論述

巨量資料與政府統計

多元的社會需要更多更詳細的資料來呈現，但資料蒐集成本增加了，預算卻很難增加。由機器自動產生的巨量資料（Big Data），如信用卡帳單列示上個月的明細支出，比起個人記憶，不僅不會遺漏每筆（何時、何地、金額…）紀錄，相較於政府統計，更凸顯了精確與即時的特性。統計方法是否已過時？被巨量資料取代？或兩者可整合運用呢？

謝仁弘（行政院主計總處綜合統計處專門委員）

壹、前言

隨著越來越多的資料透過網絡傳送，及電子感測器無所不在地裝設，產生了大量、即時、多樣性的巨量資料[1]。2012 年經濟合作暨發展組織（OECD）科技展望論壇中強調：巨量資料在科學研究（包括醫療保健）、零售、金融保險及公共服務產業具龐大潛力；然而，這些每天製造、常與日常生活有關的紀錄，需經新的資訊科技（IT）與統計分析處理才有價值，對傳統的政府統計言，是個挑戰，也是機會。

聯合國歐洲經濟委員會（UNECE）於 2011 年設立一個以荷蘭爲首的跨國 10 人小組（High-Level Group, HLG），目的爲探討巨量資料在政府統計的使用問題，並提出試行方案進而國際合作[2]；本文主要係介紹該小組於 2013 年研討的內容及後續發展，並以案例說明巨量資料如何應用於政府統計與侷限性。

貳、巨量資料來源與處理方式

政府統計的資料主要源自於行政紀錄與普調查，屬規劃下的產品，在適量、穩定的結構化資料模式下進行處理分析；常由機器生成的巨量資料則爲各活動的副產品（byproducts），取得成本較低，且多屬非結構化（約占 80 － 85％）型式，其來源可分

類爲：一、行政管理（電子病歷、醫院探訪、保險紀錄、銀行紀錄等）；二、商務交易（信用卡交易、網上交易等）；三、感測器（衛星影像、道路感測器、氣候感測器等）；四、跟蹤裝置（行動電話、GPS 等）；五、行爲（網上搜索、瀏覽等）；六、看法（社交媒體等評論）（圖 1）。

這些原始資料須經中介（格式化、過濾、清理、連結、標準化…）、分析（探勘、關聯、視覺化…）等 IT 工具與統計方法處理後，才能進一步應用，若處置得當，可從多樣性的資料中萃取有價值的統計數據。

參、巨量資料提供政府統計機會

荷蘭統計局等若干政府單位或國際組織與私部門或學界合作，應用巨量資料在選定地區展開研究與試行專案，以下說明幾個成功實例。

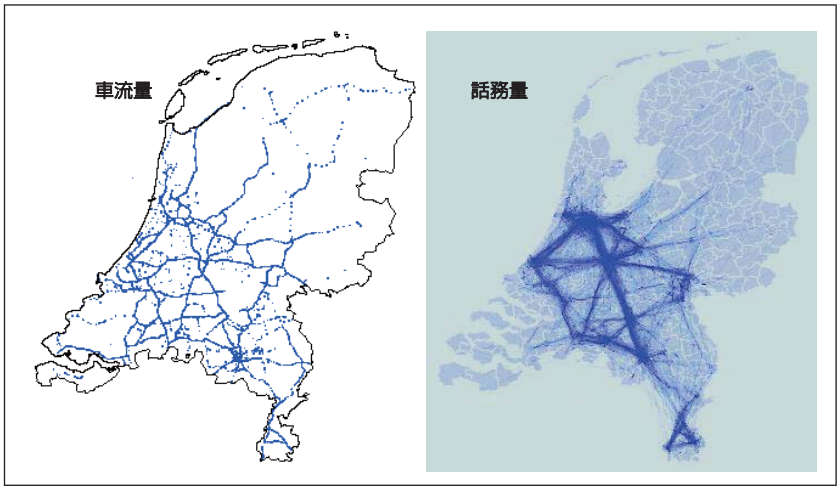
一、手機通訊與旅遊統計

在荷蘭幾乎人手一支手機且全天候開機，利用電信公司提供覆蓋率高的通訊資料，政府可進行多用途的統計，如旅遊、人群流動、道路使用及在地化的相關經濟活動等；圖 2 爲車輛在道路上 2 週的通訊紀

圖 1　大量、即時、多樣性的巨量資料來源

資料來源：Microsoft，
http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/concept_note.pdf.

圖 2　荷蘭手機話務量與車流量



資料來源：Gosse van der Veen（2013）.

論述

錄，其洞查力甚至優於調查資料。

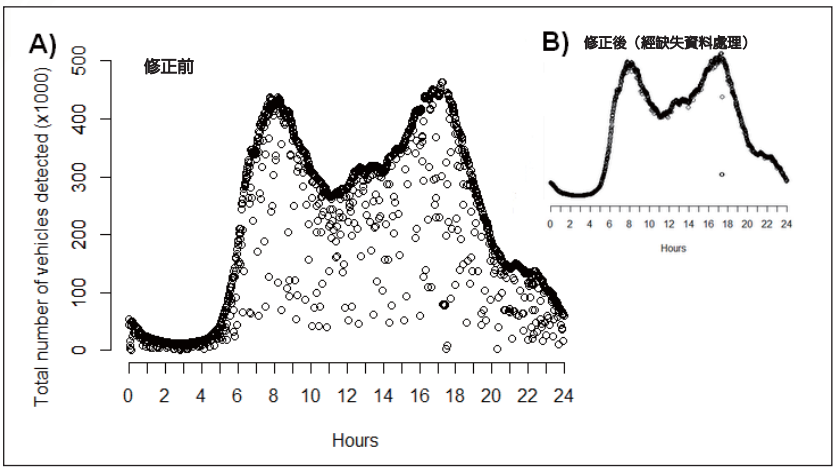
二、車輛偵測與交通及運輸統計

荷蘭的道路有設置超過 12,000 個感應線圈車輛偵測器，每天產生約 8 千萬筆車輛偵測紀錄，可作爲交通運輸統計或經濟現象的訊息來源；圖 3.1（A）爲按分鐘計算的汽車通過數，惟機器偵測並非萬無一失，如下午 5 點左右的缺失資料（missing data），必須採統計方法設算，以該時段前後 5 分鐘爲資料集插補該資料（圖 3.1（B）），插補後車輛數增加 12%。

若將偵測紀錄依汽車的長度分類：小型（≦ 5.6 公尺，黑色）、中型（＞ 5.6 與≦ 12.2 公尺，紅色）、大型（＞ 12.2 公尺，綠色），可比較 3 類汽車經缺失資料處理、標準化後的上路高峰時段（圖 3.2）。

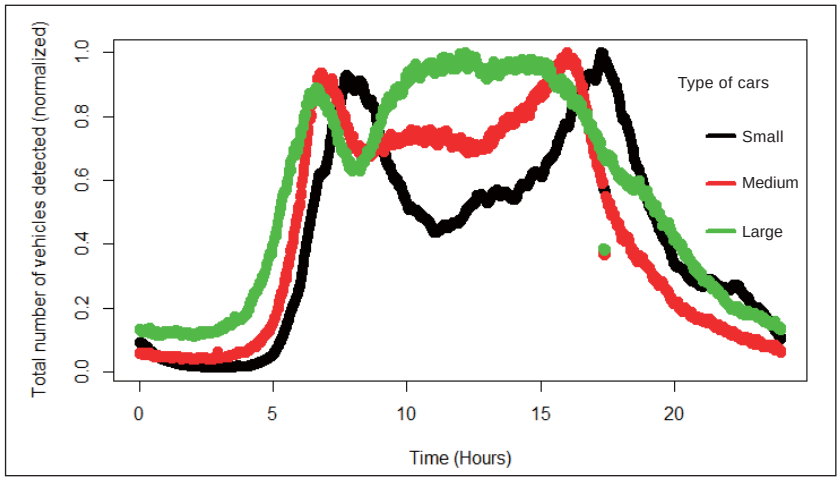
若細看個別區域（下頁圖 3.3），包括 Bergen op Zoom 附近高速公路（下頁圖 3.3 左上圖）偵測到車流量呈現不穩定，可能需要更多的統計方法。

圖 3.1　荷蘭道路偵測每分鐘通過車輛數



資料來源：Daas P.J.H., Puts M.J., Buelens B. and van den Hurk P.A.M.（2013）.

圖 3.2　3 類汽車經標準化後每分鐘通過車輛數



資料來源：同圖 3.1。

三、社交媒體統計

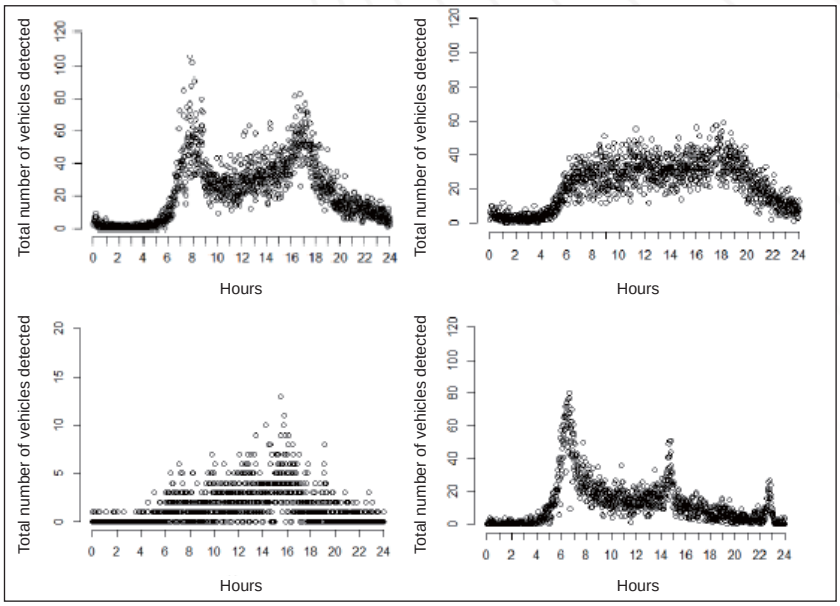
荷蘭約 70%人口在社交媒體發表留言，每天約有 100 萬則消息，經上網與親友分享、討論感興趣話題，統計局針對消息的內容與情緒進行兩面向研究，依主要來源 Twitter 的內容顯示，近 50%爲“無意義的牢騷”，其餘主要爲討論

業餘時間活動（10％）、工作（7％）、媒體（電視及電台5％）與政治（3％）等，因不太重要的“牢騷”占多數，導致阻礙了文字探勘（text mining）方法的研究。在情緒消息測定方面，發現社交媒體消息的情緒（以正面、負面、中立分類）與抽樣調查（1,500位受訪者）之消費者信心呈高度相關，特別是對經濟情勢的情緒（相關係數爲0.88），若刪除其中12月因聖誕節與對新的一年前夕較樂觀情緒資料，相關係數達0.90；其中按月、週的資料較穩定，每日資料則呈不穩定，故可在週後的第一個工作日，另產生消費者信心週線指標（圖4）。

四、網路價格與物價統計

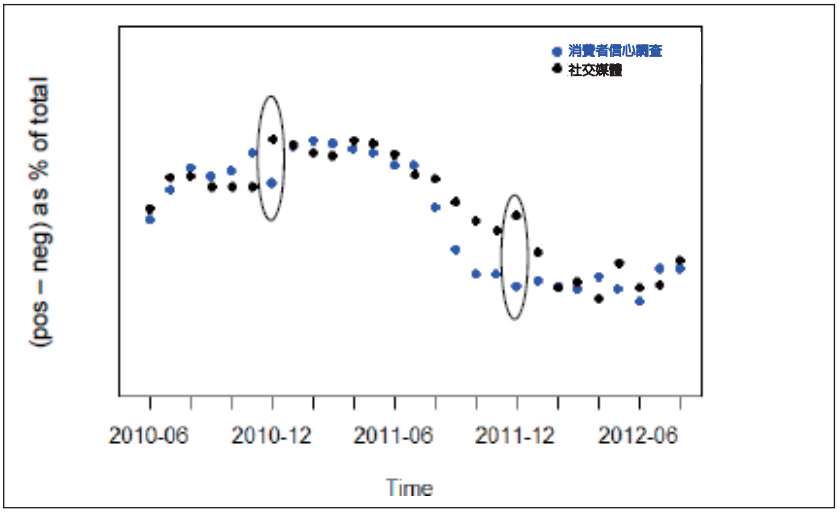
美國麻省理工學院的10億價格專案（MIT Billion Price Project, BPP）開發了應用軟體，每天自動化上網抓取超過50萬項零售商出售價格，目的爲即時（隔3天）提供按日線上物價指數（daily online price index）及估計按月、年物價變動趨勢，資料量較美國傳統的消費者物價指數（CPI）高出5倍，資料蒐集速度較快，成本也較低。如2008年9月雷曼兄弟控股公司申請破產之後，立刻發現物價出現通貨緊縮的趨

圖 3.3　荷蘭小區域道路偵測每分鐘通過車輛數



資料來源：同圖3.1。

圖 4　荷蘭社交媒體經濟情勢情緒與消費者信心



資料來源：Gosse van der Veen（2013）.

論述

勢（圖 5）；BPP 已在美、英、德、法及巴西等 22 個國家推展。

五、網路搜尋頻率統計

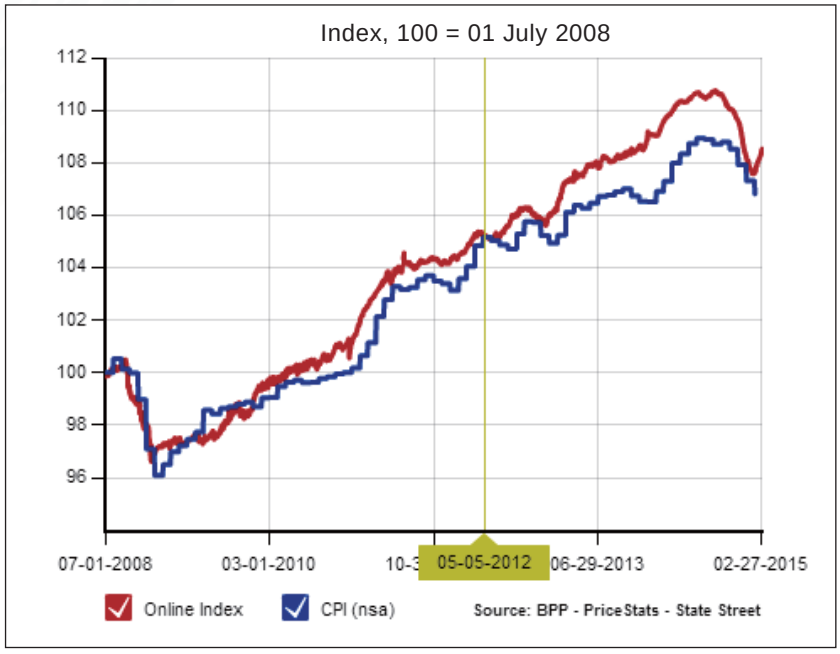
Google 公司在 2008 年推出流感趨勢工具，以搜索引擎先挑選美國人最常用前 5 千萬個搜尋字眼，再與疾病控制與預防中心（CDC）於 2003－2008 年間流感傳播資料比對，透過 4.5 億種模型測試，找出了 45 個搜尋字眼預測流感，定期更新預測模型，其結果與官方公布的眞實資料十分符合，且即時掌握流感疫情，也用在不同的國家及地區（圖 6）。

另約翰霍普金斯大學亦以類似作法搜尋 Twitter 推文，分析了超過 160 萬筆與衛生相關的推文，其資料（模型標準化後機率）與 CDC 發布 H1N1 流感率之間的相關係數達 0.958（下頁圖 7）。

六、線上話題

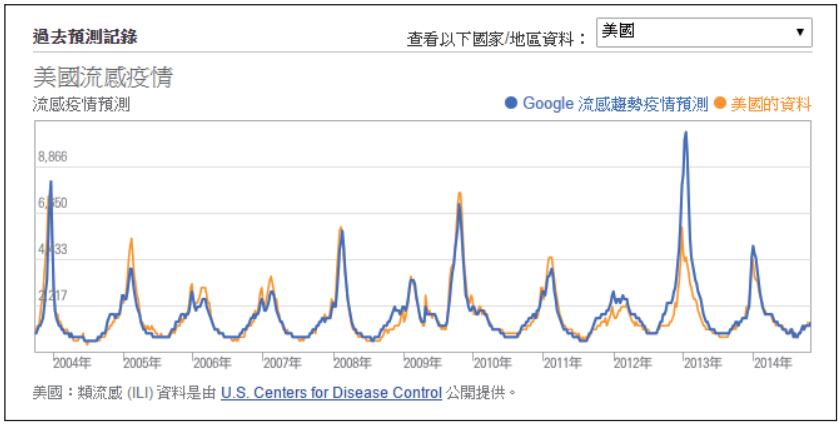
聯合國全球脈動（UN Global Pulse）與 SAS 公司合作，分析在部落格、線上論壇與新聞等社交媒體中有關就業的話題與失業率統計關係，其中愛爾蘭在失業率上升前 5、3 個月，會增加焦慮及困惑的情緒性談話，在美國失業尖峰後

圖 5　美國按日線上物價指數與 CPI



資料來源：http://bpp.mit.edu/usa/。

圖 6　美國 Google 流感疫情趨勢與實際就診資料



資料來源： Google.org. Flu Trends http://www.google.org/flutrends/intl/zh_tw/about/how.html.

2、3 個月，則增加有關房貸負擔能力及對大眾運輸與賣車的討論。

七、美國資料整合

美國普查局及勞工統計局利用電子交易、社交媒體訊息、互聯網搜索與行政資料，經混搭（mash up）、建模、資料分析以補充或改善統計品質，強化小區域的估計，如：

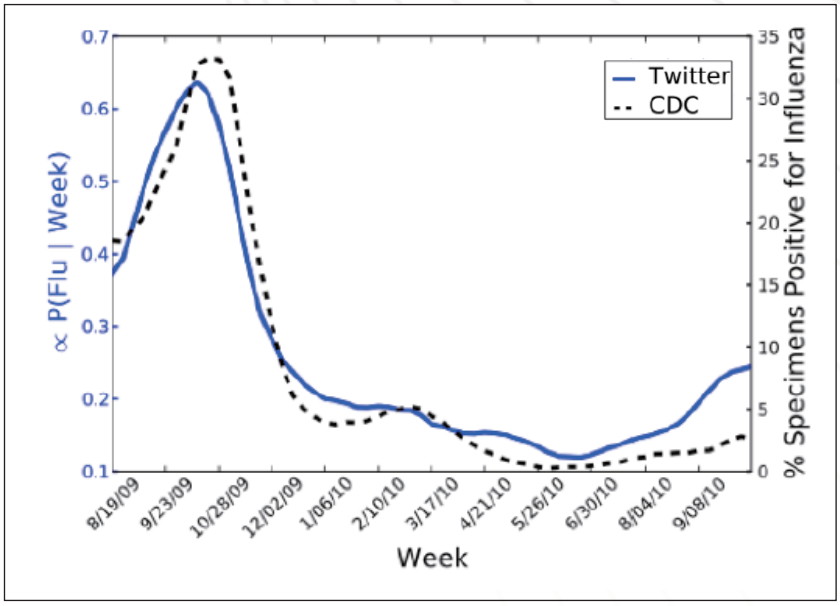
（一）2020 年普查：建立無縫與不斷更新之額外描述的背景資料（paradata），使 10 年一次的普查運作更靈活，可降低成本，減少受查者負擔。

（二）營建統計：引用建設與住房（法拍屋、公共建設及建築許可登錄）資料。

（三）零售及服務業統計：利用電子支付處理，填補如依區域、企業規模或新商品銷售的細項資料。

（四）其他如 CPI 與零售掃描資料之間的比較研究，美國社區調查結合在地資料進行廣泛測試等。

圖 7　美國 Twitter 推文與 CDC 發布流感率趨勢



資料來源：“You Are What You Tweet: Analyzing Twitter for Public Health.”M. J. Paul and M. Dredze, 2011. http://www.cs.jhu.edu/%7Empaul/files/2011.icwsm.twitter_health.pdf.

八、南韓試行專案

南韓統計局設計了自動收集媒體資料系統，進行一項試行專案，於 2013 年 3、4 月起按月試編工業生產指數，在工礦業中選出 4 個業別、162 項產品與 1,438 場所單位爲對象，其結果將與原調查資料併行分析，未來更將擴大到所有業別。

肆、巨量資料帶來政府統計的挑戰

使用巨量資料有其積極面，但也涉及到安全性、保密性、分析及解釋，甚而立法等諸多挑戰；在分析結果前，應進行適當的檢核，考量其品質、有效性與侷限性。前述案例之統計應用須關注的包括：

一、統計方法的挑戰與新思維

（一）典型的巨量資料集並非來自依統計目的所做的事前設計，很難使用傳統的統計方法與工具，如機率抽樣、統計分

論述

類等。在情感分析案例中，如何以文字探勘方法將情緒話題分類成正面、負面、中立的概念，極具挑戰性。

（二）巨量資料大多屬非結構化，即未預先定義資料型態或不適於傳統的關聯型資料庫，會有很多如缺失資料、量測誤差等難以處理的問題；如車輛偵測例子中，伺服器當機與網絡中斷導致資料遺失。

（三）巨量資料以演算法進行推論，從自動蒐集到的資料中不斷地重覆修正，找出最適模型或相關性；惟在無特定目的、事前設計及明確定義下，所得到的資訊是否足以解釋，仍有風險，況且有許多爲無用（如“牢騷”話題內容案例）或假資料。

（四）以巨量資料推計或視同母體在分析上存有危險，如瞎子摸象，當需要越大資料量時，就越難得到整體圖像（global picture），因所看到的可能是小區域（local）的拼圖。掌握整體資訊，與客製化輔助做小區域的推計，其實是兩個不同向度的分析，在統計上也是新的挑戰；案例中如車輛偵測並未普及、以年輕人爲主的 Twitter 推文以及網路價格等皆屬選擇性對象；又按分鐘計算車流量而發現了缺失資料，但若改爲按小時計算，可能會誤以爲數量減少而忽略這個問題。

（五）巨量資料有助於建立小區域估計，惟應考量資料的波動性（volatility）與解析度（resolution），小區域道路偵測的車流量波動很大，這些波動雖是事實，但過高的解析度反造成統計上的困擾；同樣地，在每天的情感分析亦有類似情形，建議開發統計方法，如採移動平均與過濾技術的時間序列分析。

二、資料驅動創新

（一）巨量資料與傳統政府統計（調查、公務）具互補作用，如美國、南韓案例；巨量資料不僅可強化公務統計，對於調查統計之輔助包括：1. 提供變量，協助抽樣調查作更好的分層與改進估計方法；2. 插補未回答資料；3. 提高資料發布的頻率與即時性；4. 改善及提供更多的小區域估計等。

（二）巨量資料以全新的統計角色，從新的視角建立趨勢指標，提供預警、即時監測的功能，而非替代；如當下量測（nowcasting），強調

的是即時量測「當下」，而非預測（forecasting）未來。

伍、結論

近年巨量資料掀起了風潮，是舊瓶裝新酒？還是新瓶裝舊酒？其實，統計學家在早期就有處理如天文、匯率、醫療保險及地震等巨量資料的經驗；所謂“大”與技術成熟有關，當儲存技術改變的時候，原來的大就不那麼大，當計算的能力增加時，原來很難計算的就不那麼難算，所以是與時俱進不斷地在改變，且大與否跟要問的或想回答的問題有關。

由於巨量資料快速成長，需要靠 IT 與統計分析的技術解決，目前 IT 已有很大程度的進展，而注重品質、透明度與合理方法的政府統計則相對保守與嚴謹。2013 年起若干國家統計機構（NSIs）與國際組織，包括聯合國統計司（UNSD）、UNECE、OECD 及歐盟統計局（Eurostat）已採取各項措施，協調解決許多挑戰與問題，國內各政府統計機構宜持續關注其發展。

註釋

1. 依國際研究暨顧問機構 Gartner 的定義，巨量資料爲「大量（high-volume）、快速累積（high- velocity）、具有多樣性（high-variety）的資訊資產，需要具成本效益、創新的處理技術，以發掘問題、提升決策品質」；同時，巨量資料的眞實性（veracity）也需要被檢驗。
2. HLG 成員包括來自荷蘭（小組主席）、澳大利亞、墨西哥、義大利、南韓、斯洛維尼亞、美國、歐盟、OECD 及 UNECE 等專家，相關研討文獻請參考：http://www1.unece.org/stat/platform/label/bigdata_resources。

參考文獻

1. 張源俊（2013），“Statistics in the Big Data Era -Going beyond the buzzword”，2013 Big Data Forum Jointly organized by CITI &IIS, Academia Sinica.
2. Daas P.J.H., Puts M.J., Buelens B. and van den Hurk P.A.M.（2013），“Big Data and Official Statistics”.
3. Gosse van der Veen（2013），“Big Data：Big Opportunity!”，High Level Group, for the Modernization of Statistical Products and Services，UNECE.
4. Jeong-Im Ahn , Young-Ja Hwang（2013），“Production of Official Statistics by Using Big Data”.
5. Martin Karlberg , Michail Skaliotis（2013），“Big Data for Official Statistics—Strategies and Some Initial European Applications”.
6. Reimsbach-Kounatze, C.（2015），“The Proliferation of “Big Data” and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis”，OECD Digital Economy Papers, No. 245, OECD Publishing.
7. UN Global Pulse（2012），*Big Data for Development：Challenges & Opportunities*.
8. William G. Bostic, Jr.（2013），“Big Data for Policy, Development, and Official Statistics”.❖