家農牧業工作人數 3 人以上所占戶數不多，因此有工作人數者分三類，從事自家農牧業工作人數 1 人占 34.58％；從事自家農牧業工作人數 2 人占 38.65％ ；從事自家農牧業工作人數 3 人以上占 19.38％。

(三)連續型關鍵變數處理

另可耕作地面積與農牧業收入兩個關鍵變數和年底面積_1 到年底面積_25、單次最大種植面積或全年數量_1 到單次最大種植面積或全年數量_50、自家初級農畜產品銷售收入、自行加工農畜產品銷售收入、委外加

工農畜產品銷售收入、休閒農業服務收入等細項連續型變數，因為分布過於右偏，類別化未必為去識別化之最適方法，因此考慮四捨五入法及衍生變數法，探討三種方法何者為最佳的去識別化方法。

1. 類別化：將上述變數先取對數後，扣除沒有值的戶數，再依等比率的區間分組。

(1) 可耕作地面積：無面積者占 0.62％；1-23 公畝占 22.92％；24-36 公畝占 19.94％；37-54 公畝占 18.76％；55-93 公畝占 18.82％；94 公畝以上占 18.93％。

(2) 年底面積_1-年底面積_25： 1-15 公畝占 19.51 ％；16-27 公畝占 20.21％；28-40 公畝占 19.59％；41-70 公畝占 20.96％；71 公畝以上占 19.72％。

(3) 單次最大種植面積或全年數量_1-單次最大種植面積或全年數量_50：

① 計量單位為「公畝」： 1-9 公畝占 20.86％、10-20 公畝占 20.53％、21-35 公畝占 20.30 ％、36-60 公畝占 19.63 ％、61 公畝以上占 18.68％。

② 計量單位為「公斤」：1-300 公斤占 34.90％、301-6,000 公斤占 32.21 ％、6,001 公斤以上占 32.89％。

③ 計量單位為「箱」：1-5,000 箱占 34.09 ％、5,001- 50,000 箱占

33.33 %、50,001 箱以上占 32.58 %。

④ 計量單位為「盆」:1-1,200 盆占 33.94 %、1,201-16,000 盆占 33.43 %、16,001 盆以上 32.63 %。

⑤ 計量單位為「包」:1- 90,000 包占 35.34 %、90,001-190,000 包 31.75 %、190,001 包以上 32.92 %。

⑥ 計量單位為「瓶」:1- 180,000 瓶占 50.00 %、180,001 瓶以上占 50.00 %。

(4) 農牧業收入:未從事農牧業者占 7.39 %;無收入者占 19.88 %;1-50 千元占 16.05 %;51-100 千元占 14.72 %;101-200 千元占 14.99 %;201-420 千元占 12.56 %;421 千元以上占 14.4 %。

(5) 初級農畜產品銷售收入:未從事農牧業者占 7.39 %;無收入者占 20.30 %;1-48 千元占 14.57 %;49-96 千元占 14.42 %;97-181 千元占 14.37 %;182-400 千元占 14.57 %;401 千元以上占 14.38 %。

(6) 自行加工農畜產品銷售收入:因為此項收入大部分為 0,因此扣除未從事農牧業者,分無、有兩類:無占 91.86 %;有占 0.75 %。

(7) 委外加工農畜產品銷售收入:因為此項收入大部分為 0,因此扣除未從事農牧業者,分無、有兩類:無占 92.05 %;有占 0.55 %。

(8) 休閒農業服務收入:因為此項收入大部分為 0,因此扣除未從事農

牧業者，分無、有兩類：無占 92.45％；有占 0.16％。

2. 四捨五入法：1-999 四捨五入至十位，1,000-9,999 四捨五入到百位，10,000-99,999 四捨五入到千位，以此類推。若個位數四捨五入後為 0，則轉換成 1；原本為 0 的值，四捨五入後一樣為 0。例如：75 四捨五入至十位為 80；275 四捨五入至十位為 280；1,275 四捨五入至百位為 1,300；3 四捨五入至十位為 0，再轉換成 1。由於資料過於右偏，為了使極端值較不易被辨認，將第 90 百分位以上的值以該區間之中位數取代（第 90 百分位以上全部以第 95 百分位取代）。

3. 衍生變數法：將上述連續型變數算各縣市占全臺灣的比率，新增變數為縣市總面積占比（或縣市總收入占比），其為一個類別化的區間，此區間依縣市個數大致相等分為以下區間，使其無法識別原始資料，如表 5。

表 5 農牧戶衍生變數法之類別化區間縣市個數

可耕作地面積類別化區間	未滿 3％	3％- 未滿 7％	7％以上
縣市個數	7	6	7
農牧業收入類別化區間	未滿 1.5％	1.5％-未滿 7％	7％以上
縣市個數	7	6	7

由於總比率已經模糊化，因此所有細項可以直接設定為占該鄉鎮市區的比率。例如：第一筆資料在宜蘭縣，宜蘭縣可耕作地總面積為 1,693,207 公畝，全臺灣可耕作地總面積為 53,768,627 公畝，因此宜蘭縣可耕作地總面積為 3％-未滿 7％；宜蘭縣農牧業收入為 5,101,105 千元，全臺灣農牧業收入為 241,500,647 千元，因此宜蘭縣農牧業收入之縣市總收入占比為 1.5 % -未滿 7 %。而所有細項皆是占鄉鎮市區的比率，如表 6 和表 7 所示，使用時宜蘭縣宜蘭市兩筆資料及臺北市士林區兩筆資料可以相互比較大小，若要全臺灣相比，則乘以縣市總面積占比(或縣市總收入占比)的組中點，即可用來比較。

表 6 農牧戶衍生變數資料說明(以可耕作地總面積為例)

編號	縣市	縣市可耕作地占全臺總面積比率	鄉鎮市區	鄉鎮市區可耕作地占該縣市總面積比率	可耕作地占該鄉鎮市區總面積比率
00008241107017	宜蘭縣	3％- 未滿 7％	宜蘭市	9.52760％	0.01612％
00069741107017	宜蘭縣	3％- 未滿 7％	宜蘭市	9.52760％	0.02542％
00232141107017	臺北市	未滿 3％	士林區	17.8418％	1.62162％
00310741107017	臺北市	未滿 3％	士林區	17.8418％	0.17230％

註 1:宜蘭縣總面積 1,693,207 公畝，宜蘭市總面積 161,322 公畝，全臺灣總面積 53,768,627 公畝。
註 2:臺北市總面積 553,011 公畝，士林區總面積 98,667 公畝，全臺灣總面積 53,768,627 公畝。

表 7 農牧戶衍生變數資料說明(以農牧業收入為例)

編號	縣市	縣市農牧業收入占全臺總收入比率	鄉鎮市區	鄉鎮市區農牧業收入占該縣市總收入比率	農牧業收入占該鄉鎮市區總收入比率
00008241107017	宜蘭縣	1.5 % -未滿 7 %	宜蘭市	8.09497 %	0.01138 %
00069741107017	宜蘭縣	1.5 % -未滿 7 %	宜蘭市	8.09497 %	0.03608 %
00232141107017	臺北市	未滿 1.5 %	士林區	27.4665 %	0.05160 %
00310741107017	臺北市	未滿 1.5 %	士林區	27.4665 %	0.29488 %

註 1:宜蘭縣總收入 5,101,105 千元，宜蘭市總收入 412,933 千元，全臺灣總收入 241,500,647 千元。

註 2:臺北市總收入 987,723 千元，士林區總收入 271,293 千元，全臺灣總收入 241,500,647 千元。

接著使用關聯分析來比較三種去識別化方法何者為最適合的方法。可耕作地總面積和農牧業收入在原始資料、四捨五入法及衍生變數法皆為非常態分布的連續型變數，因此使用 Kruskal-Wallis Test 來檢定不同組別間的中位數是否有差異；而類別化方法將可耕作地總面積和農牧業收入轉為類別型變數，因此使用卡方獨立性檢定分析其關聯。

由表 8 和表 9 可以看到，原始資料中，可耕作地總面積和農牧戶經營管理者性別、年齡、教育程度與主要經營種類四個變數皆有顯著關聯；而農牧業收入和農牧戶經營管理者性別與主要經營種類兩個變數有顯著關聯，和農牧戶經營管理者教育程度及年齡兩個變數則無顯著關聯。

表 8 可耕作地總面積去識別化方法關聯分析比較

		方法			
		原始資料(1)	類別化(2)	四捨五入法(3)	衍生變數法(4)
去識別變數	比較變數	P 值	P 值	P 值	P 值
可耕作地總面積	性別	<0.0001***	<0.0001***	<0.0001***	0.00322**
	年齡	<0.0001***	<0.0001***	<0.0001***	0.00113**
	教育程度	<0.0001***	0.00091***	<0.0001***	<0.0001***
	主要經營種類	<0.0001***	<0.0001***	<0.0001***	<0.0001***

註 1: (1)(3)(4)使用 Kruskal-Wallis Test；(2)使用卡方獨立性檢定。

註 2:*在 P 值為 0.05 顯著；**在 P 值為 0.01 顯著；***在 P 值為 0.001 顯著。

表 9 農牧業收入去識別化方法關聯分析比較

		方法			
		原始資料(1)	類別化(2)	四捨五入法(3)	衍生變數法(4)
去識別變數	比較變數	P 值	P 值	P 值	P 值
農牧業收入	性別	0.00185**	<0.0001***	0.00204**	0.27262
	年齡	0.58907	<0.0001***	0.61632	1.16634
	教育程度	0.91468	<0.0001***	0.8472	0.05969
	主要經營種類	<0.0001***	<0.0001***	0.00014***	<0.0001***

註 1: (1)(3)(4)使用 Kruskal-Wallis Test；(2)使用卡方獨立性檢定。

註 2:*在 P 值為 0.05 顯著；**在 P 值為 0.01 顯著；***在 P 值為 0.001 顯著。

類別化的可耕作地總面積和農牧戶經營管理者性別、年齡、教育程度與主要經營種類四個變數皆有顯著關聯，與原始資料結果相同；而農牧業收入和農牧戶經營管理者性別與主要經營種類兩個變數有顯著關聯，與原始資料結果相同，而和農牧戶經營管理者年齡及教育程度亦有顯著關聯，與原始資料結果不同。

四捨五入法不論在可耕作地總面積或農牧業收入和農牧戶經營管理者性別、年齡及教育程度與主要經營種類四個變數之關聯性，與原始資料

結果相同。

衍生變數法的可耕作地總面積和農牧戶經營管理者性別、年齡及教育程度與主要經營種類四個變數皆有顯著關聯，與原始資料結果相同；而農牧業收入和農牧戶經營管理者性別沒有顯著關聯，與原始資料結果不同，和農牧戶經營管理者年齡與教育程度兩個變數亦無顯著關聯，和主要經營種類則有顯著關聯，皆與原始資料結果相同。

三種去識別化的方法中，可耕作地總面積在類別化和農牧戶經營管理者教育程度之 P 值為 0.00091，與原始資料仍有些微差異，在衍生變數法和農牧戶經營管理者性別、年齡之 P 值與原始資料也有些微差異，在四捨五入法四個變數皆可保有原始資料的特性，且只有 21.21 % 的戶數可耕作地總面積未改變，因此建議面積使用四捨五入法為去識別化方法；而農牧業收入則可明顯看出類別化及衍生變數法關聯性與原始資料不相同，雖然四捨五入法有 54.30 %的戶數農牧業收入未改變，有 54.43 %的戶數自家初級農畜產品銷售收入未改變，其餘細項收入更高達 90 %以上都未改變，但僅有四捨五入法保有與原始資料相同關聯的特性，因此建議收入使用四捨五入法為去識別化方法。

上述未提及的變數，面積的部分：年底面積_1 到年底面積_25、單次最大種植面積或全年數量_1 到單次最大種植面積或全年數量_50 雖然不為重要變數，但仍有外釋風險，因此使用類別化法；而收入的部分：自家

初級農畜產品銷售收入使用四捨五入法，其餘細項收入則採用類別化法，進行去識別化。其餘的變數會與原始資料相同，未做其他處理。

(四)風險分數評估

以四捨五入法處理過後的資料計算兩種風險分數，發現大部分資料的風險分數相近，只有少數風險分數很高，如高於 99.5 百分位點的資料有較高的揭露風險，第一種風險分數第 99.5 百分位點為 1.0149，第二種風險分數第 99.5 百分位點為 1,505.9078。兩種風險分布的圖如圖 4和圖 5。四捨五入法處理後的唯一資料比率降為 65.69％，若只考慮類別型關鍵變數，唯一資料比率則降至 1.77％。因此，全檔關鍵變數既已處理，使得高風險資料之風險分數大幅降低，故不予以刪除。

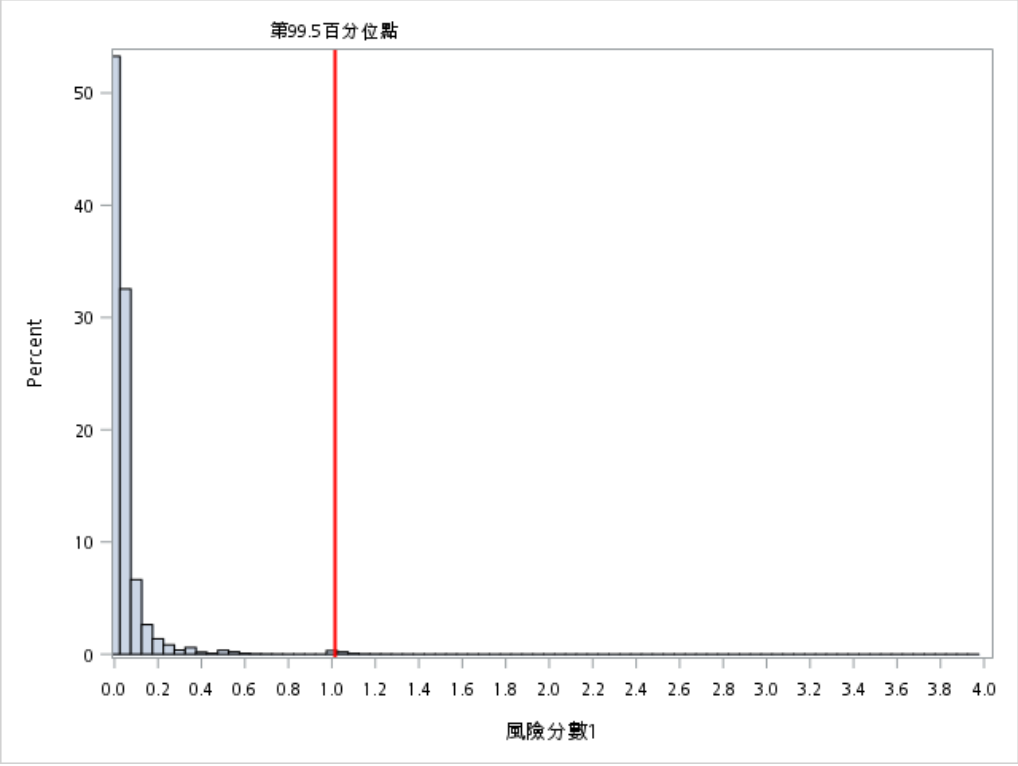

圖 4 農牧戶第一種風險分數

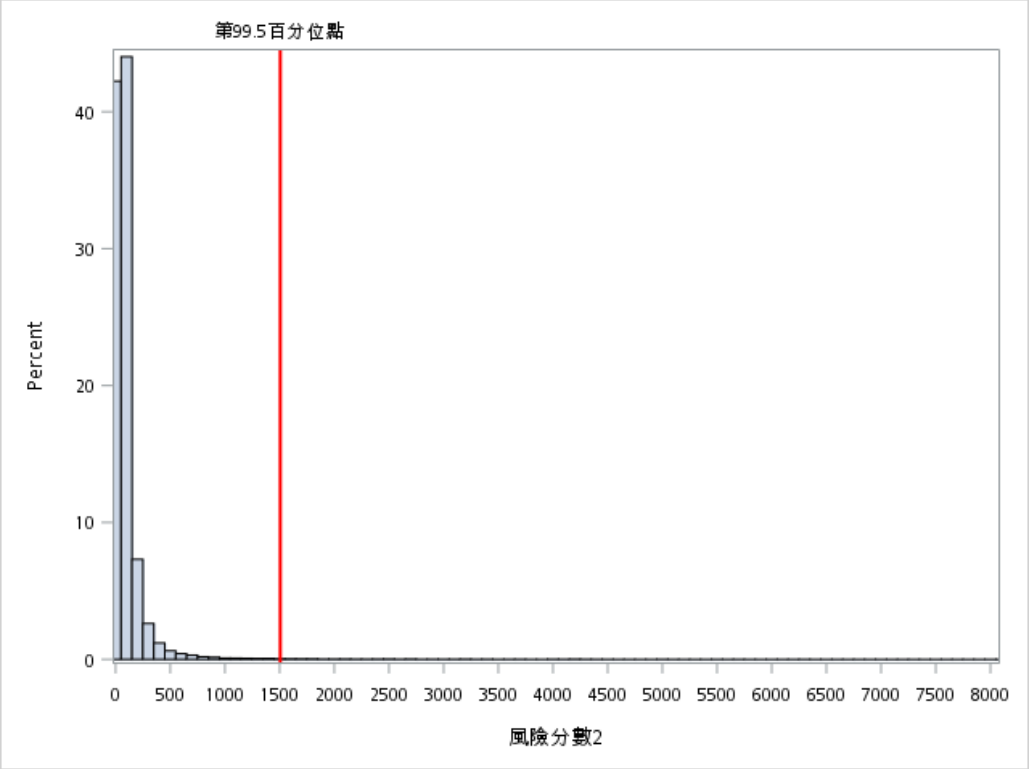

圖 5 農牧戶第二種風險分數

二、 林業

林業普查表涵蓋資源分布與用途、家庭人口、勞動力特性、森林作業情形及收入狀況等經營概況資訊，共九問項，內容如下：

1. 經營組織型態。
2. 戶內人口情形。
3. 實際負責經營管理者特性。
4. 全年主要經營林業種類。
5. 林業土地面積。
6. 仍參與政府造林補助之林地面積。

7. 全年從事森林作業情形。

8. 從事林業相關工作人數。

9. 全年林業收入。

林業依組織別可分為兩大類的資料：林戶家數占 99.64％；林場家數占 0.36%。林戶資料筆數多，總面積僅占 6.81％，占比小；但林場資料筆數少，但總面積占 93.19％，占比大。由於兩類資料合併處理會使資料產生偏誤，因此本次林業將兩類資料分開進行處理，並產出林戶及林場兩個全檔資料以供使用。

(一)林戶

由於澎湖縣林戶只有一筆資料，且其林地位於屏東縣，因此將其併入屏東縣計算。

首先，為了確認資料外釋安全性，以林業經營管理者性別、年齡、教育程度、主要經營種類、林業土地面積、面積按主要所有權屬分、面積按主要林相種類分、面積按主要功能用途分、有無參與政府造林補助之林地、有無從事森林作業等 10 個關鍵變數計算唯一資料比率，其中得出唯一資料比率高達 89.05％。因關鍵變數中包含連續型變數，會導致唯一資料比率過高，在只考慮類別型關鍵變數情況下，唯一資料比率為 34.18％。

1. 類別型關鍵變數處理

由於未經處理的原始檔唯一資料過多，為了降低唯一資料的比率，對於部分連續型式的關鍵變數進行分組處理；部分類別型式的關鍵變數中分類數或人數太少者，則需進行選項合併，詳細說明如下：

(1) 林業經營管理者性別：男性占 77.56 %；女性占 22.44 %。

(2) 林業經營管理者年齡：未滿 45 歲以下占 8.65 %；46-54 歲占 20.25%；55-64 歲占 30.36 %；65-74 歲占 21.29 %；75 歲以上占 19.45 %。

(3) 林業經營管理者教育程度：不識字占 5.64 %；小學及自修占 33.53 %；國（初）中占 22.84 %；高中（職）占 25.82 %；大專及以上占 12.17 %。

(4) 主要經營林業種類：由於經營森林遊樂業資料筆數只有 8 筆，且林相多為闊葉樹林或針闊葉混淆林，故將其併入一般林木經營業，分類結果如下：一般林木經營業占 70.67 %；特殊林木經營業占 29.33 %。

(5) 林地面積按主要所有權屬分：資料中已有總面積，此變數雖為連續型式變數，但目的是為了區分各種類土地面積利用的大小，因此以各筆資料占比最高的類別作為此變數代表，分類結果如下：所有權屬為自有林地占 79.02 %；租借國、公有林地占 20.18 %；

租借私有林地占 0.66 %；接受委託經營占 0.13 %。

(6) 林地面積按主要林相種類分：資料中已有總面積，此變數雖為連續型式變數，但目的是為了區分各種類土地利用的面積大小，因此以各筆資料占比最高的類別作為此變數代表，分類結果如下：林相種類為闊葉樹林占 47.29 %；竹林占 27.12 %；針闊葉混淆林占 18.41 %；針葉樹林占 6.66%；未立木地占 0.53 %。

(7) 林地面積按主要功能用途分：資料中已有總面積，此變數雖為連續型變數，但目的是為了區分各種類土地利用的面積大小，因此以各筆資料占比最高的類別作為此變數代表，分類結果如下：功能用途為生產林木占 81.33 %；國土保安占 9.94 %；自然保護占 8.20 %；其他占 0.3 %；伐跡地、新開墾地占 0.22 %；森林遊樂占 0.01 %。

(8) 有無參與政府造林補助：無參與政府造林補助占 85.09 %；有參與政府造林補助占 14.91 %。

(9) 有無從事森林作業：無從事森林作業占 58.14 %；有從事森林作業占 41.86 %。

2. 連續型關鍵變數處理

由於林業土地面積分布過於右偏，使用類別化方法將資料分組，可能使資料代表性不足，在此未必為合適去識別化方法，因此另外使

用四捨五入法及衍生變數法，來探討何者為最佳的去識別化方法。

(1) 類別化：總面積最小為 10 公畝，最大為 49,505 公畝，因資料極度右偏，先對其取對數使資料分布均勻，再使用等比率分為五組：24 公畝以下占 19.1 %；25-49 公畝占 18.1%；50-99 公畝占 22.7%；100-199 公畝占 20.0 %；200 公畝以上占 20.1%。

(2) 四捨五入法：1-999 四捨五入至十位，1,000-9,999 四捨五入到百位，10,000-99,999 四捨五入到千位，以此類推。若四捨五入後為 0，則轉換成 1。原本為 0 的值，四捨五入後一樣為 0。由於資料過於右偏，為了使極端值較不易被辨認，將第 90 百分位以上的值以該區間之中位數取代（第 90 百分位以上全部以第 95 百分位取代）。

(3) 衍生變數法：計算各縣市林業土地面積占全臺灣的比率，新增變數為縣市總面積占比，其為一個類別化的區間，此區間依縣市個數大致相等分為以下區間，使其無法識別原始資料，如表 10。

表 10 林業衍生變數法之類別化區間縣市個數

林業土地面積類別化區間	未滿 1 %	1 %- 未滿 5 %	5 %-未滿 10 %	10 % 以上
縣市個數	5	5	6	3

註:澎湖縣併入屏東縣計算。

由於總比率已經模糊化，因此所有細項可以直接給占該縣市的比率。例如：第一筆資料在宜蘭縣，宜蘭縣林戶林業土地面積為 503,285 公畝，全臺灣林戶林業土地面積為 12,806,393 公畝，因此宜蘭縣林業土地面積之縣市總面積占比為 1 %-未滿 5%。而所有細項皆是占縣市的比率，如表 11 所示，使用時宜蘭縣兩筆資料可以相互比較大小，若要全臺灣相比，則乘以縣市總面積占比的組中點，即可用來比較。

表 11 林業衍生變數資料說明(以林業土地面積為例)

編號	縣市	林業 土地面積	林業土地占該縣 市總面積比率	縣市林業土地占 全臺面積比率
00000541107017	宜蘭縣	19	0.00378 %	1 %-未滿 5%
00004441107017	宜蘭縣	15	0.00298 %	1 %-未滿 5%
00805341107017	新竹縣	30	0.00312 %	5 %-未滿 10%

註:宜蘭縣總面積 503,285 公畝，新竹縣總面積 961,183 公畝，全臺灣總面積 12,806,393 公畝。

使用關聯分析來比較三種去識別化方法何者為最適合的方法。林業土地面積在原始資料、四捨五入法及衍生變數法皆為非常態分布的連續型變數，因此使用 Kruskal-Wallis Test 來檢定不同組別間的中位數是否有差異，而類別化方法將林業土地面積轉為類別型變數，因此使用卡方獨立性檢定分析其關聯。由表 12 可以看到，原始資料中，林業土地面積和林業經營管理者性別、年齡、教育程度與主要經營種類四個變數皆有顯著關

聯。三種去識別化方法的關聯性與原始資料皆相同，因此以下分別討論各方法優劣。

表 12 林業土地總面積去識別化方法關聯分析比較

		方法			
		原始資料(1)	類別化(2)	四捨五入法(3)	衍生變數法(4)
去識別變數	比較變數	P 值	P 值	P 值	P 值
林業土地面積	性別	<0.0001***	<0.0001***	<0.0001***	<0.0001***
	年齡	<0.0001***	<0.0001***	<0.0001***	<0.0001***
	教育程度	<0.0001***	<0.0001***	<0.0001***	<0.0001***
	主要經營種類	<0.0001***	<0.0001***	<0.0001***	<0.0001***

註 1: (1)(3)(4)使用 Kruskal-Wallis Test；(2)使用卡方獨立性檢定。

註 2:*在 P 值為 0.05 顯著；**在 P 值為 0.01 顯著；***在 P 值為 0.001 顯著。

本次林業普查林業土地面積資料過於右偏，使用類別化對於資料分類過於籠統，可能使資料代表性不足。使用四捨五入法去識別後，資料分布改變最少，但未改變的資料占 49.12%。衍生變數法在資料應用上較不直觀，研究者可能難以使用。綜合以上，類別化可能使資料代表性不足，衍生變數法較難以應用，因此建議林業普查林業土地面積使用四捨五入法為去識別化方法。

上述未提及的變數，面積的部分：按所在地區分面積_1 到按所在地區分面積_10、新造林面積、砍採伐面積等變數雖然不為重要變數，但因其面積為原始資料，有外釋風險，因此使用類別化法去識別化，各項收入雖然不為重要變數，但因收入為原始資料，有外釋風險，因此使用類別化

法去識別化，其餘的變數則與原始資料相同，未做其他處理。

3. 風險分數評估

以四捨五入法處理過後的資料計算兩種風險分數，發現大部份資料的風險分數相近，只有少數風險分數很高，如高於 99 百分位點的資料有較高的揭露風險，第一種風險分數第 99 百分位點為 0.3771，第二種風險分數第 99 百分位點為 1,861.0078。兩種風險分布的圖如圖 6 和圖 7。四捨五入法處理後的唯一資料比率降為 40.73 %，而只考慮類別型變數的唯一資料比率降為 9.92 %。因此，全檔關鍵變數既已處理，使得高風險資料之風險分數大幅降低，故不予以刪除。

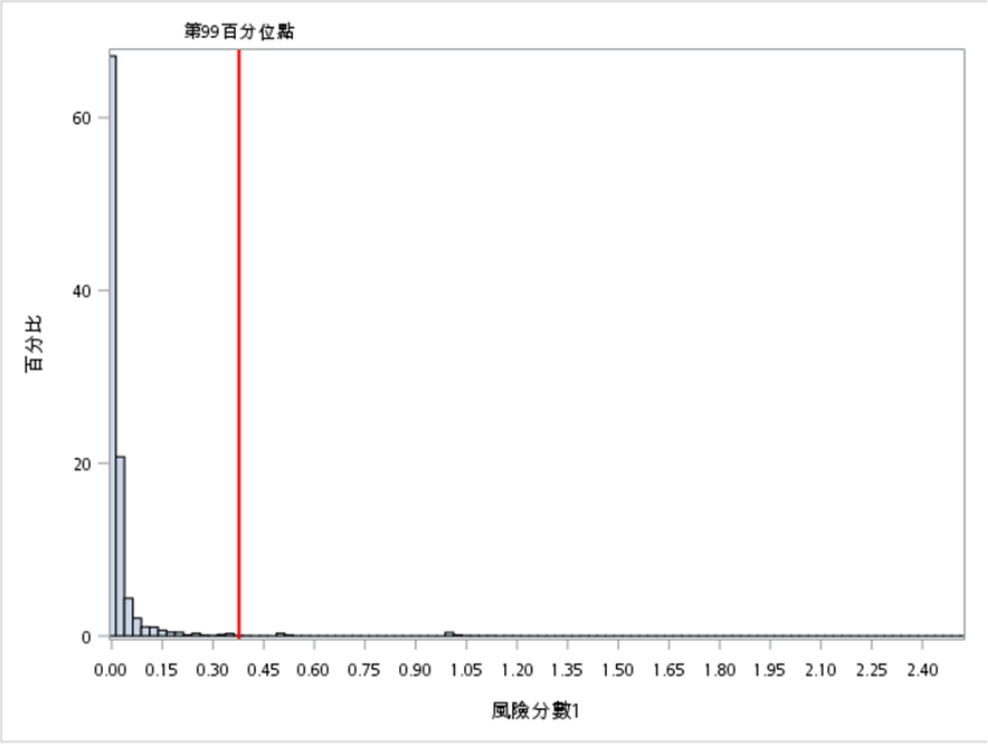

圖 6 林業第一種風險分數

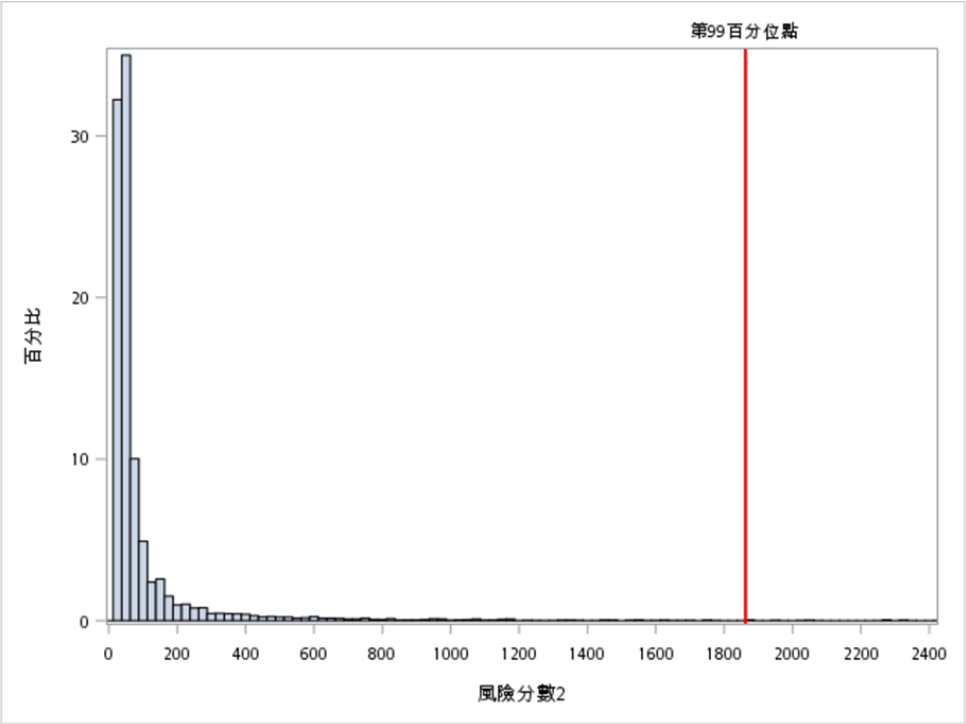

圖 7 林業第二種風險分數

(二)林場

1. 類別型關鍵變數處理

(1) 林業經營管理者性別：男性占 86.90 %；女性占 13.10 %。

(2) 林業經營管理者年齡：未滿 45 歲以下占 6.71 %；45-54 歲占 26.52 %；55-64 歲占 51.76 %；65-74 歲占 10.54 %；75 歲以上占 4.47 %。

(3) 林業經營管理者教育程度：不識字占 0.32 %；小學及自修占 6.71 %；國（初）中占 6.39 %；高中（職）占 24.92 %；大專及以上占 61.66 %。

(4) 主要經營林業種類：一般林木經營業占 80.19 %；特殊林木經營業占 6.39 %；經營森林遊樂業占 13.42 %。

(5) 林地面積按主要所有權屬分：資料中已有總面積，此變數雖為連續型式變數，但目的是為了區分各種類土地面積利用的大小，因此以各筆資料占比最高的類別作為此變數代表，分類結果如下：所有權屬為自有林地占 72.84 %；租借國、公有林地占 17.25 %；租借私有林地占 1.92 %；接受委託經營占 7.99 %。

(6) 林地面積按主要林相種類分：資料中已有總面積，此變數雖為連續型式變數，但目的是為了區分各種類土地利用的面積大小，因此以各筆資料占比最高的類別作為此變數代表，分類結果如下：林相種類為闊葉樹林占 59.42 %；竹林占 6.39 %；針闊葉混淆林占 24.28 %；針葉樹林占 9.27%；未立木地占 0.64 %。

(7) 林地面積按主要功能用途分：資料中已有總面積，此變數雖為連續型變數，但目的是為了區分各種類土地利用的面積大小，因此以各筆資料占比最高的類別作為此變數代表，分類結果如下：功能用途為生產林木占 59.42 %；國土保安占 15.34 %；自然保護占 11.82 %；其他占 1.60 %；伐跡地、新開墾地占 0.96 %；森林遊樂占 10.86 %。

(8) 有無參與政府造林補助：無參與政府造林補助占 70.61 %；有參

與政府造林補助占 29.39 %。

(9) 有無從事森林作業：無從事森林作業占 33.87 %；有從事森林作業占 66.13 %。

2. 連續型關鍵變數處理

由於林業土地面積分布過於右偏，使用類別化方法將資料分組，可能使資料代表性不足，因此非為去識別化之最適方法，另外使用四捨五入法及衍生變數法，來探討何者為最佳的去識別化方法。

(1) 類別化：總面積最小為 10 公畝，最大為 32,070,432 公畝，因資料極度右偏，先對其取對數使資料分布均勻，再使用等比率分為四組。300 公畝以下占 24.28 %；301-5,000 公畝占 29.07 %；5,001-25,000 公畝占 22.36 %；25,001 公畝以上占 24.28 %。

(2) 四捨五入法：1-999 四捨五入至十位，1,000-9,999 四捨五入到百位，10,000-99,999 四捨五入到千位，以此類推。若四捨五入後為 0，則轉換成 1。原本為 0 的值，四捨五入後一樣為 0。由於資料過於右偏，為了使極端值不易被辨認，將第 90 百分位以上的值以該區間之中位數取代（第 90 百分位以上全部以第 95 百分位取代）。

(3) 衍生變數法：計算各林場林業土地面積占全臺灣的比率，新增變數林場總面積占比，其為一個類別化的區間：未滿 0.0002 % 、

0.0002 % - 0.002 %（未含）、0.002 % - 0.015%（未含）、0.015 % 以上，使其無法識別原始資料，如表 13。

表 13 林場類別化區間次數及百分比

林場總面積占比	次數	百分比
未滿 0.0002 %	81	25.88
0.0002 % - 未滿 0.002 %	73	23.32
0.002 % -未滿 0.015 %	84	26.84
0.015 % 以上	75	23.96

例如：第三筆林場資料林業土地面積為 42,993 公畝，全臺灣林場林業土地面積為 175,046,358 公畝，因此此筆資料總面積占比為 0.015 % 以上，如表 14 所示，使用時各林場間可以組中點相互比較大小。

表 14 林業衍生變數資料說明(林場)

編號	林業土地面積	林業土地面積占比
00020341107017	1,300	0.0002 % - 未滿 0.002 %
00028241107017	5,958	0.002 % -未滿 0.015 %
00051941107017	42,993	0.015 % 以上

註:全臺灣總面積 175,046,358 公畝。

使用關聯分析來比較三種去識別化方法何者為最適合的方法。林業土地面積在原始資料、四捨五入法及衍生變數法皆為非常態分布的

連續型變數，因此使用 Kruskal-Wallis Test 來檢定不同組別間的中位數是否有差異，而類別化方法將林業土地面積轉為類別型變數，因此使用卡方獨立性檢定分析其關聯。由表 15 可以看到，原始資料中，林業土地面積和林業經營管理者性別、年齡及教育程度三個變數皆有顯著關聯。林業土地面積與主要經營種類關聯性 P 值大約為 0.06，三種方法中只有四捨五入法的關聯性與原始資料皆相同，因此建議林業普查林場林業土地面積使用四捨五入法為去識別化方法。

表 15 林業土地總面積去識別化方法關聯分析比較(林場)

		方法			
		原始資料(1)	類別化(2)	四捨五入法(3)	衍生變數法(4)
去識別變數	比較變數	P 值	P 值	P 值	P 值
林業土地面積	性別	0.00606**	0.00136**	0.00641**	0.00172**
	年齡	0.00035***	0.04095*	0.00012***	<0.0001***
	教育程度	<0.0001***	<0.0001***	<0.0001***	<0.0001***
	主要經營種類	0.06157	0.76031	0.06335	0.10638

註 1: (1)(3)(4)使用 Kruskal-Wallis Test；(2)使用卡方獨立性檢定。

註 2:*在 P 值為 0.05 顯著；**在 P 值為 0.01 顯著；***在 P 值為 0.001 顯著。

由於林場資料全部公布，因此不對其進行風險分數的計算。

三、 獨資漁戶

(一)普查項目

獨資漁戶普查涵蓋資源分布與運用、家庭人口、勞動力特性、漁撈及水產養繁殖作業情形及收入狀況等經營概況資訊，共十六問項，內容如

下：

1. 全年漁業經營情形。
2. 戶內人口數。
3. 年底漁撈使用的漁船情形。
4. 全年「在沿岸、河川或湖泊之不使用漁船採捕作業」情形。
5. 年底可從事養繁殖的面積。
6. 年底養繁殖使用的漁船。
7. 全年主要經營漁業種類。
8. 全年自家初級漁產品生產銷售分配情形。
9. 全年自家初級漁產品加工情形。
10. 全年主要經營休閒漁業類型。
11. 全年經營漁業加工、休閒以外之相關事業。
12. 各月份從事自家漁業之外僱人力。
13. 年底從事自家漁業工作之外僱人力種類及來源。
14. 戶內 15 歲以上人口特性與工作狀況，每位蒐集（1）稱謂、（2）性別、（3）出生年次、（4）教育程度、（5）漁業身分、（6）全年從事自家漁業工作日數、（7）全年主要工作狀況。
15. 全年戶內人口從事自家漁業外工作情形。
16. 全年漁業相關收入。

首先，為了確認資料外釋安全性，因考量箱網養殖者家數過少，極容易被識別，故予以剔除，再以養繁殖總面積（箱網除外）、獨資漁戶經營管理者性別、年齡、教育程度及 104 年從事自家漁業工作日數、主要經營種類、漁業收入、漁撈漁船總艘數、從事自家漁業工作人數等 9 個關鍵變數計算唯一資料比率，其中得出唯一資料比率高達 98.36％，因關鍵變數中包含連續型變數，會導致唯一資料比率過高，在只考慮類別型關鍵變數情況下，唯一資料比率為 2.84％。

(二)類別型關鍵變數處理

由於未經處理的原始檔唯一資料過多，為了降低唯一資料的比率，對於部分連續型的關鍵變數進行分組處理；部分類別型的關鍵變數中分類數或人數太少者，則進行選項合併，除了漁撈漁船總艘數變數外，其他變數中未從事漁業者占 10.99％，其餘的選項詳細說明如下：

1. 經營管理者性別：男性占 78.05％；女性占 10.96％。
2. 經營管理者年齡：未滿 45 歲占 10.35％；45-54 歲占 21.67％；55-64 歲占 28.28％；65-74 歲占 18.28％；75 歲以上占 10.43％。
3. 經營管理者教育程度：不識字占 6.27％；小學及自修占 28.40％；國（初）中占 23.98％；高中（職）占 22.82％；大專及以上占 7.54％。

4. 經營管理者 104 年從事自家漁業工作日數：1-59 日占 20.47％；60-149 日分一類占 31.00％；150 日以上分一類占 37.54％。

5. 主要經營種類：遠洋漁業占 0.71％；近海漁業占 11.35％；沿岸漁業占 18.02％；內陸漁撈業占 0.83％；海面養殖業占 5.40％；內陸鹹水養殖業占 31.39 ％；淡水養殖業占 20.97 ％；轉型休閒占 0.35％。

6. 漁撈漁船總艘數：有漁撈漁船的獨資漁戶的擁有艘數不會太多，集中在 1 艘或是沒有漁撈漁船，因此分三類，無漁撈漁船占 67.94％； 1 艘漁船占 30.85％，2 艘以上占 1.21％。

7. 從事自家漁業工作人數：資料中，從事自家漁業工作人數最少 1 人，最多 13 人，因為從事自家漁業工作人數 3 人以上所占戶數不多，因此有工作人數者分三類，從事自家漁業工作人數 1 人，占 35.62％；從事自家漁業工作人數 2 人，占 33.05％ ；從事自家漁業工作人數 3 人以上，占 20.35％。

(三)連續型關鍵變數處理

另養繁殖總面積（箱網除外）與漁業收入兩個連續型關鍵變數和年底養繁殖總面積_1 到年底養繁殖總面積_20、自家初級漁產品銷售收入、自行加工漁產品銷售收入、委外加工漁產品銷售收入、休閒漁業服務收入等細項連續型變數，因為分布過於右偏，類別化方法將資料分組，可能使資

料代表性不足，在此未必為合適去識別化方法，因此考慮四捨五入法及衍生變數法，探討三種方法何者為最佳的去識別化方法。

1. 類別化：將上述變數先取對數後，扣除沒有值的戶數，再依等比率的區間分組。

(1) 養繁殖總面積（箱網除外）：無面積者占 33.96 %；1-39 公畝占 17.34 %；40-79 公畝占 15.53 %；80-149 公畝占 15.85 %；150 公畝以上占 17.33 %。

(2) 年底養繁殖總面積_1-年底養繁殖總面積_20：1-34 公畝占 24.57 %；35-69 公畝占 23.73 %；70-129 公畝占 25.92 %；130 公畝以上占 25.77 %。

(3) 漁業收入：未從事漁業無資料者，占 10.99 %；無收入者占 9.50 %，1-249 千元占 18.68 %；250-599 千元占 18.74 %；600-1,499 千元占 20.69%；1,500 千元以上占 21.40 %。

(4) 自家初級漁產品銷售收入：未從事漁業無資料者占 10.99 %；無收入者占 9.89 %，1-249 千元占 18.65 %；250-599 千元占 18.65 %；600-1,499 千元占 20.56 %；1,500 千元以上占 21.26 %。

(5) 自行加工漁產品銷售收入：因為此項收入大部分為 0，因此扣除未從事漁業者，分無、有兩類：無占 88.88 %；有占 0.14 %。

(6) 委外加工漁產品銷售收入：因為此項收入大部分為 0，因此扣除

未從事漁業者，分無、有兩類：無占 88.96 %；有占 0.05 %。

(7) 休閒漁業服務收入：因為此項收入大部分為 0，因此扣除未從事漁業者，分無、有兩類：無占 88.32 %；有占 0.69 %。

2. 四捨五入法：1-999 四捨五入至十位，1,000-9,999 四捨五入到百位，10,000-99,999 四捨五入到千位，以此類推。若四捨五入後為 0，則轉換成 1。原本為 0 的值，四捨五入後一樣為 0。由於資料過於右偏，為了使極端值較不易被辨認，將第 90 百分位以上的值以其中位數取代（第 90 百分位以上全部以第 95 百分位取代）。

3. 衍生變數法：將上述變數算各縣市占全臺灣的比率，新增變數為縣市總面積占比（或縣市總收入占比），其為一個類別化的區間，此區間依縣市個數大致相等分為以下區間，使其無法識別原始資料，如表 16。

表 16 獨資漁戶衍生變數法之類別化區間縣市個數

養繁殖總面積(箱網除外)類別化區間	未滿 0.4 %	0.4 % - 未滿 5 %	5 %以上
縣市個數	8	6	6
漁業收入類別化區間	未滿 0.5 %	0.5 % - 未滿 5 %	5 %以上
縣市個數	7	7	6

由於總比率已經模糊化，因此所有細項可以直接設定為占該縣市的比率。例如：第一筆資料在宜蘭縣，宜蘭縣養繁殖總面積（箱網除外）為 55,974 公畝，全臺灣養繁殖總面積（箱網除外）為 4,112,641 公畝，因此宜蘭縣養繁殖總面積（箱網除外）之縣市總面積占比為 0.4 %-未滿 5 %；宜蘭縣漁業收入為 3,752,205 千元，全臺灣漁業收入為 54,134,373 千元，因此宜蘭縣漁業收入之縣市總收入占比為 5 %以上。而所有細項皆是給占縣市的比率，如表 17 和表 18 所示，研究者使用時，宜蘭縣兩筆資料可以相互比較大小，若要全臺灣相比，則乘以縣市總面積占比（或縣市總收入占比）的組中點，即可用來比較。

表 17 獨資漁戶衍生變數說明(以養繁殖總面積(箱網除外)為例)

編號	縣市	養繁殖總面積(箱網除外)占該縣市總面積比率	縣市養繁殖面積占全臺總面積比率
00002741107017	宜蘭縣	0.14828 %	0.4 % -未滿 5 %
00004841107017	宜蘭縣	0.03037 %	0.4 % -未滿 5 %
00606741107017	新北市	0.35494 %	0.4 % -未滿 5 %

註:宜蘭縣總面積 55,974 公畝，新北市總面積 19,158 公畝，全臺灣總面積 4,112,641 公畝。

表 18 獨資漁戶衍生變數說明(以漁業收入為例)

編號	縣市	漁業收入占該縣市總收入比率	縣市總收入占全臺總收入比率
00002741107017	宜蘭縣	0.00213 %	5 %以上
00003941107017	宜蘭縣	0.10660 %	5 %以上
02117241107017	新北市	0.05061 %	0.5 % -未滿 5 %

註:宜蘭縣總收入 3,752,205 千元，新北市總收入 1,383,057 千元，全臺灣總收入 54,134,373 千元。

接著使用關聯分析來比較三種去識別化方法何者為最適合的方法。養繁殖總面積（箱網除外）和漁業收入在原始資料、四捨五入法及衍生變數法皆為非常態分布的連續型變數，因此使用 Kruskal-Wallis Test 來檢定不同組別間的中位數是否有差異；而類別化方法將養繁殖總面積（箱網除外）和漁業收入轉為類別型變數，因此使用卡方獨立性檢定分析其關聯。

由表 19 和表 20 可以看到，原始資料中，養繁殖總面積（箱網除外）及漁業收入和獨資漁戶經營管理者性別、年齡、教育程度與主要經營種類四個變數皆有顯著關聯。類別化方法及四捨五入法中，養繁殖總面積（箱網除外）及漁業收入與四個變數皆有顯著關聯；但四捨五入後養繁殖總面積與原始資料相同者占 63.68 %，年底養繁殖總面積_1 與原始資料一樣的占 64.63 %，漁業收入與原始資料相同者占 70.81 %，自家初級漁產品銷售收入與原始資料一樣的占 70.89 %，其餘細項面積及收入更是有 90 % 以上都未改變，因此不建議使用四捨五入法。衍生變數法在漁業收入與經營管理者年齡的關聯與原始資料不相同，其餘關聯則相同，另外在資料應用上由於較不直觀，研究者可能難以使用，因此不建議使用衍生變數法。綜合以上，類別化方法為建議使用的去識別化方法。

表 19 養繁殖總面積(箱網除外)去識別化方法關聯分析比較

		方法			
		原始資料(1)	類別化(2)	四捨五入法(3)	衍生變數法(4)
去識別變數	比較變數	P 值	P 值	P 值	P 值
養繁殖總面積（箱網除外）	性別	<0.0001***	<0.0001***	<0.0001***	<0.0001***
	年齡	<0.0001***	<0.0001***	<0.0001***	<0.0001***
	教育程度	<0.0001***	<0.0001***	<0.0001***	<0.0001***
	主要經營種類	<0.0001***	<0.0001***	<0.0001***	<0.0001***

註 1: (1)(3)(4)使用 Kruskal-Wallis Test；(2)使用卡方獨立性檢定。

註 2:*在 P 值為 0.05 顯著；**在 P 值為 0.01 顯著；***在 P 值為 0.001 顯著。

表 20 漁業收入去識別化方法關聯分析比較

		方法			
		原始資料(1)	類別化(2)	四捨五入法(3)	衍生變數法(4)
去識別變數	比較變數	P 值	P 值	P 值	P 值
漁業收入	性別	<0.0001***	<0.0001***	<0.0001***	<0.0001***
	年齡	<0.0001***	<0.0001***	<0.0001***	0.0908
	教育程度	<0.0001***	<0.0001***	<0.0001***	<0.0001***
	主要經營種類	<0.0001***	<0.0001***	<0.0001***	<0.0001***

註 1: (1)(3)(4)使用 Kruskal-Wallis Test；(2)使用卡方獨立性檢定。

註 2:*在 P 值為 0.05 顯著；**在 P 值為 0.01 顯著；***在 P 值為 0.001 顯著。

上述未提及的變數，面積的部分：年底面積_1 到年底面積_20 雖然不為重要變數，但仍有外釋風險，因此使用類別化法處理；而收入的部分：自家初級漁產品銷售收入、自行加工漁產品銷售收入、委外加工漁產品銷售收入和休閒漁業服務收入使用類別化法處理。其餘的變數會與原始資料相同，未做其他處理。

(四)風險分數評估

最終以類別化方法處理過後的資料計算兩種風險分數，發現大部份資料的風險分數相近，只有少數風險分數很高，如高於 99 百分位點的資料有較高的揭露風險，第一種風險分數第 99 百分位點的值為 0.2569，第二種風險分數第 99 百分位點的值為 144.3635。兩種風險分布的圖如圖 8 和圖 9。類別化方法處理後的唯一資料比率降為 37.46 %，而只考慮類別資料的唯一資料比率降為 1.38 %。因此，全檔關鍵變數既已處理，使得高風險資料之風險分數大幅降低，故不予以刪除。

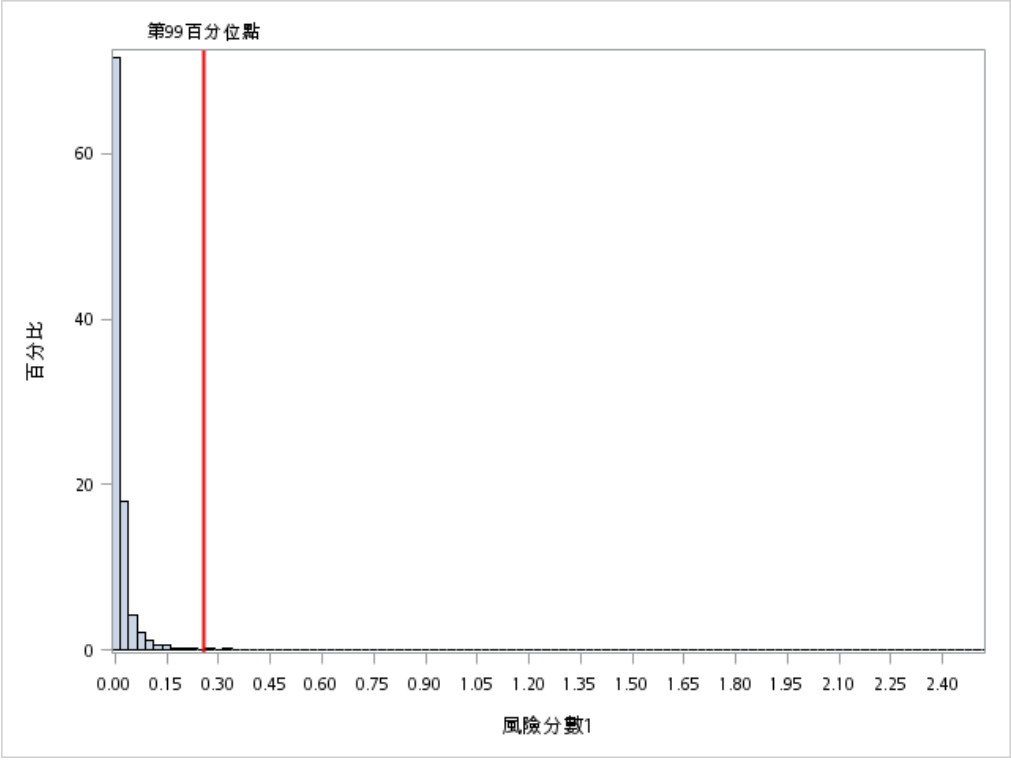

圖 8 獨資漁戶第一種風險分數

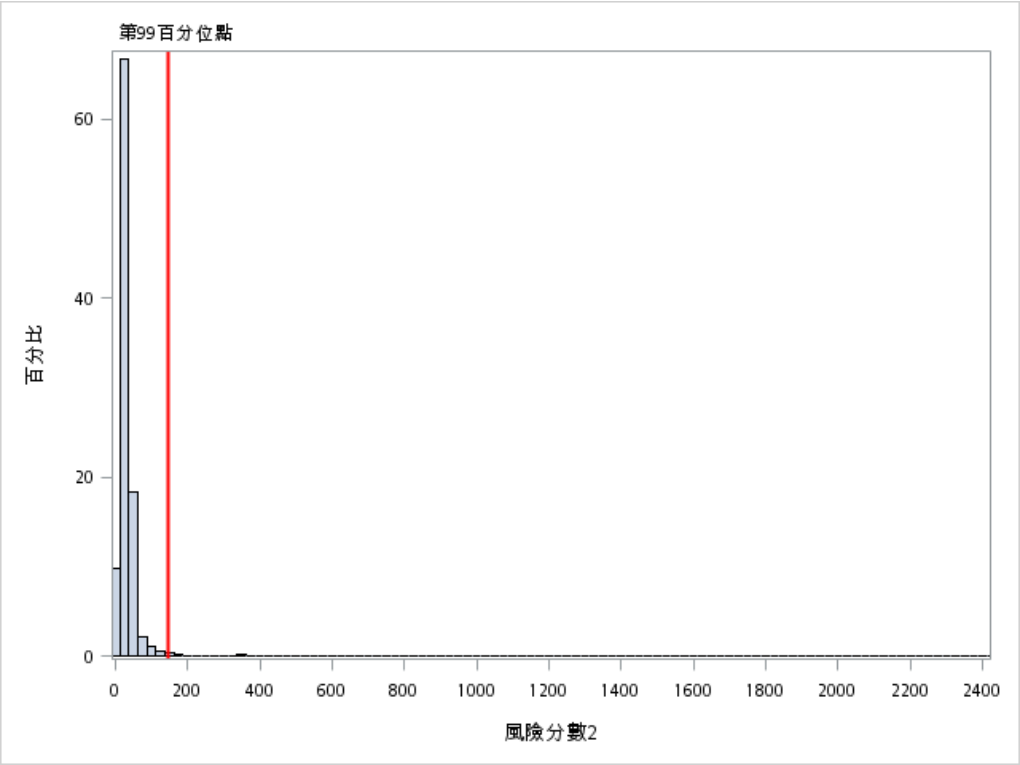

圖 9 獨資漁戶第二種風險分數

第二節 抽樣檔資料

一、 農牧戶

本次研究農牧戶抽樣檔希望能保有經營管理者特性，而未從事農牧業者在本研究中並沒有此部分資料，故只針對有從事農牧業者進行抽樣。

農牧戶考量抽樣分層變數為主要經營種類、可耕作地面積及農牧業收入，這些變數需推估至臺灣地區 20 個縣市母體總數及結構。原希望以鄉鎮市區為單位抽樣推計至縣市，以顯示各鄉鎮市區之間的差異，但大部分的主要經營種類所占比率過少，正常抽樣會抽不到這些種類的資料，為顧及此變數與全檔的一致性，因此選擇以縣市為單位抽樣，推計至全國。

由於部分主要經營種類占比過少，會有抽不到的問題，因此將主要經營種類做部分併組：稻作休耕及轉型休閒占 7.05 %；稻作占 29.88 %；雜糧占 8.22 %；特用作物占 4.71 %；蔬菜占 19.74 %；果樹占 25.81 %；食用菇蕈占 0.18 %；花卉占 0.71%；其他農作物占 1.85 %；豬占 0.72 %；雞占 0.58 %；其他畜牧業占 0.56 %。另由於可耕作地總面積及農牧業收入為連續型變數，因此除無面積者(或無收入者)外，將其等比率分為 5 組。可耕作地總面積：無面積者占 0.67 %；23 公畝以下占 21.96 %；24-36 公畝占 15.79 %；37-54 公畝占 22.25 %；55-93 公畝占 19.78 %；94 公畝以上占 19.55 %，農牧業收入：無收入者占 21.46 %；50 千元以下占 18.43 %；51-100 千元占 15.30 %；101-200 千元占 15.89 %；201-420 千元占 13.43 %；421 千元以上占 15.49 %，以便進行抽樣。

抽樣以主要經營種類、分組之可耕作地面積及分組之農牧業收入排序，以確保各類別都能被抽取到，全檔資料筆數為 717,958 筆，系統抽樣後得到 89,878 筆資料。

為避免分層變數底下各類別權重總和與原始資料不符，因此對分層變數進行權重的迭代計算，以使各類別權重總和與原始資料一致。(舉例：系統抽樣檔中主要經營種類為稻作的資料權重總和為 80，代表此類別底下有 80 戶，但原始資料中主要經營種類為稻作的資料有 100 戶，系統抽樣出的權重過少，因此將系統抽樣檔資料為稻作的權重都乘以 100/80，

以使此類別總權重與原始資料一致。）

最後從系統抽樣檔中依照分層變數比例，抽出 1 % (6,950 筆) 的檔案，由於系統抽樣檔大約為全檔的 12.52 % (89,878 / 717,958)，為了使 1 % 抽樣檔能推計至母體總數，因此各戶權重最後再乘以 12.52 為最終權重，各戶配適權重可估計母體總數及關鍵變數間之關聯。

為了確認抽樣檔的關鍵變數之分布是否與全檔相同，使用卡方適合度檢定檢視抽樣檔中所有關鍵變數與全檔的一致性，可以看到各檢定結果 P 值皆大於 0.05，表示抽樣檔的關鍵變數比率與全檔相符，如表 21。另外檢視農牧業收入與其他關鍵變數的關聯性，檢定結果皆與全檔關聯相同，如表 22 所示。

表 21 農牧戶抽樣檔變數卡方適合度檢定

變數	P 值
可耕作地面積	1.0000
性別	0.6421
年齡	0.8181
教育程度	0.4745
主要經營種類	0.9993
農牧業收入	0.9999
104 年全年從事自家農牧業工作日數	0.9498
從事自家農牧業工作人數	0.7833

表 22 農牧戶抽樣檔關聯分析與全檔之比較

分層變數	比較變數	全檔 P 值	抽樣檔 P 值
農牧業收入	性別	<0.0001***	<0.0001***
	年齡	<0.0001***	<0.0001***
	教育程度	<0.0001***	<0.0001***
	主要經營種類	<0.0001***	<0.0001***
	104 年全年從事自家農牧業工作日數	<0.0001***	<0.0001***
	從事自家農牧業工作人數	<0.0001***	<0.0001***

註:*在 P 值為 0.05 顯著；**在 P 值為 0.01 顯著；***在 P 值為 0.001 顯著。

二、 林戶

本次研究由於林場資料數過少，如正常抽樣會抽不到林場資料，如對其過度抽樣會使資料產生極大偏誤，因此本次研究只對林戶資料進行抽樣。

林業考量抽樣分層變數為主要經營種類及林業土地面積，這些變數需推估至全國母體總數及結構。因此以縣市為單位抽樣，以顯示各縣市之間的差異，並推計至全國。

由於林業土地面積為一連續型變數，若以連續型變數分層（將每個值視為一類別），其各類別之資料筆數過少，會使資料抽取偏誤，因此將其等比率分為 5 組，以確保面積大小的結構比例。24 公畝以下占 19.1％；25-49 公畝占 18.1％；50-99 公畝占 22.7％；100-199 公畝占 20.0％；200 公畝以上占 20.1％。

抽樣須以縣市為單位進行，因此先以縣市，主要經營種類及分組之林

業土地面積排序，以確保各類別都能被抽取到，全檔資料筆數為 87,152 筆，系統抽樣後得到 11,880 筆資料。

為避免分層變數底下各類別權重總和與原始資料不符，因此對分層變數進行權重的迭代計算，以使各類別權重總和與原始資料一致。

最後從系統抽樣檔中依照分層變數比例，抽出 1%的檔案（873 筆），由於系統抽樣檔大約為全檔的 13.6 %（11,880 / 87,152），為了使 1 % 抽樣檔能推計至母體總數，因此各戶權重最後再乘以 13.6 為最終權重，各戶配適權重可估計母體總數及關鍵變數間之關聯。

為了確認抽樣檔的關鍵變數之分布是否與全檔相同，使用卡方適合度檢定檢視抽樣檔中所有關鍵變數與全檔的一致性，可以看到各檢定結果 P 值，皆大於 0.05，表示抽樣檔的關鍵變數比率與全檔相符，如表 23。另外檢視林業土地面積與性別、年齡、教育程度及主要經營種類的關聯性，檢定結果皆與全檔關聯相同，如表 24 所示。

表 23 林戶抽樣檔變數卡方適合度檢定

變數	P 值
性別	0.3765
年齡	0.1259
教育程度	0.4155
主要經營種類	0.8788
林業土地面積	0.9979
主要所有權屬	0.8348
主要林相種類	0.3316
主要功能用途	0.1691
有無造林補助	0.5160
有無森林作業	0.7091

表 24 林戶抽樣檔關聯分析與全檔之比較

分層變數	比較變數	全檔 P 值	抽樣檔 P 值
林業土地面積	性別	<0.0001***	<0.0001***
	年齡	<0.0001***	<0.0001***
	教育程度	<0.0001***	<0.0001***
	主要經營種類	<0.0001***	<0.0001***

註:*在 P 值為 0.05 顯著；**在 P 值為 0.01 顯著；***在 P 值為 0.001 顯著。

三、 獨資漁戶

本次研究獨資漁戶抽樣檔希望能保有經營管理者特性，而未從事漁業者在本研究中並沒有此部分資料，故只針對有從事漁業者進行抽樣；另由於有箱網養殖家戶過少，如正常抽樣會抽不到有箱網養殖家戶的資料，如對其過度抽樣會使資料產生極大偏誤，因此本次研究只對沒有箱網養

殖的家戶進行抽樣。

獨資漁戶考量抽樣分層變數為主要經營種類、養繁殖總面積（箱網除外）及漁業收入，這些變數需推估至全國母體總數及結構。原希望以縣市為單位抽樣推計至全國，以顯示各縣市之間的差異，但主要經營種類中遠洋漁業、近海漁業、內陸漁撈業、海面養殖業及轉型休閒所占比率過少，正常抽樣會抽不到這些種類的資料，為顧及此變數與全檔的一致性，因此選擇以全國為單位抽樣，推計至全國。

抽樣以主要經營種類、養繁殖總面積（箱網除外）及漁業收入排序，以確保各類別都能被抽取到，全檔資料筆數為 38,783 筆，系統抽樣後得到 4,848 筆資料。

為避免分層變數底下各類別權重總和與原始資料不符，因此對分層變數進行權重的迭代計算，以使各類別權重總和與原始資料一致。

從系統抽樣檔中依照分層變數比例，抽出 1％（338 筆）的檔案，其中，主要經營種類中的轉型休閒並沒有抽到，會不符合一致性，因此決定外釋 5％（1,937 筆）的抽樣檔以增加代表性，由於系統抽樣檔大約為全檔的 12.5％（4,848/38,783），為了使 5％ 抽樣檔能推計至母體總數，因此各戶權重最後再乘以 2.5 為最終權重，各戶配適權重可估計母體總數及關鍵變數間之關聯。

為了確認抽樣檔的關鍵變數之分布是否與全檔相同，使用卡方適合度檢定檢視抽樣檔中所有關鍵變數與全檔的一致性，可以看到各檢定結果 P 值皆大於 0.05，表示抽樣檔的所有關鍵變數比率與全檔相符，如表 25。另外檢視漁業收入與其他關鍵變數的關聯性，檢定結果皆與全檔關聯相同，如表 26 所示。

表 25 獨資漁戶抽樣檔變數卡方適合度檢定

變數	P 值
養繁殖總面積(箱網除外)	0.9828
性別	0.8703
年齡	0.6226
教育程度	0.2557
主要經營種類	0.9979
漁業收入	0.9914
104 年從事自家漁業工作日數	0.3330
漁撈漁船總艘數	0.5733
從事自家漁業工作人數	0.9306

表 26 獨資漁戶抽樣檔關聯分析與全檔之比較

分層變數	比較變數	全檔 P 值	抽樣檔 P 值
漁業收入	養繁殖總面積(箱網除外)	<0.0001***	<0.0001***
	性別	<0.0001***	<0.0001***
	年齡	<0.0001***	<0.0001***
	教育程度	<0.0001***	<0.0001***
	主要經營種類	<0.0001***	<0.0001***
	104 年從事自家漁業工作日數	<0.0001***	<0.0001***
	漁撈漁船總艘數	<0.0001***	<0.0001***
	從事自家漁業工作人數	<0.0001***	<0.0001***

註:*在 P 值為 0.05 顯著；**在 P 值為 0.01 顯著；***在 P 值為 0.001 顯著。

第五章 結論與建議

第一節 結論

本計畫參考各國之外釋資料處理方法，提出適合我國之微觀數據建置方式。先針對農牧戶、林業、獨資漁戶三個業別分別產出全檔去識別化資料檔，再參考國際作法，建置普查外釋資料抽樣檔（微觀數據檔），以供各界研究使用。詳細說明如下：

一、 全檔

(一)農牧戶

農牧戶共有 775,258 筆資料，對於較容易被識別出的變數進行去識別化，類別型變數採用合併類別法，關鍵連續型變數採用四捨五入法，其餘連續型變數則採用合併類別法。處理後之全檔，提供 775,258 筆資料及 14 個普查項目。

(二)林業

林業共有 87,465 筆資料，其中 87,152 筆為林戶、313 筆為林場，由於此兩類資料合併處理會使資料產生偏誤，因此本次林業將兩類資料分開進行處理，並產出林戶及林場兩個全檔資料以供使用。

對於較容易被識別出的變數進行去識別化，類別型變數採用合併

類別法，關鍵連續型變數採用四捨五入法，其餘連續型變數則採用合併類別法。林戶經處理後之全檔，提供 87,152 筆資料及 9 個普查項目。林場經處理後之全檔，提供 313 筆資料及 9 個普查項目。

(三)獨資漁戶

本次獨資漁戶共有 43,587 筆資料，其中箱網養殖僅 17 家，極容易被識別，考量外釋安全性，全檔資料不外釋此類資料。

對於較容易被識別出的變數進行去識別化，類別型變數採用合併類別法，關鍵連續型變數採用類別化法，其餘連續型變數則採用合併類別法。獨資漁戶經處理後之全檔，提供 43,570 筆資料及 16 個普查項目。

二、 抽樣檔

為提升普查資料之應用價值，本計畫針對農牧戶、林業、獨資漁戶三個業別之全檔去識別化資料檔進行抽樣，依照安全性與使用方式，決定外釋抽樣檔比率，提供三個業別之微觀資料檔，以供各界研究使用。詳細說明如下：

(一)農牧戶

由於本次研究農牧戶抽樣檔希望能保有經營管理者特性，而未從事農牧業者在本研究中並沒有此部分資料，故抽樣檔只針對有從事農

牧業者進行抽樣。

依照美國 PUMS 建置微觀資料檔方法，對於農牧戶之全檔去識別化資料檔進行抽樣，並為各戶配適權重，使抽樣檔可推計至全檔總數及關鍵變數間之關聯，本次農牧戶抽樣檔提供 1 % 微觀資料檔，共 6,950 筆資料及 14 個普查項目。

(二)林戶

由於本次研究林場資料數過少，總面積占比極高，如正常抽樣會抽不到林場資料，如對其過度抽樣會使資料產生極大偏誤，因此本次研究只對林戶資料進行抽樣。

依照美國 PUMS 建置微觀資料檔方法，對於林戶之全檔去識別化資料檔進行抽樣，並為各戶配適權重，使抽樣檔可推計至全檔總數及關鍵變數間之關聯，本次林戶抽樣檔提供 1 % 微觀資料檔，共 873 筆資料及 9 個普查項目。

(三)獨資漁戶

由於本次研究獨資漁戶抽樣檔希望能保有經營管理者特性，而未從事漁業者在本研究中並沒有此部分資料，故只針對有從事漁業者進行抽樣；另由於有箱網養殖家戶過少，如正常抽樣會抽不到有箱網養殖家戶的資料，如對其過度抽樣會使資料產生極大偏誤，因此本次研究只對沒有箱網養殖的家戶進行抽樣。

依照美國 PUMS 建置微觀資料檔方法，對於獨資漁戶之全檔去識別化資料檔進行抽樣，並為各戶配適權重，使抽樣檔可推計至全檔總數及關鍵變數間之關聯，本次獨資漁戶抽樣檔提供 5% 微觀資料檔，共 1,937 筆資料及 16 個普查項目。

由於農牧戶原本外釋資料提供的最小區域單位為鄉鎮，而其他兩個業別的最小外釋區域單位為縣市；但抽樣檔若要可以維持關鍵變數間的關聯，最小區域單位需要調整。雖本研究所建置的抽樣檔之分層變數較美國 PUMS 少，但考慮最小區域單位的範圍較廣，資料異質性大，最後所採用的抽樣比例仍舊維持 PUMS 的設定。假設未來最小區域單位的範圍較小（如村里），而分層變數為 3 個時，則依照分析一個變數約需要 30 個觀察值的原則，需要至少抽樣 90 戶，則抽樣比例可設定為區域戶數少於 400 戶，抽樣率設定為 1:2、區域戶數介於 400 戶到 600 戶，抽樣率設定為 1:4、區域戶數介於 600 戶到 1,000 戶，抽樣率設定為 1:6，區域戶數超過 1,000 戶，則抽樣率設定為 1:8。

第二節 未來研究與建議

農林漁牧業普查是政府掌握農林漁牧業者經營現況的依據，進行普查時，相關業者一定要提供資料，政府則有義務確保業者所提供的資料之

安全，如何適當地將資料外釋又可保有資料的實用價值，讓研究者做農業相關研究以提供政府部門政策建議，是相當重要的。

為了保護資料安全性，本計畫在外釋資料需進行去識別化，在農牧戶、林業及獨資漁戶考量之關鍵變數分別高達 8 個、10 個及 9 個，考量間接識別變數相當嚴謹，但也因資料去識別化愈高，其可提供研究的資訊愈少，如何拿捏去識別化的程度，還需要搭配資料敏感性之認定。再者，因科技進步使資料蒐集變的相對容易，若僅考慮單一份資料去識別化，可能不夠嚴謹，未來仍需要進一步考慮到使用者可能利用多檔串接的方式，而間接識別個資的問題。

對於個資的保護歐盟一直不遺餘力，於 1995 年提出個人資料保護指令（Data Protection Directive），並於 2016 年通過「一般資料保護規則」（General Data Protection Regulation；簡稱 GDPR）來取代原本的法案，經過兩年的緩衝期後，已於 2018 年 5 月 25 日生效並全面施行。因此，歐盟外釋資料大多採用數據模擬，採議題式運用統計模擬，產生資料提供外界使用。這類的外釋資料安全性高，但研究的範圍會被限縮。若未來外釋資料是有分特定主題，且個人資料有隱私的考量，則可以參考歐盟採用數據模擬外釋資料。

本研究建置的抽樣檔包含所有全檔的變數，並針對分層關鍵變數分別設定權重，提供研究者回推全檔的估計值，回推的數據可能與全檔的數

據有些微差異。研究者可以先利用抽樣檔進行資料探索，找出合適的研究主題，之後可以再申請全檔資料。本次外釋資料的最小區域的資訊為鄉鎮或縣市，未來可以參考美國，建置兩種格式的抽樣檔，一種針對家戶，提供詳細家戶的資料，另一種則提供較詳細的區域資料。

除進行資料去識別化外，可同時參考國外採取不同程度外釋資料的方式，即外釋變數少且去識別化程度大的資料可提供所有人申請，而變數愈多且去識別程度較少的資料為限制性申請，而若需要全檔者或與其他資料庫串連的資料，則須使用雲端登入或進入資料管制室方可分析資料，如此才可能在保護個資的條件下，讓資料有效的運用，讓更多研究者投入研究，使政府決策更有依據。

第三節　資料使用注意事項

一、 代表性

目前三個業別全檔與抽樣檔的關鍵變數與分層變數分配均有通過一致性，且全檔與抽樣檔關鍵變數與分層變數的關聯有一致。但本計畫為避免揭露高端個資資料，將高端的收入與土地面積以高端的平均值取代，該部分雖有利用權重進行修正，但該修正僅可以讓農牧戶與林戶的抽樣資料回推至縣市，獨資漁戶的資料回推至全國，無法保證低一等級的區域抽樣檔的估計值會與全檔一致。再者，因部分業別的變數具有稀少性(如加

工收入)，為避免個資揭露，這些變數僅以有無註記提供資料，故無法回推至全檔。

二、 研究限制

由於外釋資料需經過去識別化處理，會使變數間喪失部分關聯，以下根據外釋檔案討論研究限制：

(一)三個業別之全檔，關鍵變數間之關聯皆與母體相符；而未討論之變數，研究上使用可能與母體變數關聯不一致。

(二)三個業別之抽樣檔，關鍵變數分布皆與全檔相符；而未討論之變數，研究上使用可能與全檔變數關聯不一致，故僅能供做探索性分析，如果要做更深入之研究，建議使用者申請全檔資料。

參考文獻

1. Akkerman, A. (1980). On the relationship between household composition and population age distribution. Population Studies, 34, 525-534.
2. Alfons, A., Kraft, S., Templ, M., & Filzmoser, P. (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. Statistical Methods & Applications, 20(3), 383-407.
3. Butrica, B. A., & Iams, H. M. (2000). Divorced women at retirement: Projections of ecoonomic well-being in the near future. Soc. Sec. Bull., 63, 3.
4. Favreault, M. M. (2002). The impact of Social Security reform on low-income and older women. Project report. AARP Public Policy Institute, 1-46.
5. Fraser, B., & Wooton, J. (2005). A proposed method for confidentialising tabular output to protect against differencing. Monographs of Official Statistics. Work session on Statistical Data Confidentiality, 299-302.
6. General Social Survey, Cycle 27: Giving, Volunteering and Participating (2013). Public Use Microdata File, Documentation and User's Guide.
7. Ito, S., & Hoshino, N. (2014). Data swapping as a more efficient tool to create anonymized census microdata in Japan. In Privacy in Statistical Databases, 1-14.
8. Ito, S., Hoshino, N., Akutsu, F. (2015) “A Quantitative Assessment of Data Confidentiality and Data Utility to Create Anonymized Census Microdata in Japan”, Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Helsinki, Finland, pp. 1-14.
9. Klevmarken, A., & Lindgren, B. (Eds.). (2008). Simulating an ageing

population: a microsimulation approach applied to Sweden. Emerald Group Publishing Limited.

10. Li, Y. (2004). Samples of Anonymised records (SARs) from the UK censuses: a unique source for social science research, Sociology, 38 (3), 553-572.
11. Münnich, R., & Schürle, J. (2003). On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series 4. https://pdfs.semanticscholar.org/33e1/3b52fb85bd5c100ad2388946564f8056590b.pdf?_ga=2.63832105.746047764.1537516381-2022527448.1537516381
12. Nadeau, S. E., Wu, S. S., Dobkin, B. H., Azen, S. P., Rose, D. K., Tilson, J. K., etal. & LEAPS Investigative Team. (2013). Effects of task-specific and impairment-based training compared with usual care on functional walking ability after inpatient stroke rehabilitation: LEAPS Trial. Neurorehabilitation and neural repair, 27(4), 370-380.
13. Namazi-Rad, M. R., Mokhtarian, P., & Perez, P. (2014). Generating a dynamic synthetic population–using an age-structured two-sex model for household dynamics. PloS one, 9(4), e94761.
14. National Household survey public use microdata file (2011). Hierarchical file Documentation and user guide.
15. Sabelhaus, J., & Topoleski, J. (2007). Uncertain policy for an uncertain world: The case of social security. Journal of Policy Analysis and Management, 26(3), 507-525.
16. Shlomo, N., Tudor, C. and Groom, P. (2010). Data swapping for protecting census tables. Domingo-Ferrer, J. and Magkos, E. (eds) Privacy in

Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2010 Corfu, Greece, September, 2010 Proceedings, Springer, pp.41-51.

17. Stephan, F.F. (1942). An iterative method of adjusting frequency tables when expected marginal totals are known. The Annals of Mathematical Statistics, 13. 166-178.
18. Sundberg, O. (2007). Model 5: SESIM (longitudinal dynamic microsimulation model). In modelling our future: Population ageing, health and aged careEmerald Group Publishing Limited, 453-460.
19. 陳惠欣、周怡伶(2014)。我國農林漁牧業普查之推展與應用。調查研究-方法與應用，第 32 期，159-185。

附件 期末報告審查意見表

審查意見	處理情形
美國人口普查長問卷 1%及 5%抽樣檔，二者並無相關，報告所列抽樣模式(圖 1)應予修正及說明。	感謝委員建議，已參照委員建議修正。
英國國家統計局採用 5 種方法來預防個別資料外洩，請說明其優、缺點及適用情形。	感謝委員建議，已對此增加說明。
國外人口普查外釋資料，有釋出比率高則釋出變數較多之情形，請增加原因說明。	感謝委員建議，已對此增加說明。
農林漁牧業普查資料依風險分數評估建議刪除者，於建置全檔及抽樣檔時並未刪除，請補述原因。	感謝委員建議，已補述原因。
依本研究各縣市農牧戶抽樣檔之抽樣比率除基隆市外均相同，請增列比率設定之原則或說明。	感謝委員建議，已對此增加說明。
本研究所建置之全檔及抽樣檔，因可耕作地面積、作物、畜禽及戶內人口等為逐筆填寫資料，其變數個數重複計算，請改以普查項目數呈現，以避免有筆數少、變數太多之疑慮。	感謝委員建議，已參照委員建議修正。
農牧戶抽樣檔之關鍵變數中，經檢定後教育程度與全檔並不一致，請依變數特性再予處理。	感謝委員建議，修正後所有關鍵變數與全檔皆一致。
於安全前提下，各國外釋資料有不同等級之提供方式，請補述農林漁牧業普查外釋資料分類及適用原則。另請於報告及技術文件(Technical Document)中，清楚說明外釋資料之使用方法、限制及代表性。	感謝委員建議，已對此增加說明。
請提供本研究全檔及抽樣檔相關資料分析、抽樣及程式等技術文件，俾利後續應用。	感謝委員建議，將提供技術文件供後續應用。