論述 》專論 · 評述

統計學在大數據時代的角色

在谷歌的流感趨勢（Google Flu Trend）發表之後，大數據逐漸成為新的潮流，一時之間大數據似乎無所不在。隨著大數據的廣泛運用，大數據的一些盲點也逐漸出現。本文從統計學的觀點出發，討論大數據分析的優劣，以統計模式修正谷歌流感趨勢的案例，點出統計方法在大數據分析中可以扮演的角色，希望在這個大數據的浪潮中，統計方法能帶領我們避開荊棘，從大數據分析中真正獲益。

王鴻龍（國立臺北大學統計學系副教授）

壹、前言

Ginsberg J et al.（2009）利用 Google 搜尋引擎的關鍵字資料建構一個流感預測模式，在科學期刊（Science）上發表。這個預測模式，現稱爲谷歌流感趨勢（google flu trend，縮記爲 GFT）於 2010 年的及時預測結果正確性很高，且比美國疾病管制局（CDC）的通報系統快了 1-3 週。這樣令人振奮的結果，帶動了各界對大數據的重視。

業界的反應最快，大公司紛紛思考如何利用大數據開發商機，較早起步的公司甚至建構雲端資料庫及雲端處理系統，打造了雲端服務商機，例如 Amazon web service。大中華地區也不遑多讓，鴻海與阿里巴巴合作建構網路銷售資料庫，爲小企業提供相關產業大數據分析的服務。在政府的鼓勵下，臺灣公部門積極整合數據、開放數據，並鼓勵各界使用大數據開發研究議題，郭昌儒（2015a, b）、饒志堅（2015）、謝邦昌、謝邦彥（2015）、丘昌泰（2016）等撰文分享或呼籲大數據在政府部門的使用。業界也紛紛向大數據靠攏，建構雲端資料庫，舉辦大數據分析競賽培養人才，大數據分析儼然成爲顯學。各類的資料分析活動、研討會，只要冠上「大數據」，就如同桂冠加持，立即大受歡迎。然而大部分的資料分析活動只是傳統的資料分析軟體，或是較新的資料探勘（data mining）等已經發展十數年資料分析工具的產物罷了，似乎都不具備大數據的 3V 特質的任一項。

從資料分析因此備受重

視的角度，我們也都正面看待這樣的發展。但是我們更關心的是如何在眞正的大數據分析時代中受益，要如何準備才能迎接大數據時代來臨。另外我們也關心大數據發展是否影響到目前我們的決策行爲，沒有大數據是否落伍了。然而大數據分析也有一些分析錯誤的情形，例如，GFT 的預測也曾經歷了多次的模式修正，使用大數據分析結果做決策又是否太躁進？

貳、大數據分析和傳統的資料分析差異

根據維基百科（Wikipedia）的定義，大數據（Big Data）多具備 3V 的特質，包括：

(1) 數量 Volume：資料規模大到無法用傳統資料分析工具處理的數據。

(2) 速度 Velocity：資料來源具即時性，自動連續的收集。

(3) 多樣 Variety：資料來源、格式多樣，有數字、文字、文件、聲音、影像等。

另外又被加入了變化性（Variability）、眞實性（Veracity）成了 5V。一般只要具備 3V 中的數量大，或是非傳統資料型態的資料，都被歸類爲大數據。

據有大數據特質的資料分析，由於資料量大、收集累積速度快，或是資料格式結構雜亂，只要資料具備上述的要件之一，傳統資料分析工具已不敷使用。因應大數據分析的需求，新的工具也陸續被開發出來，包括資料庫系統、演算法等等，且暫時通稱爲大數據分析工具。

數據要多大，傳統的分析工具才會受到影響？一般多透過程式語言或套裝軟體來分析資料，而程式語言或套裝軟體只能分析記憶體（RAM）容量內的資料，部分程式語言套件可以用多工運算或虛擬記憶體方式增加可運算的資料量。陳景祥（2015）以 R 所用的記憶體估算，10 個變數 1 億筆資料約 7.6GB，也相當於 100 個變數 1 千萬筆資料的量。在高階單機上或許還可以應付這樣的資料量，不過如果有圖片或是影片檔案就會很快的超過單機的容量。

巨量資料分析除了量大還有資料多源的特性。資料的存取常透過不同的資料庫進行，傳統的 SQL（Structured Query Language）語言已不敷使用。在多源資料下，不須預設資料結構的 NoSQL 因應而起。另外傳統的資料分析都以歷史資料進行分析，能因應源源不斷資料收集（例如網路資料、高速公路 e-Tag 資料等）的快速分析，也都是大數據分析工具需解決的課題。

大數據分析工具的開發，Google 的貢獻不可忽視。爲了解決快速準確搜尋結果，Google 開發了支援海量結構化管理的平台 BigTable，包括平行計算程式設計模式（MapReduce）、分散式檔案系統（Google File System）等核心技術。Hadoop 專案所開發的平台，包括了 MapReduce 平行運算模組及 HDFS（Hadoop Distributed File System）分散式檔案系統。由於 Hadoop 開放免費的原始碼，現在已成爲

最廣泛使用的大數據分析工具之一。資策會也定期開班教授 Hadoop 平台技術。

大型的程式語言系統 matlab 及套裝軟體 SPSS、SAS 也都開發了因應大數據分析的模組。這些新產品，對於剛跳脫傳統資料分析的使用者而言，應該受惠不少，隨著資料量的增加，資料結構的多源性的增加，這些產品的大數據處理能量還有待觀察。

透過大數據分析工具，我們可以受惠於大數據所帶來前所未有的分析成果。大數據到底比傳統的資料分析好在哪裡，我認爲最少有三方面可以觀察：

(1) 資料數量大，幾乎等於母體，可以看到傳統抽樣調查看不到的小地方。

(2) 資料累積快速，可以即時呈現各種機會。

(3) 彙整非傳統的資料來源（例如網路、多媒體、文字等），可以看到前所未見的資訊。

在量大幾乎等於母體的部分，Mayer-Schonberger 等（2012）在大數據（Big Data）書中舉了一個很傳神的例子。作者以日本相撲大賽 11 年，超過 6 萬 4 千場的相撲比賽紀錄爲例，這個比賽紀錄資料雖然不到所謂大數據的等級，不過幾乎涵蓋了期間所有的賽事，分析發生在季末、不爲人注意的場次。發現當甲方爲 7 勝 7 負，乙方爲 8 勝 6 負時，甲方獲勝的情形比平常高出約 25%，比對後續的比賽，下次對戰時乙方的獲勝機率大增。進一步探討原因發現，原來相撲選手必須在每賽季的 15 場比賽中取得過半勝場，才能保留下一年度的級別和收入。因此在季末無關晉級時，選手們都樂於做個順水人情讓對手有機會保留級別和收入，而受惠選手則多在下次對戰時回報。這樣的場次並不太多，相對於 6 萬 4 千多的場次，可能不到 1%，沒有透過母體資料的分析，是不太可能呈現這種比賽可能作假的數據。

在非傳統資料來源的部分，陳昇瑋（2015）以計算社會學（computational social science）的角度，探討 NGO 分析透過慈善募款影片成功案例，找到募款成功的因素，作爲日後募款影片製作的參考，有效的提升了募款的額度。另外由於通訊科技的進步，人與人之間的互動只要透過網路或通訊設備進行，都會留下紀錄；這些紀錄下來的非傳統資料，多是人自然而然地做個平常自己做的事的眞實紀錄。沒有調查誤差也沒有抽樣誤差，社會科學研究者只需從旁被動的蒐集資料，透過適當的大數據資料分析，例如，手機通聯觀察人際關係、推特訊息研究所透露的情緒，更能觀察到過去無法想像的人際互動關係。

大數據分析結果如何解讀，成爲本文接下來的探討重點：

(1) 大數據所代表的母體是哪一個母體，是原先預期的母體嗎？

(2) 大數據分析結果準確嗎？錯了怎麼辦？影響有多大？

(3) 非傳統資訊的量化是否正確？誤判的後果爲何？

解答似乎就在統計的理論中。

參、統計方法可以扮演的角色

一、大數據資料到底是在說哪一個母體，分析結果到底要推論到哪一個族群？

GFT 使用全美國數十億則網路搜尋字眼，建構了一個流感發生數預測模式，由於 Google 使用的搜尋字眼的網路，涵蓋了幾乎大部分人口密集的地區，由於流感也都發生在人口集中的地區，GFT 預測模式的資料涵蓋母體與流感發生地區母體幾乎相同，加上網路搜尋字眼可以即時取得，GFT 的預測模式的即時性（比美國疾病管中心（Centers for Disease Control, 縮記爲 CDC）的通報數快將近 1-3 週），也因此聲名大噪。然而在網路流通不夠，或是入口網站涵蓋地區不足的地方，GFT 建構模式的方式就未必可行。

郭昌儒（2015c）使用 e-Tag 的大數據資料，分析高速公路車流行爲。郭昌儒分析的資料，包括了自 2013 年底到 2015 六月底所有的 e-Tag 行車資料共約 85 億筆資料，這些資料記錄了所有行駛一、二高的車輛，進出的交流道地點及時間。e-Tag 的超高辨識系統，成就了這些資料就是收集資料期間進出一、二高的母體資料。包括各時段車流量、各路段在特定路段的需求等的分析結果，對於用路資訊必然極具參考性。然而一、二高以外的道路使用情形，如東西向高速公路，各交流道的聯絡道路等，並未在 e-Tag 的紀錄範圍。一、二高周邊的交通狀況就無法據此大數據分析做出完整的推論。

健保資料庫蒐集了 1995 年至今近 20 年民眾就診的健保給付相關資料，僅在住院醫療費用、醫令、門診處方及治療、醫令等紀錄檔案，截至 2013 年底已累積 360 億筆資料約 5.6TB 的資料量，且每年以 48 億筆約 1TB 的資料量增加中，堪稱是大數據級的資料，對於健保相關的研究具有極高的價值。不過資料庫並未涵蓋非健保給付的醫療行爲，也就沒有足夠的資料來分析非健保項目的需求，例如瘦身醫美的市場需求。業界常希望透過網路資訊取得商機，不過由於各入口網站的涵蓋區域未必完整，且入口網站的資訊也未必願意完整釋出，使用的網路耙文工具，耙出的資訊所代表的族群就值得討論了。業界的目的如果只是爲了取得商機，那麼只要找到足夠的資訊分析出夠多的潛在顧客族群，這些方法確實可以讓業界開闊市場。如果政府要做產業趨勢分析以作爲政策制定的參考，那麼網路耙文的資訊涵蓋性就值得進一步探討。選舉時使用網路資訊打選戰也是一樣的道理，只要找到足夠的支持族群，透過無所不在的媒體傳播，選戰效果就會出現。但是如果要作完整策略制定，那麼大數據分析的資料所代表的族群，就應該審慎檢視確認所做的推論所能涵蓋的族群。這個部分在統計的抽樣調查理論就有完整的討論，而大據數的差別是在，大數據的資料是涵蓋母體的所有資料，抽樣調

論述 》專論 · 評述

查是先確定推論母體再抽出資料作分析，大數據則是先有資料，我們只需確認資料是否涵蓋母體，就可以做出適當的推論。

二、大數據分析是否準確，正確性有多高？錯誤的風險有多大？

引爆大數據風暴的 GFT 預測模式建構，除了運用了大量的網路關鍵字，也使用了統計相關性的假設，並比較了上億個數學模式，而找到了最終的預測模式。Google 的分析團隊使用每天收集到超過 30 億筆的搜尋紀錄，透過美國人最常用的前五千萬個搜尋字眼，再比對 CDC 在 2003 ～ 2008 年間的流感傳播資料。團隊先做了流感關鍵字搜尋與流感感染有關的假設，透過搜尋字眼的搜尋頻率與流感傳播的時間、地區的統計相關性（correlation），總共用了 4 億 5 千萬種不同的數學模式，成功地找到可以準確預測流感病例數的數學模式。這個模式使用了 45 個搜尋字眼，預測的結果與 CDC 公布的 2007 ～ 2008 的病例數十分吻合。預測模式在 2009 年 H1N1 危機時，發揮了即時準確的預測功能，比疾管局掌握的資料更快（快了大約 2 週以上）。讓公衛當局在疫情的控制上更爲快速有效。然而這個模式在 2009 年底與 2010 年初的預測值偏低，Google 團隊依照類似程序，修正關鍵字與數學預測模式。修正後的 GFT 模式在 2011 年準確預測，不過這個模式在 2012 年時卻發生了預測值偏高的情形。GFT 又需再次修正。這麼大的大數據分析工程所建構的預測模式，曾發生兩次的嚴重預測錯誤，且預測錯誤的方向不同。這也凸顯了大數據分析上的缺陷－大數據分析似乎缺少了風險評估機制。我們不知道大數據分析正確性有多少，使用這樣的大數據分析的結果做後續的決策風險爲何？

統計理論也提供推論，包括估計與檢定，而這兩大推論都有分析結果正確性與錯誤率的評估。使用統計推論結果做決策的風險，可以事先推算並加以控制。多年的實務驗證也呈現，統計推論的結果未必每次都正確，不過長期下來，錯誤率是可以在控制範圍內。

統計模式結合大數據分析也有成功的案例，寇星昌（2015）、Yang, etc.（2015）將谷歌流感趨勢的方法加入 CDC 的資料，以時間數列分析方法建構新的預測模式，稱爲 ARGO（Auto-Regression with Google search data）流感模式

$$y_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t} + \varepsilon_t$$

其中 y_t = logit（CDC reported flu activity at time t）

$X_{i,t}$ = log（Google search frequency of term i at time t + 0.5）

ε_t 爲隨機常態誤差。

模式的前段以 CDC 的歷史通報數建構了時間數列的 AR（Auto-regression）模式，後段則加入了谷歌搜尋資料，ARGO 流感模式的預測結果比 GFT 2014 版預測模式更準確。以時間數列方法結合 GFT 的模式建構了一個夠準確的模式，也提供了預測模式的錯誤估計。

三、非傳統的資訊（網路、多媒體、文字等），提供了更多元的數據來源，雖然大大的增加了視野，但也引起量化是否正確，錯誤風險有多大的問題。

透過非傳統的資訊，社會科學的研究邁入了新的境界。Onnela（2007）透過歐洲占有某國五分之一市場的行動通訊商，取得四個月內數百萬人的所有通話紀錄，建構了人際聯絡網。研究這些人際網絡中形成的社群網絡，有個重大的發現。將社群中連結眾多的人移除，剩下的社群網絡緊密度降低，但並不會整個崩潰；如果移除的是與社群之外有聯結的人，社群反而會突然崩解。這樣的觀察結果顯示，任何團體或整個社會來說，多樣性都至關緊要。而這樣的研究沒有大量、長時間、局部完整（1/5 市占率的行動通訊商）的通話紀錄是無法發現這樣的現象。然而從通話紀錄建立人際聯絡網絡的過程，如果有分析的誤差，會不會產生完全不同的結果也值得關注。

在非傳統資料來源部分，前述的 NGO 募款案例，是透過成功募款案例的須幫助家庭的影片量化數據，例如影片人物的年齡、親子關係、拍攝場景、家庭狀況等，找到影響募款的因素，作爲日後募款影片製作的參考，分析結果也確實有效提升了募款額度。然而從影片量化爲數據，測量誤差似乎沒有進一步討論過，是否會有更重要的因素在量化過程中被忽略了，也無法確認。網路資料上常用的文字探勘 （text mining） 也有量化測度誤差的問題。透過非傳統的資訊，社會科學的研究確實邁入了新的境界。如果能夠在非傳統資訊量化過程中多一些測度的討論，那麼分析結果的可靠性應該可以大幅提升。

肆、結語

二十世紀初，卡爾·皮爾森（Karl Pearson）的相關性討論，科莫科洛夫發表的機率公理（Axiom of Probability），開啓統計學的歷史新頁。費雪（Fisher）的實驗設計，更是讓科學的種子在非傳統科學領域中成長茁壯的重要推手。透過統計方法分析資料，我們做出具有錯誤率控制保證的推論，統計分析成了決策的重要依據。大數據浪潮之初，資料處理的技術似乎超越了一切，統計分析模式參與嚴重落後。在快速分析出結果的需求下，找到關聯性建立預測模型主導了一切。資料處理後的數據，是否眞正傳達了原始資料所提供的訊息，很少被認眞的討論；預測結果是否正確，無法從模式中探討，只能從事後的驗證得知；大數據代表的是哪一個母體的資訊，很少被提及；分析結果的適用範圍，似乎也被使用者有意無意的無限擴大。

統計學家也憂心在這一波大數據浪潮中被邊緣化，紛紛提出看法。Jordan、Lin（2014）呼籲統計學家拋開自我設限，接受不確定性與資訊科學合作共創大數據的新樂章。陳章榮、陳逸凡、趙衛中、鄒文（2015）

詳盡的討論大數據分析中所需要的統計學與數據挖掘（data mining）方法，也提出了統計學家面臨的挑戰，及統計學家應學習的電腦計算技能。Yang, etc.（2015）更以加入統計模式後的 ARGO 模式，建構更具正確預測能力的流感預測模式進化版，呈現統計方法可以讓大數據分析升級的例證。本文藉由一些大數據分析的案例，呈現大數據分析的一些盲點，也提出統計學在大數據分析上可以扮演的角色。藉此拋磚引玉，希望在這個大數據的浪潮中，統計方法能帶領我們避開荊棘，從大數據分析中眞正獲益。

參考文獻

1. Chen, James, Eric Evan Chen, Wei-Zhong Zhao, Wen Zou（2015）," Statistics in Big Data", 中國統計學報 , Vol.53, No. 3, pp. 186-202.
2. Ginsberg, J. et al.（2009）Detecting influenza epidemics using search engine query data. Nature 457:1012-1014.
3. Jordan, J. M. and Lin, D. K. J.（2014）. Statistics for Big Data: Are Statisticians Ready for Big Data? ICSA Bulletin 26, January, 2014, 58-65.
4. Mayer-Schonberger, Viktor, Kenneth Cukier（2012）, BIG DATA - A revolution that will transform how we live, work, and think. 大數據 -「數位革命」之後，「數位革命」登場：巨量資料掀起生活、工作和思考方式的全面革新。林俊宏譯，天下文化出版。
5. Onnela, J. P. et al.（2007）, Structure and Tie Strengths in Mobile Networks, *Proceedings of the National Academy of Sciences of the United States of America,* 104, p. 7332-36.
6. Yang, Shihao Mauricio Santillana, and S. C. Kou（2015）, Accurate estimation of inuenza epidemics using Google search data via ARGO, Proceedings of the National Academy of Sciences, arXiv:1505.00864v2 [stat.AP] 16 Nov 2015.
7. Wikipedia 維基百科、Big Data 大數據 https://zh.wikipedia.org/zh-tw/ 大數據 .
8. 丘昌泰（2016），以大數據挖掘主計資料金礦，主計月刊，721，p. 46-52。
9. 陳昇瑋（2015） 當電腦科學家遇上社會學－從計算社會學來看慈善募款，2015/3/11 台北大學。
10. 郭昌儒（2015a），探勘交通統計大數據（Big Data） - 高速公路一壅塞路段概況分析，主計月刊，710，p. 78-85。
11. 郭昌儒（2015b），首創巨量分析技術（Hadoop）探勘交通大數據，主計月刊，714，p. 95-98。
12. 郭昌儒（2015c），交通統計之 Big Data 大數據應用，第 24 屆南區統計研討會，國立彰化師範大學。
13. 寇星昌（2015）Big Data, Google and Disease Detection: the Statistical Story，第 24 屆南區統計研討會，張文豹先生講座，國立彰化師範大學。
14. 陳景祥（2015） R 軟體巨量資料分析：單機模式，第 24 屆南區統計研討會，國立彰化師範大學。
15. 謝邦昌、謝邦彥（2015），大數據分析在政府決策之應用，主計月刊，720，p. 26-32。
16. 饒志堅（2015），交通統計大數據首次應用分析省思，主計月刊，719，p. 78-82。❖