多類型頻繁樣態探勘應用於國軍雲端主財系統數據分析之研究

陳良駒 [a*] 傅祖銘 [b]

[a] 國防大學資訊管理學系

[b] 國防部主計局財務中心

論文編號: NM-44-02-03
DOI: 10.29496/JNDM.202411_45(2).0004
來稿 2023 年 6 月 5 日→第一次修訂 2023 年 12 月 13 日→同意刊登 2023 年 12 月 27 日

摘要

在全球化及網際化演進的趨勢下，企業組織及政府機關均仰賴管理制度及資訊系統來協助組織各項業務之推展。在面臨組織多元系統整合及多樣數據分析的複雜環境下，組織內部控制與稽核等機制之有效運作，將提昇作業流程服務之效能。然而，傳統以人工方式進行的內部控制程序及作業，恐不易及時反應內部作業之漏洞及適時預警財務業務之風險。因此，資料探勘及數據分析技術將有助於組織內部控制與稽核領域之發展。

本研究植基於多類型的關聯規則演算方法，以國軍主財資訊雲端服務系統下的人員申領休假補助費退件缺失數據集為範例，探索不同缺失樣態的連結關係，除揭露不同演算機制下的缺失樣態關聯外，也嘗試進行樣態處理之差異性分析。相關成果可納入系統防呆檢核，或是內部稽核第一線人員之訓練教案，應可有效提升各項稽核作業之效率。

關鍵詞：內部稽核、國軍雲端主財系統、休假補助費、頻繁樣態探勘、關聯規則

* 聯絡作者：陳良駒 email: ndmchorse@gmail.com

The Study on the Application of Multi-Class Frequent Pattern Mining to the Data Analysis of the National Defense Comptroller Cloud Information System

Chen, Liang-Chu [a*] Fu, Tsu-Ming [b]

[a] Department of Information Management, National Defense University, *Taiwan, R.O.C*

[b] Finance Center, Comptroller Bureau Ministry of National Defense, *Taiwan, R.O.C*

Abstract

In the era of globalization and the rapidly evolving digital landscape, businesses and government agencies increasingly rely on management and information systems to support their diverse operations. Within the complex environment of integrating multiple systems and analyzing heterogeneous data, the effective functioning of internal control and auditing mechanisms is crucial for optimizing operational efficiency. However, traditional manual internal control procedures often fail to address operational vulnerabilities or provide timely warnings for financial risks. As a result, leveraging data mining and analysis techniques can empower organizations to enhance their internal control and auditing processes.

This study focuses on employing multi-type association rule calculation methods to analyze a dataset containing erroneous data from vacation subsidy and refund applications within the National Defense Comptroller Cloud Information System. The primary objective is to investigate the association patterns among various types of errors in the dataset. In addition to identifying these patterns using different algorithmic approaches, the study aims to conduct a differential analysis of the discovered associations. The findings are expected to offer valuable insights for developing system-level error-proofing mechanisms and designing targeted training programs for frontline internal audit personnel. Ultimately, these efforts are intended to improve the overall efficiency of audit operations.

Keywords: Internal Audit, Comptroller Cloud Information System, Vacation Subsidy and Refund Applications, Frequent Pattern Mining, Association Rule

[*] Correspondng Author: Chen, Liang-Chu email: ndmchorse@gmail.com

一、前言

網際網路及大數據發展的時代趨勢下，企業組織與政府部門面臨諸多內、外在環境的挑戰，對於組織內部流程設計、資源管理配置、財務稽核管控等作業均需要有適度的管理規範及完善的制度機制來提升組織服務效能。其中內部稽核（internal audit）係以風險管理為基礎的審查機制，藉由獨立、客觀及系統化的方法來協助企業組織評估業務流程及認知管理風險，以及早進行營運改善之作為（王威智，2004），內部稽核是維持組織穩定運作的重要機制。

然而，資訊科技的快速發展已造成企業組織的內外部環境發生巨大的變化，組織各項事務若以傳統人工方式進行內部控制與稽核作業，恐不易及時反應內部作業之漏洞與適時預警財務業務之風險（孫嘉明等，2017）；同時，企業面對來自內、外部網際資訊系統的龐大數據，想要快速發現組織內部業務中的可能疑點，已成為當前稽核作業的困難之處（Yan et al., 2019; Nan, 2022）。因此，為提升組織營運價值及作業流程效率，應加強內部稽核及控制的應用與發展（傅祖銘，2020）。其中，資料探勘（data mining）技術及演算法的應用有助於稽核人員從事業務風險的評估，並提升流程稽核之效能（Bose et al., 2022），該技術係從大量複雜數據中進行結構化的歸納分析，以挖掘未知的潛在資訊，做為組織決策或業務改善之參考。然而，儘管學術界對資料探勘及分析已有諸多討論，但在內部控制與稽核領域中運用數據分析的實證研究仍處於起步階段（Earley, 2015）。

國軍為因應全球網際化的發展，分別以電子郵件、國軍智慧卡為基礎來建置及整合國軍雲端之應用服務。「國軍主財資訊雲端服務網」係彙整各類主財業務資訊系統來提供雲端化的作業介面與資訊服務，以支援各預算支用單位、財務單位、業務單位及支薪單位進行國軍官兵個人給與撥戶等服務（林永能，2015）。其中薪餉次系統包含月支薪給、年終工作獎金、考績獎金及各項補助（結婚、生育、喪葬、殮葬、教育及休假）等費用給與核結與發放作業；此外，另有薪資所得稅、軍人保險費、退休撫卹基金、強制執行等代扣款項計算及系統檔案維護作業等，可有效提供各單位作業承辦人依實際業務辦理即時薪餉發放或資料異動等作業（傅祖銘，2020）。各項補助費發放係為國軍人員受理臨櫃案件數量最多的業務，其中休假補助費須填製內容資料相較其他補助費要求更多，因此發生缺失退件之情形佔各項補助費之冠，也是近年 1985 申訴專線重點檢討項目。故本研究即以國軍休假補助費缺失頻繁項為數據探勘與分析之標的。

而頻繁樣式資料探勘技術，包括分類（classification）、分群（clustering）、關聯規則及循序樣態等方法（Tsai et al., 2013），其中關聯規則目標係為發現非循序性的事件樣態（Angeline, 2013），適合挖掘國軍財務數據的潛在關聯，以利組織內部稽核之參照。而關聯規則演算模式一般可分合併為基演算法（join-based algorithms）、樹狀為基演算法（tree-based algorithms）及樣型成長演算法（pattern growth algorithms）三大類（Chee et al., 2019）。因此，本研究目的即藉由多類型的關聯規則演算方法，探索及分析國軍休假補助費缺失頻繁項在不同演算模型中的樣態關聯性，以實證為基礎來瞭解各項缺失樣態間之問題，進而精煉內部稽核效率，亦可避免類似案例的持續發生。

二、文獻探討

2.1 內部稽核

內部稽核概念起源自歐美國家的內部控制作業機制，我國政府的內部稽核發展原則亦在其影響下，逐步建立制度。相關概念及方法最早見於 1968 年訂頒之「加強政治經濟工作效率計劃綱要」所列改進會計審計職能，以會計人員執行事前審核，控制收支等事項（陳浚明，2009）。內部稽核係以獨立、客觀的諮詢及審查活動，協助企業組織透過系統化的方法來評估作業流程及管理風險，以增加業務價值並改善營運流程，達成組織內部管控的目標（王威智，2004）。近年來，以風險為基礎之持續性稽核與監控受到許多學者的關注，持續性稽核的概念包括持續性風險評估、持續性控制監督模組及持續性異常偵測模組等三面向（石淑暖，2016）。

謝家蘭（2015）認為持續性稽核係各項查核作業資訊化，以減輕企業稽核人員工作負擔，並提升稽核作業效率，同時減少企業營運或決策上錯誤的損失。石淑暖（2016）探討主計資訊系統導入持續性稽核技術研究成果的運用，認為持續性稽核係以自動化的資訊科技進行組織業務風險評估或監控，以及早發現可能的風控問題，並適時修正相關作業流程。一般具有強化內部控管機制、提升稽核成效、提供即時預警、避免風險等效益。

在資訊科技的快速發展下，企業的內外部環境發生了巨大的變化，以傳統人工方式將個人經驗和判斷進行內部控制與稽核作業，較不易檢測內部稽核之錯漏問題及潛在風險（孫嘉明等，2017）。而在資訊網路高速發展的時代下，企業面對來自內、外部單位的龐大數據，想要快速發現財務會計業務中的疑點，已成為當前稽核作業的困難之處（Yan et al., 2019; Nan, 2022）。

然而，資訊技術的發展給審計作業帶來了契機，資料探勘技術的應用和演算法的改良有助於審計風險評估過程的順利進行。隨著大數據議題的興起，學者 Earley（2015）認為在稽核作業中使用數據分析有四個主要好處：（1）可增加稽核證據的充分性，亦即可以測試比現在更多的交易資料；（2）可以通過更深入地了解客戶的交易來提高稽核品質；（3）稽核人員利用工具和技術來分析數據，可改進稽核作業中潛在的欺詐檢測問題；（4）稽核人員可以使用非財務數據或外部資料，為稽核作業提供更多的訊息；同時藉由數據可以建立預測模型，以提供業務處理問題的解決建議。Bose et al.（2022）則認為數據分析可以提供稽核人員處理下列工作事項：（1）深入分析公司的總賬系統以提供稽核證據；（2）檢測財務欺詐並進行法務會計的修正；（3）協助檢測異常和趨勢，以及在風險評估中比較行業數據；（4）通過整合外部數據，可以為客戶提供超出其當前能力範圍的服務和解決方案。

因此，資料探勘與數據分析技術可協助稽核人員辨識出感興趣的異常態樣，並透過數據關聯性發現其他類似的異常資料，以利及早採取因應作為並可大幅度改善內部稽核的工作效率（Capriotti, 2014）。

2.2 資料探勘技術應用於內部稽核之研究

資料是組織中的關鍵資產，也是協助企業決策的重要依據，因此資料探勘或資料庫知識發現（Knowledge Discovery in Databases, KDD）技術扮演著企業各項服務或應用的

重要角色（Frawley et al., 1992）；其目的係針對組織數據資料庫中挖掘隱含、未知、潛在或具有關聯性的有用訊息（Keyvanpour et al., 2011）。迄今為止，資料探勘已經被廣泛應用於製造業品質改善（Köksal et al., 2011）、金融詐欺偵測（Ngai et al., 2011）、犯罪行為預測（Hassani et al., 2016）、醫療保健服務（Islam et al., 2018）及教育學習分析（Romero and Ventura, 2020）等領域作業。然而，雖然已經有部分研究利用資料探勘與人工智慧相關技術於財務報表審查、業務流程及內部控制檢測、政策遵循及舞弊/詐欺偵查等面向之研究（Shabani et al., 2021），但在內部控制與稽核領域中運用數據分析的實證研究卻仍處於起步階段（Earley, 2015）。

Kirkos et al.（2007）探討了決策樹、神經網絡和貝葉斯信念網絡等三種資料探勘分類技術在檢測已發布的財務數據中，發現欺詐性財務報表的有用性。Bai et al.（2008）蒐集國際虛假財務報表（False Financial Statements, FFS）事件之技巧、指標等資訊，同時檢視 10 家中國有 FFS 歷史的公司，從其中萃取特徵，並使用分類和迴歸樹（Classification and Regression Tree, CART）技術進行學習及虛假財務報表之預測。

黃士銘等（2012）學者認為電腦稽核領域常使用班佛法則（Benford Law）、二階班佛法則（Second-order Benford Law）、齊普夫定律（Zipf's Law）與串相似度算法（Levenshtein Distance）等文字探勘規則與技術來協助查核財務舞弊事件，並以電子業製造商重複付款查核為例進行說明。Albashrawi（2016）回顧 2004-2015 年間使用資料探勘工具檢測金融欺詐的研究文獻，發現財務報表欺詐和銀行欺詐是該領域正在調查的兩個最大的金融應用，而其探勘工具較常使用監督式學習工具，其中邏輯回歸模型（logistic regression model）是檢測金融欺詐最常使用的方法。Al-Hashedi and Magalingam（2021）透過文獻蒐整與回顧來探索植基於資料探勘技術的財務詐欺偵測（financial fraud detection）研究，發現多數文獻聚焦於銀行及保險的詐欺研究，而 SVM、Naive Bayes 及 Random Forest 是前三項最常使用的資料探勘技術。

Saglar and Kefe（2021）介紹多項資料探勘技術應用於財務稽核流程之研究，包括類神經網路（Artificial Neural Networks, ANN）、邏輯迴歸（Logistic Regression, LR）、決策樹（Decision Trees, DT）、支援向量機（Support Vector Machines, SVM）、基因演算（Genetic Algorithms, GA）及文字探勘（txt mining）等，同時強調對於財務稽核之目的著重於預測。Boskou et al.（2018）蒐集 133 家希臘公司財報，分別以文字探勘、SVM/ PCA/ Regression 等方法進行分析，發現職責分離（separation of duties and responsibilities）等 11 項內部稽核的重要因素。Ren and Chen（2021）以聚類分析方法進行公務差旅費用數據之探勘，發現 20 組不合法及 75 違規的差旅使用情形。Shan et al.（2022）針對某商業銀行的 1500 個客戶信貸資料下的 24 項屬性，以多隨機決策樹方法建立財務稽核（financial audit）預測之模型，驗證該方法優於傳統 C4.5 的決策演算。Xuanyuan et al.（2022）蒐整中國壽險公司財務資訊，以 C4.5 決策樹演算方法挖掘影響保險金融的關鍵因素，協助建立決策模式的優化。

而 Tsai et al.（2013）依據問題特性（classify pattern/ find events）、分類標籤（labeled data/ unlabeled data）及樣態循序性（sequential/ nonsequential）等指標來區分不同的頻繁樣式資料探勘技術，包括分類、分群、關聯規則及循序樣態；其中關聯規則目標係為發

現非循序性的事件樣態，適合挖掘國軍財務數據的潛在樣態。

傅祖銘（2020）嘗試以結構式的概念範例來說明導入 FP-growth 資料探勘演算法至國軍人員休假補助費之數據分析可能解決方案，該研究雖宣稱可透過演算流程發現數據中缺失態樣、錯誤樣本及可能風險等關聯性，但並未具體呈現數據來源及關聯分析之成果。陳良駒與傅祖銘（2022）則企圖蒐整國軍人員休假補助費之歷年數據，藉由 Apriori 演算規則的特性來發現缺失態樣的相關性，雖然挖掘出部分缺失規則，但其方法與數據的完整度並不足夠。本研究將擴展關聯規則之特性，以進行更為全面性的解析。

2.3 頻繁樣式探勘：關聯規則

資料探勘與分析是智慧型社會發展的重要技術，許多學者也積極開發不同領域之應用；在多種資料探勘技術中，頻繁樣式探勘（Frequent Pattern Mining, FPM）是其中一種重要的發展模式。頻繁樣式（frequent pattern）是指在特定支持度門檻下所頻繁共同出現的項目集合（Agrawal et al., 1993）。Yun and Ryu（2011）發現頻繁樣式探勘也運用於許多不同類型的組織數據，包括關聯規則（association rules）、相關性（correlations）、循序樣態（sequential patterns）、串流資料（stream data）、圖形樣態（graph patterns）等，其中關聯規則目標係為發現非循序性的事件樣態，是應用最廣並受到關注的一項技術（Angeline, 2013）。

頻繁樣式探勘能夠挖掘資料庫中不同項目之間的重複樣態，並透過關聯（association）形式來表達；藉由迭代的逐次掃描並計算資料庫中每個頻繁項目集，在滿足設定條件的最小支持度下，直到不可能再掃描出更多的 k 個項目集為止（Chee et al., 2019）。其運作基本概念如圖 1 所示。

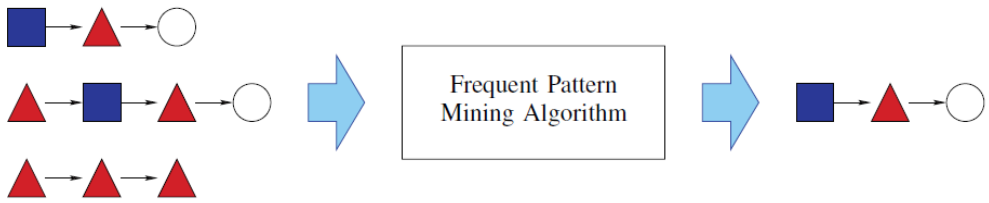


圖 1　頻繁樣式探勘運作流程概念
資料來源：Tsai et al.（2013）

關聯規則主要目的是從龐大交易記錄資料庫中，以產生大量規則以尋找未知項目組合與關聯性（Agrawal et al., 1993）；亦即藉由適當的邏輯或規則找出同時發生的項目或事物。相關分析已展現於客戶關係管理、網站入侵偵測、醫學病歷紀錄、氣象預報分析、財務服務推薦、教學現場分析、基因定序關聯性等多樣性的應用（Pruengkarn et al., 2017）。

一般來說，關聯規則演算模式可以分為 3 大類（Chee et al., 2019）：（1）合併為基演算法（join-based algorithms）是以不斷將標的加入項目集中掃描並得出那些滿足相關聯之最小支持閾值；（2）樹狀為基演算法（tree-based algorithms）是將探勘標的以樹狀圖方式不斷排序及逐次修剪，並得出相關聯之最小支持閾值；（3）樣型成長演算法（pattern growth algorithms）以 FP-growth 演算法為基礎演化而來，主要是先建構條件 FP 樹及以後綴模式對資料庫的串聯掃描生成的，而因頻繁項目集（frequent itemsets）都是掃描在

頻繁模式的對應路徑中的樹，並以此得出那些滿足相關聯之最小支持閾值。各項關聯類型及代表性演算法如圖 2 所示。

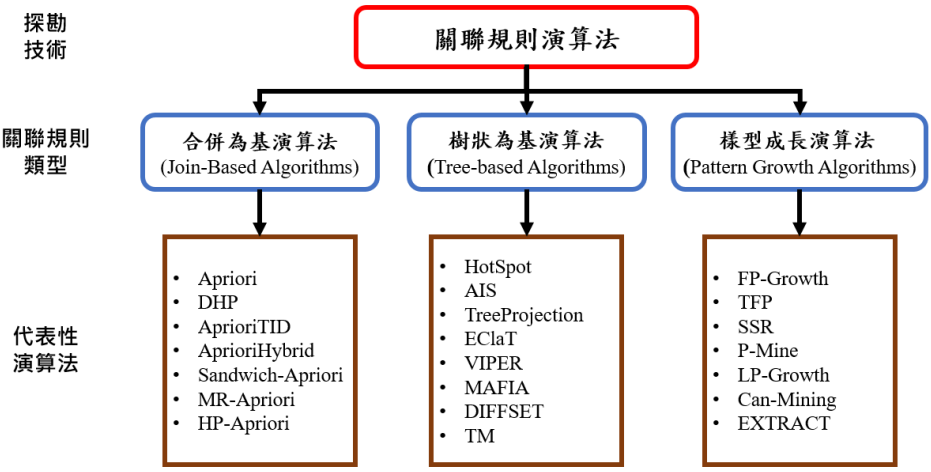


圖 2　關聯演算法的分類圖
資料來源：改繪自 Chee et al.（2019）

合併為基演算法（join-based algorithms）利用項目組合來生成頻繁項目集中的候選者集合，進而發現項目間的關聯性（Agrawal et al., 1993）。其演算法包含 Direct Hashing and Pruning（DHP）及 Apriori 的各式改良演算法，例如：AprioriTID、AprioriHybrid、Sandwich-Apriori、MapReduce-Apriori（MR-Apriori）、Horizontal Parallel-Apriori（HP-Apriori）等。其中最具有代表性且運用最廣泛的演算法即為 Apriori 演算法。

樹狀為基演算法（Tree-based Algorithms）基於集合概念列舉並假定多面向策略（如以深度或廣度為優先搜尋方式）探索候選項目集的樹狀結構資料，而為了識別此候選項目集之關聯性，進一步生成字典樹或列舉樹等點陣式圖型；樹狀為基演算法以生成中的項目間的排列順序及建構方式，來推論得出隱藏的關聯性（Borah et al., 2019）。樹狀為基演算法包含 Hotspot、Artificial Immune System(AIS)、TreeProjection、Equivalence CLAss clustering and bottom-up lattice Traversal(EClaT)、VIPER、MAFIA、DIFFSET 及 Transaction Mapping（TM）等多個演算法。

樣型成長演算法（pattern growth algorithms）：在現有的樣型成長演算法中，大部分是由 FP-growth 模型算法演變而來的。這是因為 FP-growth 僅使用對數據集進行兩次掃描，用壓縮樹結構表示整個數據集，並通過消除生成候選項目集的需要來減少執行時間（Mittal et al., 2015）。樣型成長是透過條件 FP 樹及以後綴模式對資料庫的串聯掃描生成的，而由於全部頻繁項目集都是掃描在頻繁模式的對應路徑中，此模式保證了樹的完整性結果（Chee et al., 2019）。

三、研究方法與架構

3.1 研究架構與流程設計

本研究以次級資料分析方法為基礎，參照相關學者觀點，並融合本研究標的之設計，將分析步驟區分為資料蒐集/彙整、關聯探勘演算運用、成果分析與建議三大階段（詳如

圖 3），並簡述研究步驟如下：

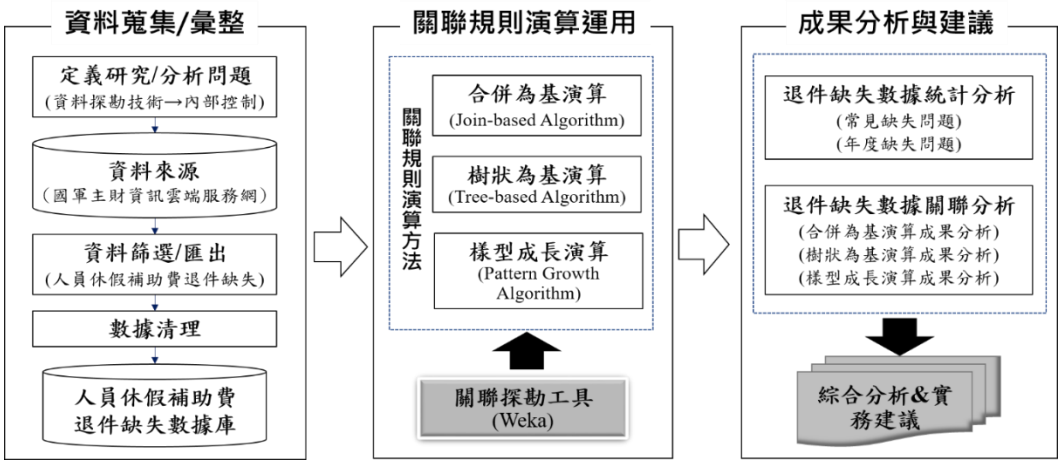


圖 3 主財資訊數據探勘架構與流程

一、 資料蒐集/彙整：以國軍主財資訊雲端服務網的薪餉次系統為主要目標，蒐集其中的休假補助費發放數據中的退件缺失態樣，並將資料匯出準備分析。

二、 關聯規則演算運用：根據本研究缺失態樣的研究標的資料型態，分別以 Chee et al.（2019）提出的合併為基演算法（join-based algorithms）、樹狀為基演算法（tree-based algorithms）、樣型成長演算法（pattern growth algorithms）等三類關聯規則探勘方法，並選取代表性演算法進行探勘分析，關聯運算則以開源探勘軟體 Weka 為主要分析工具。

三、 成果分析與建議：分別針對休假補助費退件缺失數據進行統計分析及關聯分析作業，並提出綜合分析與實務建議。

3.2 研究流程說明

一、 資料蒐集/彙整

本研究以國軍主財資訊雲端服務網為研究標的，該平台係以國軍雲端服務為基礎整合各類主財業務資訊系統，提供全國各財務、業務及支薪等單位使用。然鑒於國軍主財資訊雲端服務網系統相關資料涉及個人隱私且具機敏性，同時考量數據分析的清晰度，本研究僅蒐集薪餉次系統下之休假補助費退件缺失態樣為探勘範例資料集，並將部分資料轉換為流水代碼進行後續分析。然而，因休假補助費資料於民國 105 年 1 月才逐步建置，故將篩選民國 105 年迄 111 年 12 月間，經查核具有缺失項而遭退件之資料紀錄，總計蒐整 20,033 件。休假補助費缺失共分為代碼 A（申請名冊）、代碼 B（申請表）、代碼 C（憑證）三大類，同一大類代碼均以連續編號表示之。休補費缺失代碼摘要如表 1。

休假補助費缺失相關樣態數據，係由日常辦理申領作業所檢附之紙本資料轉置並記錄，形成休假補助費退件缺失態樣檢核紀錄，同時透過代碼及布林值型態之轉換，形成休補費退件缺失的二元邏輯關係，以利後續探勘分析。

表 1 休假補助費缺失資料摘要說明

缺失項目類別	缺失項目個數	缺失代碼	缺失態樣（範例）
申請名冊（A）	15	A01~A15	名章不符、階級不符、身份證字號不符、申請金額不符、檢附單據張數/金額不符等缺失
申請表（B）	20	B01~B20	名章不符、身分證字號不符、申請日期不符、休假日期不符、塗改漏蓋主官章、年度內已超支等缺失
憑證（C）	17	C01~C17	收據買受人地址漏填、申領人單據未簽名、單據數量、單價與總價不符、單據日期非休假日等缺失

二、 關聯規則演算運用

關聯規則演算通常植基於兩項重要的指標：支持度（support）與信賴度（confidence）。若 X,Y 被視為交易資料庫中的物件項目，則支持度係指在所有交易數據紀錄中，同時出現{X, Y}物件項目所佔的比率；而信賴度是定義此關聯法則可以信賴的程度，也就是在發生 X 物件項目的條件下，也會發生 Y 物件項目所佔的比率（Harikumar and Dilikumar, 2016）。計算公式分別如（1）（2）所示（Lin and Tseng, 2006; Manimaran and Velmurugan, 2015）：

$$\text{Support}(X \rightarrow Y) = P(X \cup Y) \quad (1)$$

$$\text{Confidence}(X \rightarrow Y) = \text{Support}(X \cup Y) / \text{Support}(X) \quad (2)$$

接著說明演算規則的共通性術語（陳垂呈，2017）：（1）滿足最小支持度的項目集，稱為頻繁項目集（frequent itemset）；（2）若一個項集包含 k 個項目，則稱之為 k-項目集（k-itemsets）；（3）若某 k-項集滿足最小支持度，則稱之為頻繁 k-項目集（frequent k-itemsets）。

本研究參採學者 Chee et al.（2019）提出的合併為基、樹狀為基、樣型成長等三類型的關聯演算模式進行探勘分析，並嘗試以休補費退件缺失項目為範例，分別說明不同演算機制之探勘流程，同時介紹代表性演算法。

(一) 合併為基演算方法：Apriori

Apriori 演算法涉及挑選出未知的相互依存關係數據並找出這些項目之間的規則，以萃取出各種記錄的關聯，其係為合併為基關聯規則的代表性探勘方法。Apriori 關聯規則演算的探勘流程如下（曾憲雄等，2007；Agrawal and Srikant, 1994），同時以範例呈現相關運算邏輯（如圖 4）。

1. 計算各缺失項目的支持度及信賴度，同時給定關聯規則的最小支持度（minimum support）與最小信賴度（minimum confidence）數值。本範例以缺失項目若出現 2 次（含）以上（Min_Support = 2）為支持度的門檻值。
2. 讀取資料集中所有的記錄，分別以 1-項目計算支持個數所形成的項目集列為候選項目集合（candidate itemset），若候選項目集合的支持度高於設定的最小支持度（即刪

除個數少於最小支持度的項目集），則該候選項目集合視為頻繁項目集合（frequent itemset）。本範例資料集數據均符合標準，故候選 C1 即視為頻繁項目集合 L1。

3. 進行頻繁項目集合 L1 的結合，產生候選集合 C2；接著掃描資料集並計算每一個候選 2-項目集的支持度，刪除支持個數少於最小支持度的項目集，產生頻繁項目集 L2。。
4. 持續利用 L2 的結合來產生候選集合 C3，並重覆掃描資料集計算候選 3-項目集的支持度，依條件產生新的頻繁項目集，直到沒有新的候選項目集合為止。
5. 計算 頻繁 k-項目集所形成的關聯規則，若滿足最小信賴度，則關聯規則成立。

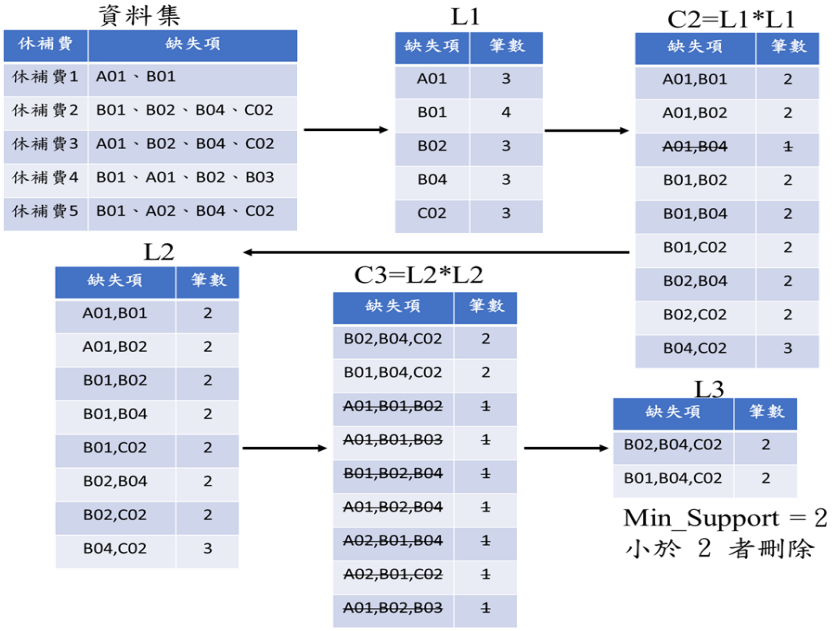


圖 4 Apriori 演算運用於休補費之執行流程（範例）
資料來源：陳良駒與傅祖銘（2022）

(二) 樹狀為基演算方法：Hotspot

HotSpot 演算法係參照 Friedman and Fisher（1999）的理論所開發的 Weka 關聯規則擴充套件，其可針對離散或類別數據進行探勘以滿足最小支持度的關聯區間，並將探勘結果以樹狀結構的形式進行規則的呈現（Qing-dao-er-ji et al., 2020）。HotSpot 演算法可依據使用者感興趣的目標項目（target item）為根節點，在最小支持度的限制下搜尋機率最大的其他項目，依序列出與目標項目中出現頻率最高的相關項目做為第一層的項目節點，接著透過最大分支度（Max Branching Factor, MBF）的演算產生最大化或最小化的分支，以發掘與目標項目對應的前提規則集（Left Hand Side, LHS），同時建立一套樹狀結構的規則。

本研究將對 Hotspot 演算法中樹狀節點的 MBF 設到最大，並以支持度及信賴度排序得出頻繁關聯程度。其簡易執行流程如圖 5 所示。

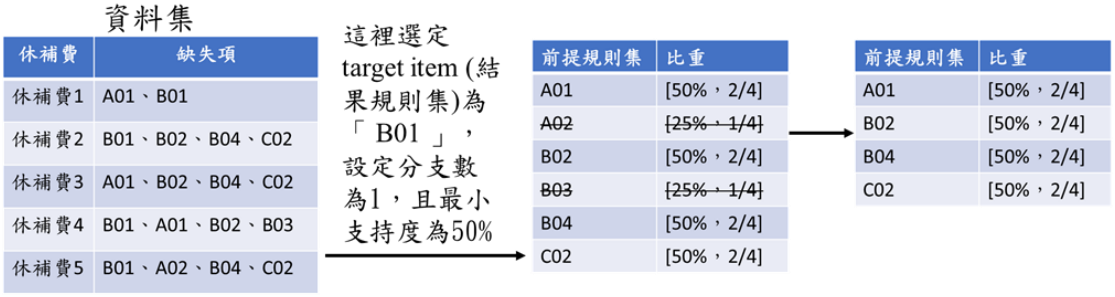


圖 5 Hotspot 演算運用於休補費之執行流程（範例）

(三) 樣型成長為基演算方法：FP-growth

FP-growth 是樣型成長為基關聯規則中最常見的演算類型(Borgelt, 2005)，FP-growth 分為兩個演算過程（Kumar et al., 2010）：首先根據原始資料建構 FP 樹（FP-Tree），然後在 FP 樹上擷取頻繁關聯模式。該方法只需要掃描資料庫兩次，可改善 Apriori 演算法在執行時需不斷掃描資料庫以產生大量候選項目集，而導致探勘效率不佳之問題。相關演算探勘流程如下說明（Han et al., 2004），同時以圖 6 為範例進行簡易說明。

1. 建構 FP-tree：將資料集壓縮成一個緊湊的頻繁模式樹結構。
   (1). 設定資料集及最小支持度後，掃描該資料集，並獲得頻繁 1-項目集和支持度。
   (2). 對頻繁 1-項目集按支持度排序，並將支持度小於最小支持度的項目刪除，獲得頻繁項表來建立一個以 null 值為根節點的 FP-tree。
2. 從 FP-tree 挖掘關聯規則：開發一種基於 FP-tree 的高效頻繁模式探勘方法。
   (1). 第二次掃描整個資料庫，並依照先前排序的順序建立成 FP-tree 的資料結構。
   (2). 之後演算法即可透過 FP-tree 反覆運算找出所有的大型項目集。

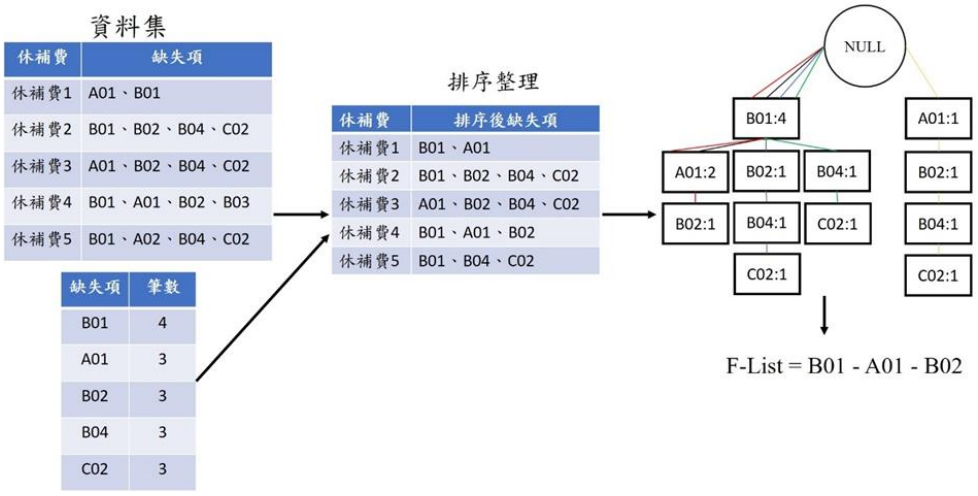


圖 6 FP-growth 演算運用於休補費之執行流程（範例）

三、 成果分析與建議：

研究成果主要涵蓋「退件缺失數據統計分析」及「退件缺失數據關聯分析」兩大類。前者包括常見缺失問題研討與年度缺失問題統計，後者則聚焦合併為基、樹狀為基、樣型成長等三種關聯規則演算之成果，最後透過綜合分析並提出實務建議。

四、研究成果與分析

4.1 缺失態樣年度統計資訊

本研究之缺失態樣係由 105 年開始累計迄 111 年底，相關缺失次數統計如表 2 及圖 7 所示。年度內常見缺失態樣數量多無太大變化，僅申請名冊（A）部分略有逐年增加隨後持穩的情形，申請表部分（B）及憑證部分（C）則有逐年下降的趨勢；此外，申請表部分（B）缺失問題數量普遍高於其他類型缺失；而總缺失筆數則以申請表部分（B）之問題最多，其次為憑證部分（C），申請名冊部分（A）的缺失數量相對較少。

表 2 年度缺失項目之樣態數量統計表

缺失項目	缺失態樣年度合計筆數							
	105	106	107	108	109	110	111	合計
申請名冊部分（A）	25,351	25,318	27,820	28,227	28,819	27,714	27,277	190,526
申請表部分（B）	37,141	35,642	34,348	34,827	33,617	32,531	32,480	240,586
憑證部分（C）	29,992	29,643	29,389	28,842	28,284	27,343	26,121	199,614

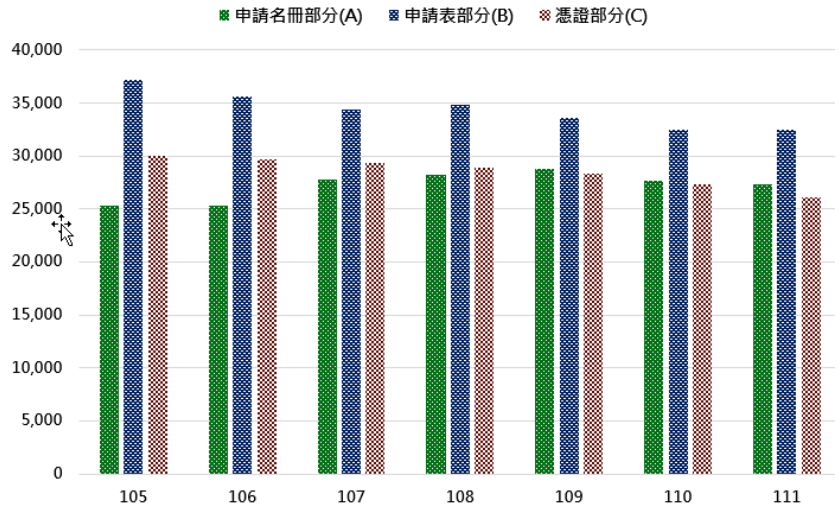


圖7　年度缺失項目之樣態數量統計圖

經統計彙整歷年最常見的前 20 項樣態缺失相關數據（詳如表 3），發現休假日數不符（申請名冊部分）（A05）、休假日為國定休假日（B14）、收據買受人、買受人地址漏填（C01）分別為發生頻率最高的前三項缺失。若以缺失代碼來看，A 類申請名冊部分的缺失問題最多（8 項），B 類申請表部分（6 項）及 C 類憑證部分（6 項）的缺失數量相當，綜合來看各類的缺失分布較為平均；若以缺失態樣來觀察，休假日數不符、名章

不符及其他等缺失問題在 A 類申請名冊與 B 類申請表均同時出現，顯示這些項目確實為經費申請人及業務承辦者極為容易忽略之問題。此外，C 類憑證部分則主要發生於資料漏填、單據金額及發票日期不符規定等缺失問題。相關缺失應要求各單位注意此缺失樣態之發生。

表 3 歷年常見之前十項缺失態樣

排名	缺失代碼	缺失態樣	合計筆數
1	A05	休假日數不符（申請名冊部分）	18,846
2	B14	休假日為國定休假日	17,672
3	C01	收據買受人、買受人地址漏填	17,256
4	B04	休假日數不符（申請表部分）	17,130
5	A04	申請金額不符	17,111
6	B01	名章不符（申請表部分）	16,576
7	C05	單據金額未達申領數二分之一	16,567
8	A01	名章不符（申請名冊部分）	16,503
9	A15	其他（申請名冊部分）	15,841
10	A09	合計大寫金額不符	15,802
11	A02	階級不符	15,560
12	A03	身分證字號不符	15,085
13	C11	收據未註明負責人姓名	15,069
14	C08	一次開立多聯發票有缺漏	14,876
15	B20	其他（申請表部分）	14,841
16	C13	發票日期非當期之月份	14,784
17	B03	申請日期不符	14,402
18	C02	申領人單據未簽名	13,757
19	B09	申領破月人員未註明原因及加蓋私章	13,663
20	A07	檢附單據金額不符	13,521

4.2 缺失態樣探勘分析

因配合研究期程，本論文以民國 105-110 間的年度缺失項目為基礎，透過不同演算模型進行缺失樣態探勘分析，並進行各項缺失關聯規則之說明，分述如下。

一、合併為基 Apriori 演算成果：

透過前述的樣態格式轉換及 Apriori 關聯規則算法之流程，探勘相關成果總計有 56 組配對規則。其中前三組經常出現之規則為 B12（申請年度、日期不符）→C14（電子發票不得有統一編號）、C15（買受人非當事人）→B19（申領金額錯誤）及 B19（申領金額錯誤）→ C15（買受人非當事人）。第一組顯示規則型態為當 B12 出現時，有較高機率出現 C14 的錯誤缺失樣態；而第二、三組則呈現雙向之規則，顯示這兩種錯誤經常共同發生。

此外，觀察並彙整所有配對組之關聯規則，發現有 7 組頻繁項目集的缺失組合（信賴度均為 95%以上），表 4 呈現各項缺失組合的形式與說明，其中除項次 6 為同類型（A）缺失樣態外，其餘規則組均為跨類型的關聯規則型態。

表 4　Apriori 頻繁群集彙整

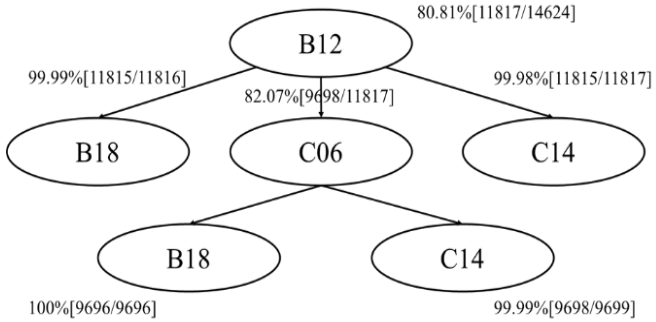
項次	缺失代碼及規則（前 3 項）	關聯說明
1	B12、B18、C14 ● B12→C14 ● B18→C14 ● （B12, B18）→C14	「申請年度、日期不符」、「檢附單據張數錯誤」及「電子發票不得有統一編號」具高度關聯性，由此可知申請表申請日期錯誤容易伴隨申請表檢附單據張數錯誤及發票憑證錯誤。
2	B09、C08、C09 ● B09→C09 ● B09→C08 ● （B09, C08）→C09	「申領破月人員未註明原因及加蓋私章」、「一次開立多聯發票有缺漏」及「未依規定開立發票」具高度關聯性，可知申請表申領破月人員未加註原因及未加蓋私章容易伴隨憑證發票檢附錯誤。
3	A14、C16 ● A14→C16 ● C16→A14	「支薪單位全銜或代號錯誤」及「未註明中文品名」具高度關聯性，由此可知申請名冊支薪單位全銜或代號錯誤容易伴隨憑證空白發票未註明中文品名。
4	B11、B19、C15 ● C15→B19 ● B19→C15 ● B11→B19	「支薪單位全銜錯誤」、「申領金額錯誤」及「買受人非當事人」具高度關聯性，由此可知申請表支薪單位全銜或申領金額錯誤容易發生憑證買受人非當事人。
5	A07、B06、B07 ● B06→A07 ● B07→A07 ● （B06, B07）→A07	「檢附單據金額不符」、「合計大寫金額不符」及「服役年資請重新計算」具高度關聯性，由此可知申請名冊造冊金額錯誤容易發生申請表服役年資計算錯誤。
6	A06、A12 ● A06→A12 ● A12→A06	「檢附單據張數不符」及「造列之表冊份數不足」具高度關聯性，由此可知申請名冊填製單據張數錯誤容易發生申請名冊檢附份數不足情事。
7	A13、C07 ● A13→C07 ● C07→A13	「非為在職期間發生之事實」及「發票未附收執聯」具高度關聯性，由此可知申請名冊在職事實錯誤容易發生憑證發票未附收執聯情事。

二、樹狀為基 Hotspot 演算成果

由於 Hotspot 演算法係以目標項目為探勘標的，並以此找出目標項目對應的最常出現的對應前提規則。因此，本研究分別以時間缺失 B12（申請年度、日期不符）及人員缺失 C15（買受人非當事人）為範例，進行 Hotspot 演算之成果分析，其成果如下說明。

(一) 目標項目：時間缺失 B12（申請年度、日期不符）

圖 8 自葉節點 B18 往上看，退件單含有 B18 缺失項有 100%（9696/9696）會有 C06 缺失項，而進一步看同時含 B18、C06 及 B12 缺失項共有 9696 筆，另外父節點的部分以 C06 為例，退件單含 C06 缺失項者有 82.07%（9698/11817）會有 B12 缺失項，綜上可知「B12 申請年度、日期不符」、「B18 檢附單據張數錯誤」及「C14 電子發票不得有統一編號」具有高度關聯性，建議做為爾後內部稽核重點檢核項目。



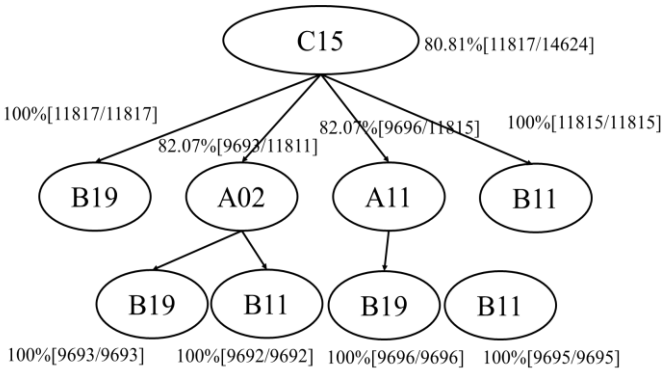
B12=1（80.81% [11817/14624]）
B18 = 1（99.99% [11815/11816]）
C14 = 1（99.98% [11817/11819]）
C06 = 1（82.07% [9698/11817]）
| B18 = 1（100% [9696/9696]）
| C14 = 1（99.99% [9698/9699]）

（a）樹狀結構圖　　（b）數據呈現邏輯

圖 8　B12 關聯（a）樹狀結構圖（b）數據邏輯呈現結果

(二) 目標項目：人員缺失 C15（買受人非當事人）

根節點以 C15 為例，退件單含有 B19 缺失項有 100%（9693/9693）會有 A02 缺失項，而進一步看同時含 B19、A02 及 C15 缺失項共有 9693 筆，另外父節點的部分以 A02 為例，退件單含 A02 缺失項者有 82.07%（9693/11811）會有 C15 缺失項，綜上可知「C15 買受人非當事人」、「B19 申領金額錯誤」及「B11 支薪單位全銜錯誤」具有高度關聯性，成果展現如圖 9，建議做為爾後內部稽核重點檢核項目。



C15=1（80.81% [11817/14624]）
B19 = 1（100% [11817/11817]）
B11 = 1（100% [11815/11815]）
A02 = 1（82.07% [9693/11811]）
| B19 = 1（100% [9693/9693]）
| B11 = 1（100% [9692/9692]）
A11 = 1（82.07% [9696/11815]）
| B19 = 1（100% [9696/9696]）
| B11 = 1（100% [9695/9695]）

（a）樹狀結構圖　　（b）數據呈現邏輯

圖 9　C15 關聯（a）樹狀結構圖（b）數據邏輯呈現結果

三、樣型成長為基演算法：FP-growth

依據演算步驟建構 FP 樹，並從其中擷取頻繁關聯規則後，發現符合 FP-growth 演算模型探勘之關聯規則計有 60 項，表 5 呈現前 10 項的重要關聯， 分別代表前提規則集（Left Hand Side, LHS）的出現會同時出現結果規則集（Right Hand Side, RHS）的關聯。以前三項關聯規則為例，分別出現申請年度/日期不符（B12）→電子發票不得有統一編號（C14）、檢附單據張數錯誤（B18）→ 電子發票不得有統一編號（C14）、一次開立多聯發票有缺漏（C08）→ 未依規定開立發票（C09）等關係，顯示若前提規則集 LHS（B12, B18, C08）之缺失發生時，結果規則集 RHS（C14, C14, C09）的缺失項目同步發生的機率極高。因此，在執行內部稽核作業時，可注意同步檢查各項具關聯性的缺失項目，以確認休補費紀錄的正確性，同時提升各項作業的執行效率。

表 5　FP-growth 模型執行結果彙整表（前 10 項）

項次	LHS	RHS	關聯性
1	B12	C14	申請年度/日期不符 →電子發票不得有統一編號
2	B18	C14	檢附單據張數錯誤 → 電子發票不得有統一編號
3	C08	C09	一次開立多聯發票有缺漏 → 未依規定開立發票
4	B09	C09	申領破月人員未註明原因及加蓋私章 → 未依規定開立發票
5	A14	C16	支薪單位全銜或代號錯誤→未註明中文品名
6	C15	B19	買受人非當事人→申領金額錯誤
7	B19	C15	申領金額錯誤→買受人非當事人
8	B11	C15	支薪單位全銜錯誤→買受人非當事人
9	B11	B19	支薪單位全銜錯誤→申領金額錯誤
10	A02	A11	階級不符→塗改漏蓋主官章

4.3　綜合分析

一、 頻繁規則成果比較

研究結果通過關聯規則分析（Apriori 及 FP-growth）執行頻繁項目集分析，並分別綜合出 7 組及 9 組頻繁項目集（信賴度均為 95%以上），其中兩個關聯規則所挖掘的頻繁項群集有 7 組相同，而差異出現於（A02、A11）及（B10、B15）兩組關聯規則，分別呈現「階級不符」及「塗改漏蓋主官章」具高度關聯性、「年度內已超支」及「休假日數合計不符」具高度關聯性的特性。究其原因係為 FP-growth 演算法在第一階段 FP 樹中即完成此兩項組合的編排，隨後第二階段的掃瞄是依據第一階段的 FP 樹來形成；而 Apriori 掃描資料的過程中則是不斷重複掃描直至頻繁項目集出現，造成頻繁項次規則的差異。兩類關聯規則探勘結果之差異如表 6 所示。

表 6　Apriori 與 FP-growth 挖掘頻繁項差異比較

項次	缺失代碼及規則	關聯規則類型	
		合併為基演算（Apriori）	樣型成長為基演算（FP-growth）
1	B12、B18、C14	**V**	**V**
2	B09、C08、C09	**V**	**V**
3	A14、C16	**V**	**V**
4	B11、B19、C15	**V**	**V**
5	A07、B06、B07	**V**	**V**
6	A06、A12	**V**	**V**
7	A13、C07	**V**	**V**
8	A02、A11	-	**V**
9	B10、B15	-	**V**

二、 頻繁樣式探勘演算法比較

依據國軍主財資訊雲端服務網的資料量及伺服器運算效能及研究探勘的過程，以 FP-growth 演算法來探勘頻繁項效果最佳，其執行速度快，對比 Apriori 演算法對資料型態要求更低，減少探勘目標事前的清洗及彙整時間，另外 Hotspot 演算法則適用於針對已知單一探勘標的來做頻繁程度分析，可用於特殊任務及個案需求上，最後就各演算法研究分析後發現的特點實施說明，相關資料如表 7。

表 7 各演算法成果與優缺點分析

類型	合併為基演算	樹狀為基演算	樣型成長演算
代表性演算法	Apriori	Hotspot	FP-growth
發現規則組數	56 組	依目標項目之頻繁關聯程度而定	60 組
特性	廣度優先（建立 K 項目集→掃描資料庫）	-	深度優先（掃描資料庫→發現頻繁項目集）
優點	執行邏輯簡單，適用於大量資料庫	● 可以處理多維度/數值型的連續資料 ● 聚焦研究標的執行關聯規則分析技術	● 不需要產生候選集 ● 只需對資料庫掃描 2 次來產生並壓縮至 FP-Tree ● 執行速度快
缺點	● 過程產生大量候選集，執行效率較慢 ● 複雜的資料型態，運算效果較差	只針對單一目標項目關聯性做分析，對大量或複雜的資料型態較難處理，侷限性大	● 利用 FP-Tree 壓縮數據階段，內存記憶體負載消耗大 ● 建立 FP-Tree 壓縮數據階段對效能要求高

資料來源：Singh et al.（2014）；本研究整理

(一) 合併為基演算的特性：代表演算法為 Apriori。透過該方法探勘出的關聯計有 7 大類 56 組規則，由上述成果分析結論可知因該演算法邏輯程序簡單，較適用在純文字或標準化數值型態的資料下執行探勘運算，但因需產生大量候選集進行篩選，故執行效率相對較差。若未來採用該演算法執行探勘任務，可將重點放在探勘標的資料型態之數據清洗及態樣呈現等階段。

(二) 樹狀為基演算的特性：代表演算法為 Hotspot。該演算法僅針對單一目標關聯性作分析，可針對多維度的連續數值進行探勘，故較適合想要探索特定品項關聯之情境。若以此演算方法執行關聯探勘分析，重點與合併為基演算雷同，唯可發揮其探勘優點，聚焦在感興趣的探勘標的來實施關聯分析，同時結合樹狀圖展現其相關程度，對於關聯的呈現能力強。

(三) 樣型成長演算的特性：代表演算法為 FP-growth。本研究探勘出的關聯計有 9 大類 60 組規則，由上成果分析結論可知雖然內存記憶負載大，但因其執行速度快且非常適用於特徵關聯探勘的特性，故相當適合執行本研究資料探勘任務。建議未來針對國軍主財資訊雲端服務網相關數據分析，可利用該演算法來進行各項探勘分析。

五、結論與建議

本研究針對不同類型之探勘方法應用於休假補助費退件數據進行關聯規則之探索，並分別比較各類型演算規則及探勘成果之差異。經數據分析後有數點綜合性的發現：（一）以缺失類型而言，申請表（B）問題的缺失數量最多，普遍高於申請名冊（A）及憑證（C）等問題的缺失量；若以缺失態樣來看，則主要以休假日期或天數等問題發生錯誤的頻次較高。（二）合併為基及樣型成長兩類演算方法所挖掘出的重要規則大致趨於一致，但樣型成長探勘技術在類似條件下，可以發現較多的關聯規則。（三）樹狀為基的演算方法可聚焦於興趣或特定項目之規則挖掘，適合針對特定缺失項目進行關聯分析。

國軍預算支用單位及支薪單位主財資訊雲端服務系統執行各項財務業務任務，其資料處理所產生之業務數據量龐大，若能藉由數據探勘及分析技術之運用，進行交叉比對，應可有效了解國軍財務各項作業之實況，據以協助財務單位各項管控任務及審查發放等作業之遂行。綜上所述，分別提出值得進一步探討的數點建議。

一、演算機制運用之建議：

(一) 雲端主財業務所產生的數據量極為龐大，建議應針對不同子系統之數據，利用合併為基與樣型成長為基等演算方法進行定期性的數據分析，以發掘不同時期的關聯特性，了解缺失項目的樣態變化。

(二) 而若發現某些缺失項目頻繁產生，或因數據可處理時間有限等狀況，則可聚焦於興趣或重點目標之缺失問題，以樹狀為基的演算方法進行稽核規則之探索。

二、實務操作運用之建議：

(一) 由缺失問題關聯性來看，大致上可以區分為發票內容缺失（例如：電子發票不得有統一編號）、行政業務缺失（例如：檢附單據金額不符、合計大寫金額不符等）、承辦人員缺失（例如：身份證字號或階級等資料不符）等面向之問題。其衍生的共現關聯多半呈現跨行政/承辦及發票內容的現象。故除積極訓練業務承辦人熟悉相關規定外，也要主動制定結報發票內容注意事項供申請單位（人）參照，以事前避免發票開立缺失等現象，進而降低各項申報作業之缺失問題。

(二) 由數據所挖掘出來的各項缺失規則、分類特性等問題，可做為爾後內部稽核教育訓練之重點宣教項目，或依據國軍財務單位特性，建立填表流程及檢附資料須知等檢查機制，協助財務收支、會計帳務或資訊稽核人員快速發現財務行政之缺失。

三、未來建議：

(一) 雲端主財業務與國軍部隊維運有著重要的連結關係，國軍主財資訊雲端服務網的資料庫仍然在成長階段，其所涉及的系統類型與數據量，均可能隨著部隊運作的複雜性而需要大幅增加。本研究僅聚焦於關聯規則的探勘與分析，如何運用不同的資料探勘或人工智慧相關技術來探索多元的數據分析服務，以強化內部稽核作業，是未來可持續投入研究的議題。

(二) 未來國軍主財資訊雲端服務網相關子系統將陸續建置及整合，建議針對主財資訊雲端服務系統建立自動化的稽核程序，同時運用電腦系統進行智慧型防呆及檢核服務之設計，以期能有效預警執行單位，避免違失案件產生。

六、國防領域之實務應用

在數據科技發展的時代，針對大量資料的分析與自動化的稽核程序就相當重要。國軍主財資訊雲端服務網係提供國軍各單位便利性化的主財作業介面與資訊服務，目前系統資料正不斷的成長中。主財系統數據可透過各式資料探勘技術，挖掘資料間的潛在特性，進而協助稽核人員有效分析數據或查察異狀，以利及時而有效的提供預警，避免違失案件產生。本研究從合併為基、樹狀為基、樣型成長等三類關聯規則進行國軍休補費數據之挖掘，經實證可發現多組相關聯的缺失態樣。本論文為應用實務研究，成果可提供單位進行休補費申請時之參照，亦可擴展至主財雲端平台的其他系統數據進行全方位的分析，提供稽核單位進行缺失檢核之依據。

誌謝

感謝匿名審查委員的諸多寶貴意見，使本論文之內容更臻完善；本研究承蒙行政院國科會專題研究計畫經費支持（計畫編號：MOST 109-2410-H-606-008-MY2），謹致謝忱。

參考文獻

王威智（2004）。*內部稽核、內部控制與採購作業之關聯性研究-以我國公務機關為例*。國防大學資源管理研究所未出版碩士論文，臺灣，臺北市。

石淑暖（2016）。主計資訊系統導入持續性稽核技術研究成果之應用。*主計月刊*，726，94-99。

林永能（2015）。國軍主財資訊雲端服務網整合發展現況，*主計季刊*，*56*（1），60-67。

黃士銘、周玲儀、黃秀鳳（2012）。最新文字探勘技術於稽核上的應用。*會計研究月刊*，（323），112-119。

陳浚明（2009）。*國軍管理階層影響主計人員執行內部審核因素之研究*。輔仁大學應用統計學研究所未出版碩士論文，臺灣，新北市。

陳垂呈（2017）。高效率 Apriori 演算法探勘關聯規則。*資訊與管理科學*，*10*（2），21-29。

陳良駒、傅祖銘（2022）。國軍人員休假補助費缺失樣態之探勘與分析。*第 30 屆國防管理學術暨實務研討會*，臺灣，臺北市。

傅祖銘（2020）。內部稽核運用資料探勘 FP-growth 演算法之研究-以國軍人員申領休假補助費缺失態樣為例。*主計季刊*，*61*（4），61-73。

孫嘉明、邱靜宜、林宜隆（2017）。持續性稽核技術整合架構-以主計資訊系統為例。*電腦稽核*，（35），80-95。

曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯（2007）。*資料探勘*。臺北市：旗標。

諶家蘭（2015）。風險為基礎之持續性稽核與監控。*主計月刊*，（719），36-40。

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large database. *Proceedings of the 20th International Conference on Very Large Data Bases*, 487-499.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data,* 207-216.

Albashrawi, M. (2016). Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015. *Journal of Data Science*, *14*(3), 553-569.

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402.

Angeline, D. M. D. (2013). Association rule generation for student performance analysis using apriori algorithm. *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, *1*(1), 12-16.

Bai, B., Yen, J., & Yang, X. (2008). False financial statements: characteristics of China's listed companies and CART detecting approach. *International Journal of Information Technology & Decision Making*, *7*(2), 339-359.

Borah, A., & Nath, B. (2019). Tree based frequent and rare pattern mining techniques: a comprehensive structural and empirical analysis. *SN Applied Sciences*, *1*(9), 1-18.

Borgelt, C. (2005). An implementation of the FP-growth algorithm. *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, 1-5.

Bose, S., Dey, S. K., & Bhattacharjee, S. (2022). Big data, data analytics and artificial intelligence in accounting: An overview. In Akter, S. & Wamba, S. F. (Eds.). *Handbook of big data methods* (1-34). United Kingdom: Edward Elgar Publishing.

Boskou, G., Kirkos, E., & Spathis, C. (2018). Assessing internal audit with text mining. *Journal of Information & Knowledge Management*, *17*(2), 1850020.

Capriotti, R. J. (2014). Big Data bringing big changes to accounting. *Pennsylvania CPA Journal*, *85*(2), 36-38.

Chee, C. H., Jaafar, J., Aziz, I. A., Hasan, M. H., & Yeoh, W. (2019). Algorithms for frequent itemset mining: A literature review. *Artificial Intelligence Review*, *52*(4), 2603-2621.

Earley, C. E. (2015). Data analytics in auditing: Opportunities and challenges. *Business Horizons*, *58*(5), 493-500

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, *13*(3), 57-57.

Friedman, J. H., & Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, *9*(2), 123-143.

Han, J., Pei, J., Yin, Y. & Mao, R., (2004). Mining frequent patterns without candidate generation: a frequent-pattern tree approach. *Data Mining and Knowledge Discovery, 8*(1), 53-87.

Harikumar, S., & Dilipkumar, D. U. (2016). Apriori algorithm for association rule mining in high dimensional data. *In 2016 International Conference on Data Science and Engineering (ICDSE)*, 1-6.

Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(3), 139-154.

Islam, M. S., Hasan, M. M., Wang, X., & Germack, H. D. (2018). A systematic review on healthcare analytics: application and theoretical perspective of data mining. *Healthcare*, *6*(2), 54.

Keyvanpour, M. R., Javideh, M. & Ebrahimi, M. R. (2011). Detecting and investigating crime by means of data mining: A general crime matching framework. *Procedia Computer Science*, *3*, 872-880.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, *32*(4), 995-1003.

Köksal, G., Batmaz, I., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, *38*(10), 13448-13467.

Kumar, B. S., & Rukmani, K. V. (2010). Implementation of web usage mining using Apriori

and FP growth algorithms. *International Journal of Advanced Networking and Applications*, *1*(6), 400-404.

Lin, W. Y., & Tseng, M. C. (2006). Automated support specification for efficient mining of interesting association rules. *Journal of Information Science*, *32*(3), 238-250.

Manimaran, J., & Velmurugan, T. (2015). Analysing the quality of association rules by computing an interestingness measures. *Indian Journal of Science and Technology*, *8*(15), 1-12.

Mittal, A., Nagar, A., Gupta, K., & Nahar, R. (2015). Comparative study of various frequent pattern mining algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(4), 550-553.

Nan, N. (2022). Integration and development of enterprise internal audit and big data based on data mining technology. *Computational Intelligence and Neuroscience*, *2023*, 9825603.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559-569.

Pruengkarn, R., Wong, K. W., & Fung, C. C. (2017). A review of data mining techniques and applications. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *21*(1), 31-48.

Qing-dao-er-ji, R., Pang, R., & Chang, Y. (2020). An improved HotSpot algorithm and its application to sandstorm data in Inner Mongolia. *Mathematical Problems in Engineering*, *2020*, 4020723.

Ren, L., & Chen, Y. (2021). Research on the application of data mining technology in military audit. *Proceedings of the 2021 International Conference on Education, Information Management and Service Science (EIMSS),* 277-283.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1355.

Saglar, J., & İlker, K. E. F. E. (2021). A review on data mining methods used in internal audit and external audit. *Ekev Akademi Dergisi*, *88*, 259-274.

Shabani, N., Munir, A., & Mohanty, S. P. (2021). A study of big data analytics in internal auditing. *Proceedings of the 2021 Intelligent Systems Conference*, 362-374. Springer International Publishing.

Shan, R., Xiao, X., Che, J., Du, J., & Li, Y. (2022). Data mining optimization software and its application in financial audit data analysis. *Mobile Information Systems*, *2022*, 6851616.

Singh, A. K., Kumar, A., & Maurya, A. K. (2014). An empirical analysis and comparison of Apriori and FP-growth algorithm for frequent pattern mining. *Proceedings of the 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, 1599-1602.

Tsai, C. W., Lai, C. F., Chiang, M. C., & Yang, L. T. (2013). Data mining for internet of things:

A survey. *IEEE Communications Surveys & Tutorials*, *16*(1), 77-97.

Xuanyuan, S., Xuanyuan, S., & Yue, Y. (2022). Application of C4.5 algorithm in insurance and financial services using data mining methods. *Mobile Information Systems*, *2022*, 670784.

Yan, J., Wang, X., Wang, B., & Zhang, Y. (2019). Research on application of data mining technology in risk assessment process of audit. *Proceedings of the 2019 International Conference on Economic Management and Model Engineering (ICEMME)*, 487-491.

Yun, U., & Ryu, K. H. (2011). Approximate weighted frequent pattern mining with/without noisy environments. *Knowledge-Based Systems*, *24*(1), 73-82.