社會科學的「大數據」

連賢明*

近二十年來經濟學研究興起了「大數據」熱，越來越多經濟學者透過行政資料來進行政策分析，以圖解決調查資料無法克服的困難。最典型的兩個例子就是所得分配以及階層流動的相關研究。過往的所得分配相關研究，大多數都透過政府的調查資料進行分析。這些調查資料雖有其價值，但存在一個致命的局限：調查資料無法涵蓋超高所得家戶。這些超高所得家戶數目雖然不多，僅占家戶數目的 1% 或 0.1%，卻在整體所得占有巨大份額。結合美國歷年報稅資料和國民所得帳資料，Emmanuel Saez 和 Thomas Piketty 分析最有錢 1% 或 0.1% 的所得比例，發現超級有錢人的所得份額自 1980 年後的 10% 逐漸上升，成長到 1990 年的 15%，繼續成長到金融風暴（2008 年）到達 20% 的高點；相較之下層 50% 的人口所占所得比例卻穩定下降，從 1980 年的 20% 下滑到 2008 年的 14%。而在金融風暴後，所得前 1% 的所得占比很快回穩，但下層半數人口占比仍繼續下降達到 12% 的歷史新低，這些 50% 的人的份額竟然不到高所得 1% 的份額一半！這個超級有錢人所得占比越來越高趨勢，不但在美國出現也存在歐洲、日本、甚至臺灣。這些所得越來越不均的數據，也成為 Thomas Piketty 在《二十一世紀資本論》有錢人財富累積較快速的重要證據。

相較於所得分配的不均，不少學者則認為社會階層是否流動才是關鍵。理由是所得不均僅是一個靜態的描述，階層流動才能反映動態的過程。社會階層若具備相當流動性，即使存在相當貧富差距，還是能提供不同階層往上爬的機會，符合機會均等的公平要求。個人去年在史丹佛進修，恰巧結識一位研究階層流動這議題的翹楚：Raj Chetty 教授。Chetty 教授是印度裔的經濟學家，現年雖不到四十歲，已經是美國國家科學院院士、美國文理科學院院士、拿到經濟學界的克拉克獎（給四十歲以下最傑出經濟學家，俗稱小諾貝爾獎），據說也是研究結果最廣泛被媒體報導的經濟學家。他許多研究是使用不同行政資料分析在跨世代、地區、以及種族下的社會不公平和機會均等。舉例來說，他結合過

* 國立政治大學財政學系教授

去數十年的戶口普查資料和稅務資料，發現美國不同世代所得水準超越他們父母世代的機率顯著不同。對 50 年出生世代，所得水準超越父母世代機率超過八成，可說是充滿希望的一個世代；但這個機率隨著時間降低，到了千禧世代（1980 年出生的）的機率則幾乎不到一半，這解釋了為什麼相對其他世代，千禧世代有更多的不安全感，也對社會存在更多憤怒。

Chetty 教授也使用大數據來分析不同地區的跨世代所得階層流動。他使用美國的歷年稅務資料連結父母和小孩的所得。透過稅務資料所申報所得，他將父母和小孩所得各分成五等分，研究不同所得家庭小孩翻身成功的機率。根據他的發現，美國家庭所得最低 20% 的子女，能夠成功翻轉為最高 20% 所得階層機率只有不到 8%，遠低於加拿大的 13.5%；而令人驚訝的是，最能夠翻轉的小孩成長地區不是如同大家想像的都會區，反而是許多具有濃厚宗教信仰的中級都市（如鹽湖城），後續研究也肯定小孩成長地區的確對將來所得具關鍵影響，小孩越早從不利的居住環境遷出，將來所得改善的機率也就越高。這些結果對政府執行像社會住宅來改善居住環境的政策，提供一個選擇成效較高的居住城市名單。

筆者曾和 Chetty 教授聊到美國政府對這類數據整合的支持程度，他同意這些資料使用的確比較敏感，但許多國家都知道其價值所以在推展資料整合和數據分析；更重要的是這些研究成果能造成反饋，有效改善政府政策的品質。這也是美國對行政資料（稅務、普查、醫療和教育等）串接越來越開放理由。更何況資料保密的技術有長足進步，許多敏感性資料可透過將個體資料加權平均後提供（如同一里的所得資料），或是將資料內容轉換為分組後提供（如所得分組後提供）；即便是使用個人資料，目前要求都在資料中心進行操作，電腦可以詳細記錄所有操作細節，事後也能追蹤使用者是否違反規定，能有效保障個人隱私。

相較於國外研究，臺灣其實相當具備透過大數據討論政策研究的潛力。臺灣從日據時代開始就保存許多完整資料。而在政府行政資料上，臺灣不但有長達二十年的健保資料、社會保險、交通安全、高等教育和稅務資料，甚至更長的農業土壤和氣象水文數據，這些不同數據資料庫若串聯後，能大幅加值來作為臺灣發展大數據的利基，也可提供未來人工智慧判讀的基礎。可惜的是受限於個資使用的疑慮，臺灣政府目前除健保資料外，很少願意開放其他行政資料供學界使用和整合，遑論提供民間機構來利用這些數據。倘若主管機關能協助解決行政資料使用問題，不但能對臺灣從事社會科學研究大有助益，更能夠使用研究成果改善政策品質，提高政策效益。