淺論歷史資料與資料品質之關係

朱昌綸／金融聯合徵信中心資訊部資料品質小組組長

仰首回顧日常生活中最常有的動作，就是察看手錶時間或是檢視行事曆，舉凡：學校上課、電視節目、上班簽到、公司會議、火車時刻、結婚紀念日…等等，這顯示著現實世界就是一連串時間（timestamps）與時間區段（time intervals）堆砌之組合；呱呱墜地的嬰兒長成學步的小孩，頑皮的青少年長成忙碌的上班族，進入職場之後，我們的知識、技能、職稱、薪資逐年成熟增進，企業產品銷售的流行或退潮，股票價值的起落波動，企業營收的得失增減，市場景氣的盛衰榮枯，春夏秋冬的交互替換…等等，這代表著現實世界中，任何事件都是隨著時間潺潺逝去而更迭。

歷史資料（historical data）在資料的生命周期中扮演著相當關鍵的角色，歷史資料不僅提供資訊應用人員更豐富的訊息做出分析決策，同時，資訊處理人員也可利用歷史資料做爲前後期資料檢核時勾稽比對的工具。具體而言，歷史資料可謂由「資料庫建置」及「資料庫內容」協同架構而成的資訊系統；資料庫的建置既是長期性、連續性及延續性的工作，必須經過欄位設計、系統規劃、傳遞、彙集、檢核、建置、儲存、揭露等一連串作業流程；而資料庫的內容本身亦是由一連串與時期相關的事件逐漸累積出來的資訊匯集；申言之，歷史資料不論在資料彙集處理過程之中，或是在資訊分析應用之下，倘若資料歷程的規劃不妥適，勢必增高資料品質不良的機率，況且，此類資料歷程規劃之不妥適，也往往囿於短時期內不易被察覺，進而導致日後必須付出更高的代價來修補。本文將以「資料的時間相關屬性」、「前後期資料值的變化限制」及「前後事件變化的因果關係」三種層面，來介紹歷史資料對於資料品質之關聯。

一、資料的時間相關屬性（time-dependent attribute）

資料的時間相關屬性在資料品質評鑑過程中，屢屢受到較多的關注與省視。這些與時間相關的屬性包括：資料的新鮮度、資料的回溯度、時區的一致性、時區的連續性、時區的頻度等五項屬性，分別說明如下：

1、資料新鮮度（currency）

「資料新鮮度」係指該項資料最近一次的觀測時間與現在的時間差距。例如：J機構存

有某君97/12時之薪資資料，A機構僅存有該君97/09時之薪資資料，此時，J機構該君薪資之資料新鮮度較A機構為佳。

2、資料回溯度（retention）

「資料回溯度」係指該項資料歷次觀測軌跡被保存的時間深度。例如：J機構保存某公司96年第四季至97年第四季之會計師財務簽證資料，B機構僅保存該公司97年第二季至第四季之會計師財務簽證資料，此時，J機構該公司會計師財務簽證資料之資料回溯度較B機構為佳。

3、時區一致性（granularity）

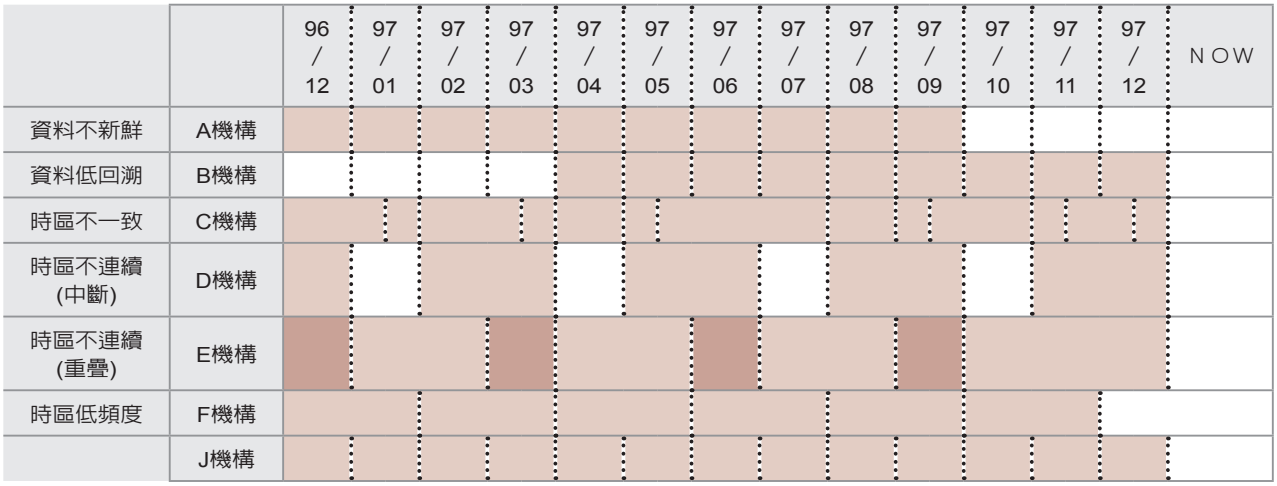
「時區一致性」係指該項資料之各觀測時間區段是否有固定長短。例如：J機構保存某君96年至97年間，每月份月底之授信餘額資料，C機構保存該君96年至97年間，長短不固定時區之授信餘額資料，此時，雖然J、C機構之資料新鮮度及回溯度相同，但J機構該君授信餘額資料之時區一致性較C機構為佳。

4、時區連續性（continuity）

「時區連續性」係指該項資料之各時觀測時間區段前後期是否保持連續，所謂時區「不連續」，又分為時區「中斷」及時區「重疊」二種類型。例如：J機構保存某商店96年至97年每月之網路交易總金額資料，D機構保存該商店96年至97年每季之前二個月網路交易總金額資料，E機構保存該商店96年至97年每季之前四個月網路交易總金額資料，此時，D機構之資料時區呈現「中斷」，而E機構之資料時區呈現「重疊」，此時，J機構該商店網路交易總金額資料之時區連續性較D、E機構為佳。

5、時區頻度（frequency）

「時區頻度」係指該項資料之觀測時間區段之周期是否短暫密集。例如：J機構保存某君97年每月份月底之授信餘額資料，F機構保存該君97年每單數月份月底之授信餘額資料，此時，雖然J、F機構之資料回溯度相同（資料新鮮度最多有一個月的時間落差），資料時區



長短固定且前後期保持連續，但J機構該君授信餘額資料之時區頻度較F機構為佳。

二、前後期資料值的變化限制（value constrain）

資料於前後期時區中，資料值的變化常常是資料品質檢核過程中想要一探究竟的對象，有關資料值變化的限制包括：資料變化的方向性、資料變化的幅度、資料變化的穩定度、資料變化的因果關係等三種形態，分別說明如下：

1、資料變化的方向性（direction）

「資料變化方向性」係指資料於前後期時區中，當資料值隨著時間的改變，有否遞增、遞減或維持不變趨勢之限制。例如：「年齡」資料值應逐年增加，「出生日期」資料值應固定不變，若某君的生日前後期不一致，則該資料可視為異常；再者，若某君「學歷」資料值隨著年度增加而降低，則該資料亦可視為異常。

2、資料變化的幅度（magnitude）

「資料變化幅度」係指資料於前後期時區中，當資料值隨著時間的變化，有否上限（max.）及下限（min.）程度之限制。例如：若某君「身高」資料值相連二年度之差異超過15公分，則該資料可視為異常；再者，若某上市公司之股票相連二交易日之收盤價格漲、跌幅度超過7%，則該資料亦可視為異常。

3、資料變化的穩定度（volatility）

「資料變化穩定度」係指資料於前後期時區中，當資料值隨著時間的變化，有否必須循序漸進，禁止遽然振盪之限制，所謂振盪係指短時間內資料值瞬增後又瞬降，或者瞬降後又瞬增。例如：職員之「薪資」資料值，除了偶爾因轉換跑道或升遷調職而出現較高幅異動現象，通常緩步循序漸增，若某君在同一家公司之薪資短時期內大起大落，則該資料可視為異常；再者，若某君短時期內在同一家銀行之信用卡資料，先為一般停卡戶，再為強制停卡戶，又為一般停卡戶，則該資料亦可視為異常。

三、事件前後變化的因果關係（cause & effect constrain）

事件（event）相對於資料值（value），其定義較為盤根錯節，首先，事件牽涉非單一對象（例如：債權人、債務人），其次，事件不僅徒具一個時間點，而是跨越一段時期，最後，事件成立的條件是由多項特定資料值的描述組合而成；有鑑於此，事件因果關係之檢驗於資料品質檢核過程中尤屬一項嚴竣的挑戰，有關事件因果關係之檢驗包括：由事件原因檢驗結果、由事件結果檢驗原因等二個面向，分別說明如下：

1、由事件原因檢驗結果（event dependence validation）

「由事件原因檢驗結果」係指當事件隨著時間的變化，是否存在有其因就必有其果之限制。例如：某君授信餘額資料上期為結案（原因），若本期該戶同帳號之授信餘額資料仍有金額（結果），則本期資料可能為異常。

2、由事件結果檢驗原因（event condition validation）

「由事件結果檢驗原因」係指當事件隨著時間的變化，是否存在有其果就必有其因之限制。例如：某君授信餘額資料前期素未有逾期或催收等不良紀錄（原因），若本期該戶授信餘額資料為呆帳（結果），則前期資料可能為異常。

誠如前言，事件成立的條件（event-specific attribute）係屬若干具體資料值的集合（例如：授信違約戶、信用卡道德風險、首購自用住宅貸款…等等），惟此事件成立條件亦唯有「後端」資訊應用人員才熟稔究竟為何，此現象將導致事件因果關係之檢驗，往往於「前端」資料建置處理的階段較無從落實；為了降低前端執行事件因果關係檢驗的複雜性及困難度，可考慮於資料欄位規劃時，另行設計「事件」欄位，將事件的描述以「代碼」表示，例如：債務協商註記（Y/N）、結案註記（Y/N）、催收（科目=A）、呆帳（科目=B）…等等，旋即再將此「代碼」視同為前述之「資料值」，裨益執行前後期時區資料值變化之檢核。不可諱言，此事件代碼欄位之設計雖然看似重覆及累贅，但實務上，此設計卻也可能提供「事件成立條件」與其相關「資料值」組合另類交互勾稽（cross check）的檢核機制。

結語

盱衡國外信用資料庫機構，其資料的彙集除了自行輸入建置外，大都藉由同業交換或購買而來，由於原始資料來源不同，有關資料的新鮮度、資料的回溯度、時區的一致性、時區的連續性、時區的頻度往往彼此互有差異，導致原始資料必須再進行整合（integration）或轉換（migration）程序，然而，倘若於此資料處理程序中，相關假設條件拿捏失準，則易滋生資料內容失真，進而顛簸爾後歷史資料之品質。

聯徵中心在金融監理主管機關的指示與要求，並承蒙各會員金融機構的充分支持，將各項信用資料依據報送作業要點之規範，「定期全體報送」至聯徵中心進行彙集，再由聯徵中心提供查詢介面，揭露彙整後之信用資訊，供各會員金融機構參考利用；如前述，聯徵中心受惠於資料必須「定期」並且「全體」報送規範之賜，其「資料的時間相關屬性」（包括：資料新鮮度、資料回溯度、時區一致性、時區連續性、時區頻度），相較於國外信用資料庫機構，先天上就已具備較健全的制度與基礎來保存相關信用資料之歷史紀錄，也就是說，當資訊使用者進行聯徵中心信用資料庫之歷史記錄分析應用時，得以享有較高的資料品質。然而，聯徵中心確也不能因此自滿而誇口，自當秉持會員機構資訊互惠共享的信念，積極致力將各種信用資料「前後期資料值的變化限制」（包括：資料值變化方向性、資料值變化幅度、資料值變化穩定度）及「事件前後變化的因果關係」（包括：由事件原因檢驗結果、由事件結果檢驗原因）等邏輯，設計於資料品質檢核的需求範圍之內，評鑑資料值及事件隨著時間的變化是否合理，確實提升信用資訊的內容品質，落實增進信用資訊的附加價值。