創新變革精進獎勵項目

# 擘建基石－運用大數據把關海關巨量資料品質

本專案運用大數據技術進行資料探勘，將紙本檢核作業全面電子化，融合統計學、歐盟技術文件與我國貿易統計資料特性，擘建自動化資料分析模型，偵測異常情形，配合視覺化結果呈現及自動派送作業，大幅提高檢核效率，確保海關進出口貿易統計資料品質。

財政部統計處（黃科員敏慈 、魏稽核政毅、吳主任佩璇）

## 壹、前言

### 一、海關資料的巨量特性

海關進出口貿易統計爲我國重要經濟統計，由關務署統計室專責編製，以海關進出口報單爲主要資料來源（海外售魚資料另由農委會漁業署提供），按進出口、稅則或貨品分類號列、國別（進口爲生產國別，出口爲目的地國別）等分類，按月彙總金額、重量及數量發布。海關每月報單量約百萬份（不含快遞簡易報單），依申報之貨品分類號列 11 碼總計 12,120 項[1]，貿易夥伴國（或實體）則有 252 個，因此各進出口別下之貨品分類號列與夥伴國約有 280 萬種組合，數據量相當龐大。

### 二、貿易統計之挑戰

貿易統計非屬調查統計，不存在抽樣誤差或推論誤差，一般認爲資料品質理應正確無虞，實則不然。美國普查局（Census Bureau）在「美國進出口貿易統計指南」（Guide to Foreign Trade Statistics）中即指出，即使在資料蒐集、處理與製表各階段都做了品質確保程序，資料仍會存在各種非抽樣誤差，其中首要爲申報誤差（reporting error），申報誤差係指進、出口人或其代理人於填報進出口資料時所發生之錯誤或遺漏，通常爲貨品分類號列、金額、數量、重量及國

別之遺漏或不正確申報，我國亦無例外。

我國貿易統計依附於海關既有制度，而近年海關爲順應國際趨勢，積極推動通關自動化與便捷化，「免審免驗（C1）」已成爲進出口報單最主要的通關方式，空運進出口及海運出口 C1 通關比率均逾 8 成，且比率仍持續攀高，此類未經估驗或「先放後核」報單依進出口人申報內容直接進入貿易統計，倘有申報誤差即可能威脅進出口貿易統計品質。錯誤申報屬個案，雖不致對整體進、出口總量值產生顯著影響，亦可能對個別貨品統計造成重大影響，進而涉及商民營業利益或消費權益，使貿易統計資料檢核工作益顯重要。

## 貳、檢核方法研究

考量進出口資料之複分類組合繁多，且數據量極爲龐大，在資料須按日 / 週檢核、按月發布之極短週期下，檢核方法之統計理論應以資訊系統頻繁運算承載能力內，具備可操作性者爲前提。本研究經評估時間數列分析法、信賴區間檢定法及線性迴歸法，採用後二者。

### 一、信賴區間檢定法

信賴區間檢定法不考慮資料的時間先後，較時間數列分析法簡便易行，係以既有資料建構信賴區間（Confidence Interval）後，再以此區間評估欲檢核資料的合理性。

### 二、線性迴歸法

由於個別貨品之統計用數量與重量之間具線性關係，統計用數量越多，重量亦等比例增加，此特性即可採線性迴歸法（Linear Regression）進行檢核。其中統計用數量係依據貿易管理或產業需要而建立標準化之計量單位，依各貨品屬性不同而異；現行需申報統計用數量之貨品分類號列共 3,050 項，示例如表 1。

## 參、自動化分析系統

本專案自力擘建全新的檢核模型，並配合視覺化判讀及自動派送作業，大幅提高效率。檢核模型之建構，係以統計理

表 1　貨品分類號列與統計用數量單位對照表（節錄）

| 貨品分類號列 | 中文貨名 | 統計用數量單位 |
|---|---|---|
| 01013000002 | 驢 | HED （頭） |
| 22041010006 | 香檳 | LTR （公升） |
| 27111200002 | 液化丙烷 | TNE （噸） |
| 39262000298 | 其他塑膠製手套 | DZN （打） |
| 44031100100 | 針葉樹類橡膠原木 | MTQ （立方公尺） |
| 52094200009 | 牛仔布 | MTK （平方公尺） |
| 64041100105 | 網球鞋 | NPR （雙） |

資料來源：中華民國輸出入貨品分類表。

# 創新變革精進獎勵項目

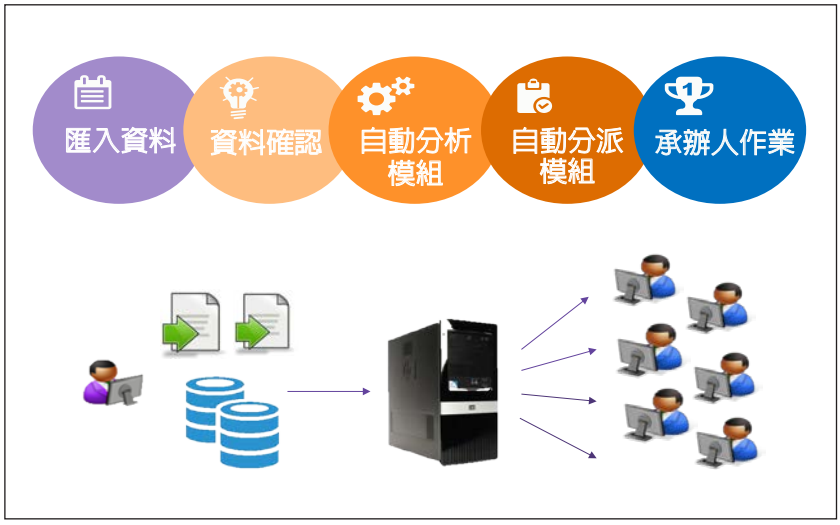
論爲基礎，運用大數據資訊技術（R 軟體等），由多個構面，建置適用不同檢核時點（週或月）之自動化資料分析模型，篩選出疑似異常分類或數據，同時呈現歷史資料與檢核結果，利於比較資料異常程度，輔助檢核人員快速篩選異常報單，降低人工檢核負擔及疏漏情形。系統運作流程僅須由操作人員匯入資料源，系統即會完整運行（圖 1）。

資訊工具選擇上，使用目前最普及的資料分析軟體：R 語言，其內建統計學和數學演算功能，且具有開源（Open Source）程式碼特性，可依資料特性量身訂做適合之統計模型並將結果產生視覺化圖形。

## 一、匯入資料

依照資料探勘標準作業程序（Cross － Industry Standard Process for Data Mining , CRISP － DM），進行資料前置作業（Data － Preparation）。將每月貿易統計初步彙整資料、每週由進出口報單產製之檢核用清單、貨物價值增減比較表等資料源，共同匯入檢核系統。

圖 1　自動化分析系統流程圖




資料來源：本研究自行整理。

## 二、資料確認

系統進行資料確認，將各來源資料進行屬性與欄位統一，並交叉驗證各來源資料完整性。

## 三、自動分析模組

依據我國貿易統計資料特性建置自動化分析模組，於不同檢核時間點，使用統計方法建立各式模型（Modeling and Evaluation），檢視資料正確性。

自動分析模組包括兩部分（下頁圖 2），依照不同檢核標的建立各式子模型，選擇適當的統計方法或演算機制，進行異常資料判別。各子模型依性質分可分爲兩部分：

第一部分源於數據科學作爲判斷標準，將各種貨品分類號列與國家之組合，分別對總金額、數量、重量進行檢核；另外兩兩組合而成之延伸指標亦可成爲檢核標的，總金額除以總重量爲平均重量單價，總

金額除以總數量爲平均數量單價，數量與重量之比值爲數重比。搭配不同的統計方法，進行異常值檢定。

第二部分源於專家經驗而來之演算機制，對於國際組織、各產業公會、輿論關注之特定貨品，設立監控機制，使承辦人能快速掌握這些貨品之進出口情形。另外針對國內進出口量極低之貨品以及容易被申報錯誤之特定國別，進行資料撈取，供承辦員檢視資料合理性。

## 四、自動分派模組

檢核結果由資料中心執行自動分派作業（Deployment），將電子清單檔與視覺化分析圖檔依照貨品範圍上傳至各檢核人員之網路雲端空間供檢核人員使用。分派完畢後，產製電腦日誌（LOG 檔），紀錄每次執行時間與分派至各檢核人員之檔案數，利於管理人員監控系統是否正常運行。

## 五、承辦人作業

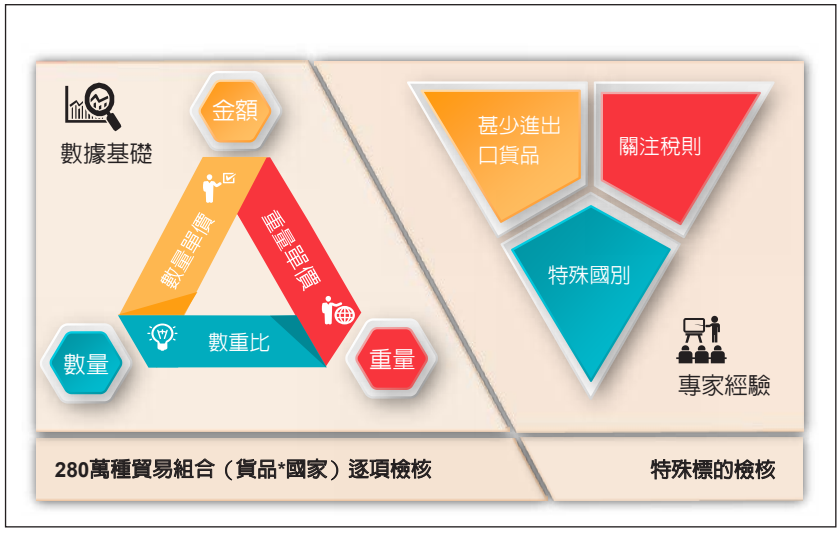
各承辦人由不同構面交叉分析總值（量）、平均單價及數重比等資料，透過圖像及顏色等視覺化資訊呈現，迅速掌握巨量資料中各式異常態樣。綜合判讀結果如有異常，即進行處理決策。

## 肆、系統特點與具體效益

本系統翻轉過去檢核思維，由以往「由下而上（Roll－up）」的單筆檢核方式，改變爲「由上而下（Drill－down）」的下鑽檢核思維（下頁圖 3），彌補了由單筆報單開始檢核容易見樹不見林之不足，提升統計品質。分析模組中植基統計理論，由數據科學基礎做爲決策依據，配合實務研提增進檢定效能，提出包括不計入進出口量值低於特定值之月份、以單尾檢定（Cantelli inequality）取代雙尾檢定（Chebyshev's Theorem）、進出口廠商單價風險矩陣等增進運作效能之策略，將抽象的統計理論、大數據資料處理技術，體現實踐在政府統計中。

本系統由規劃至程式開發

圖 2　自動分析模組內涵




資料來源：本研究自行整理。

# 創新變革精進獎勵項目

設計，均由同仁自力完成，運作效益高，系統維護成本低。使用大數據資訊技術包括資料處理、自動分析、自動派送，將紙本作業全面電子化，紙張列印大幅減少。另經由電腦輔助進行分析，篩選異常資料，降低人工檢核負擔，搭配視覺化圖形呈現檢核結果，提高檢核效率，檢核工作由勞力密集升級爲知識密集，人力資本大幅提升。

## 伍、結語

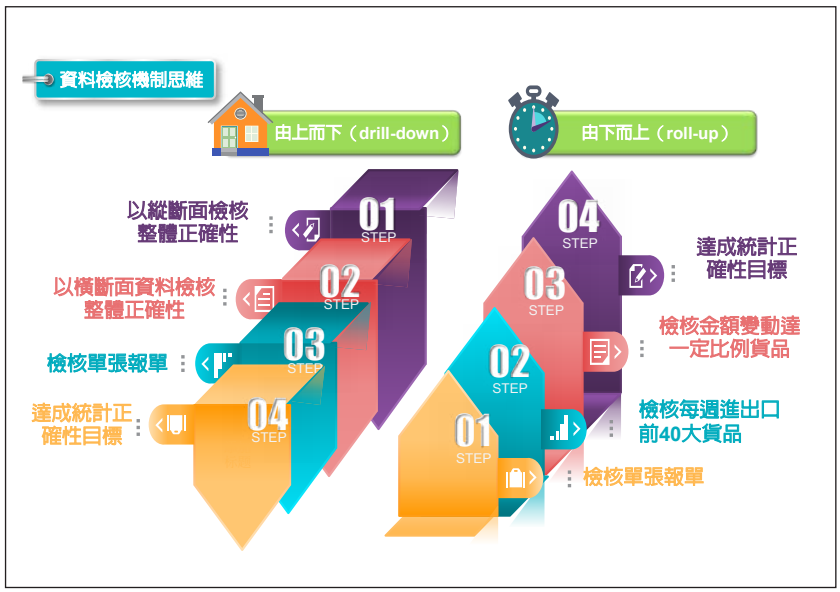
資料品質是後端應用分析的基石，進出口統計爲研判經貿情勢、編製國民所得及國際收支統計的重要基礎資料，資料品質若不佳，可能影響後端應用分析之參考價值。此外，各類鉅細靡遺貨品的進出口數量多寡與價格高低，均有特定利益團體及跨境移動貨品主管機關關注，貿易統計均需完整、正確且即時地滿足其資料需求。

透過大數據方法，通關資料不再只是一堆放行後歸檔的歷史紀錄，藉由建立模組及分析過程，逐漸賦與其生命力。海關業務範圍繁雜，貨品種類眾多，導入資料科學這個利器，協助我們做出更精準的判斷，提供更高品質的貿易統計資料。

## 註釋

1. 經濟部國際貿易局截至 107 年 12 月 13 日公告數，並配合世界關務組織 (WCO) 稅則分類及各主管機關建議時有修訂。❖

圖 3　資料檢核機制思維




資料來源：本研究自行整理。