空間資料庫之關聯樣式探勘

Mining Association Pattern from Spatial Database

鄒明城*　　孫志鴻**

Ming-Cheng Tsou　　Chin-Hong Sun

摘　要

隨著資訊科技的迅速發展，地理資訊系統在環境議題的研究上扮演越來越重要的角色。然而，傳統的資料庫管理系統與地理資訊系統，已不能充分滿足在空間決策支援上的需要，空間資料探勘的技術已越來越受重視。再者，空間資料通常是彼此交互關聯的，一種空間現象的分佈不只跟自己本身的特徵有關，更與其它鄰近空間現象有著某種程度的關聯或相互依存。本研究以集集大地震引致的山崩爲案例，利用地理資訊系統強大的空間資料處理能力，將相關資料建立成提供資料探勘使用的資料倉儲。再利用關聯法則分析與等級相關分析等資料探勘技術，探索隱含於地震山崩與相關背景環境因子間的關聯樣式，獲致重要的結果，相信這樣的方法可以應用於日後其它具豐富地理資料庫的環境議題上。

關鍵字：資料探勘、地理資訊系統、地震引致山崩

Abstract

Geographic Information Systems (GIS) is growing popularity and gaining importance in environmental applications, as the technology advances in spatial data handling and analysis. However, GIS technology still fails to meet the demand for spatial decision support fully, because it is incapable of eliciting meaningful information from volumetric multidimensional geospatial data. Recently, geospatial data mining emerges as a promising approach to meet the challenge. In particular, mining patterns of spatial association can reveal variables that may be related in some way and, therefore, can be set forth to formulate hypotheses. In this study, we analyze the association pattern between landslides induced by

*　實踐大學高雄校區資訊管理系助理教授
Assistant Professor, Department of Information Management, Shih Chien University Kaohsiung Campus.

**　國立台灣大學地理環境資源學系教授
Professor, Department of Geography, National Taiwan University.

Chi-Chi earthquake and environmental variables to demonstrate the use of two spatial data mining techniques, association rules mining and Spearman rank correlation, to gain insights into environment settings vulnerable to landslides after Chi-Chi earthquake in central Taiwan. While the case study is specific, our approach is applicable to associate environmental variables that may address other environmental issues of interest.

Keywords: Data Mining, Geographic Information System, Earthquake-Induced Landslide.

前　言

隨著資訊科技的進步，資訊系統在巨量資料處理和高速運算功能上有非常顯著的突破與進步。資料探勘 (Data Mining) 藉由此類優異的資料處理與運算能力，從所掌握的大量資料中進行多維度搜尋，並更進一步地擷取出隱含於其中之知識樣式 (Knowledge Pattern)，正成爲一門新興的研究領域 (Koperski *et al*. 1996)，爲資料庫領域上新的應用，企圖將隱藏於資料庫中有用、有益於使用者所需的資訊挖掘出來，以提供企業及使用者決策之需，幫助企業獲取商機 (Chen *et al*., 1996; Fayyad *et al*., 1996; Berry and Linoff, 1997; Keissner, 1998; Han, 1999; Miller and Han, 2001)。它不像演繹法是用來證明或反證假說 (Mennis and Liu, 2003)，而是以歸納的方式從資料中找出內隱的樣式並產生假說。資料探勘所能提供的智慧型資料分析，將可讓吾人透過對資料內涵的更通透了解，進而充分且有效的解決所面對的各種問題。從九〇年代起，資料探勘已廣泛且成功地被用在市場調查、行銷分析研究、經營決策分析、製造工程控制、生物資訊研究等領域。同樣地，各種空間資料也以驚人的速度快速累積，如何有效的萃鍊這些資料提供更具見解的知識，也是當今地理資訊科學研究者所面臨的一大課題。空間資料探勘 (spatial data mining) 是資料探勘的一個新的研究分支，其本質是從空間資料庫中挖掘、發現時空系統中內在的、有價值的規律和知識的過程，包括空間資料的樣式與特徵、空間與非空間資料之間的概要、關聯關係等 (Lu *et al*., 1993)。

現有的地理資訊系統一般已經具有空間資料管理、製圖、查詢、統計與空間分析功能，但缺乏知識的表達、獲取和應用的機制。因此，現有地理資訊系統的智慧化能力仍然較低，空間資料探勘技術正好可以彌補此一不足，因爲空間資料探勘所獲得的知識與現有地理資訊系統分析工具所獲得的資訊相比，更加的概括化與精練，可以找出現有地理資訊系統分析工具無法獲取的隱含樣式和規律，可以具備與人類思維相近的推理模式。雖然目前地理資訊系統已開始與各類專家系統結合起來 (Zhu, 1999；鍾新南，2003)，且與專家系統整合後可以在較大程度上解決知識的表達與應用的問題，但總的來說，在知識獲取方面依然存在一大瓶頸。如果在地理資訊系統中引入空間資料探勘技術，就有可能自動、半自動地從大量的空間資料中發現一些特定的知識或規律來輔助決策過程，提高複雜系統決策的科學性、合理性和決策效率。另外，統計方法可以說是空間資料分析中最常用的方法，雖然它也是資料探勘的一個較傳統的分支，但它通常假設某種統計分佈以及資料間彼此獨立，由於空間資料通常是彼此交互關聯，一種空間現象的分佈不只跟自己本身的特徵有關，更與其它鄰近空間現象有著某種程度的關聯或相互依存，這就限制了統計方法在空間資料探勘上的應用。

Han and Kamber (2000)、Berry and Linoff (1997) 將資料探勘技術歸類爲預測型 (predictive) 與敘述型 (descriptive) 二種。預測型包含像類神經網路、決策樹、案例推理、貝氏分類器等模式，可用於分類、預測，而敘述型資料探勘則著重在現象的描述、了解、解釋與知識發現 (Berry, 1997)。過去已有多位學者對於地震引致之山崩進行預測型模式上的探討 (Keefer, 1984; Refice, 2002)，而鄒明城、孫志鴻 (2004) 也曾針對相同地區與相同之背景條件，進行決策樹與類神經網路預測型模式

之建立，期能預測其它地區相同條件下發生山崩之可能下。本研究不同於以往，主要針對敘述型資料探勘進行探討，特別著重於關聯樣式 (association pattern) 的探索，分別以關聯法則與 Spearman 等級相關分析作爲探勘的工具。關聯法則屬於敘述型的資料探勘技術，主要用來探勘輸入因子與地震山崩間的關係組合規則，而 Spearman 等級相關分析則屬於敘述統計，用來探討地震山崩與各單一影響因子間的關聯性，二者皆爲敘述型，透過這樣的搭配組合，既可探討個別因子也可探討組合因子對於山崩的影響，彌補地理資訊系統在相關智慧型空間分析的不足，找出各背景空間因子與地震崩塌地之間的關聯性，探討造成地震崩塌地的可能因素，並建立成易懂的法則，作爲自動的法則產生器，提供建立規則庫，可供後續模式建立與防災決策上的參考。

研究架構與方法

(一) 研究架構

本研究以 Roiger and Geatz (2002) 及 Han (1999) 等人對資料庫知識探索 (Knowledge Discovery in Database, KDD) 所定義的七個步驟做爲研究進行的流程，過程如下：

(1) 訂定目標：了解想要進行資料探勘標的的專業領域，必須要能清楚的敘述出要實踐的目標，並提出可能的假設或想要達成的成果。

(2) 建立目標資料集：選擇所要分析的初始資料集。如相關的空間背景及屬性資料。

(3) 資料前處理：用有效或可取得的辦法來處理資料雜訊，並且須決定要如何處理資料的遺漏。

(4) 資料轉換：刪除或新增目標資料群的屬性和資料。這個步驟必須決定一些標準化、轉換以及修飾資料的方法。如向量資料轉網格資料以及空間資料倉儲的建立。

(5) 資料探勘：使用一個或多個資料探索演算法，將資料處理成最佳的表現模式。

(6) 解釋與評估：審查模式所做出的結果，找出是否有用且有趣的資訊，若沒有，則再以新的資料屬性與範例重複之前的步驟。

(7) 採取行動：假如探索到的知識被認爲是有用的，則這些知識會被整合並直接應用到適當問題的解決方案上。

本研究的架構與方法是從資料探勘的角度進行 (如圖 1)。研究的主要目標，是探索內隱於地震山崩資料庫中的造成山崩因子間關聯樣式 (如流程步驟 1)，因此在進行研究之前，先儘可能蒐集相關學者在空間資料探勘與地震引致山崩之研究文獻，了解相關技術應用的限制與可行性及引致山崩機制的環境因子探討，以便建立接下來研究上相關的基礎背景知識 (如圖 1 地震山崩 domain knowledge)。有了這些背景知識後，便可進行相關影響因子資料的蒐集與建立 (如流程步驟 2)，資料來源多樣化，包括地理圖層資料以及屬性資料，先進行必要的前處理 (如流程步驟 3)，再將這些資料分別建置進入地理資訊系統與關聯式資料庫中，做爲資料探勘的基礎資料。相關的基礎資料必須做進一步的處理，以建立適合資料探勘技術使用的乾淨資料。因爲本研究以網格式資料爲探勘的基礎，因此，必須先將各個向量型地理圖層資料轉化爲網格資料、清理空的資料，並且將不必要的資料屬性剔除，同時也利用現有的資料依空間關係再衍生出新的資料欄位。這些資料經過清理乾淨後必須加以整合再輸入資料倉儲中 (相當於流程步驟 4 與圖 1 之資料清理轉換)，由於資料龐大，因此必須有一適當的管理系統來管理這個資料倉儲，本研究以微軟公司的 SQL Server 資料庫軟體作爲資料倉儲建置與管理的平臺。經過適當的採樣技術由資料倉儲中進行資料採樣後，做爲各資料探勘模式使用的資料來源。接著就以關聯法則及 Spearman 等級相關分析從資料中找出隱含的關聯樣式並加以評估其有效性 (如流程步驟 5、6)，最後這些法則及知識可以建立成爲知識庫，用以分析預測與建置爲防災決策支援系統的一部份 (如流程步驟 7)。

(二) 關聯法則

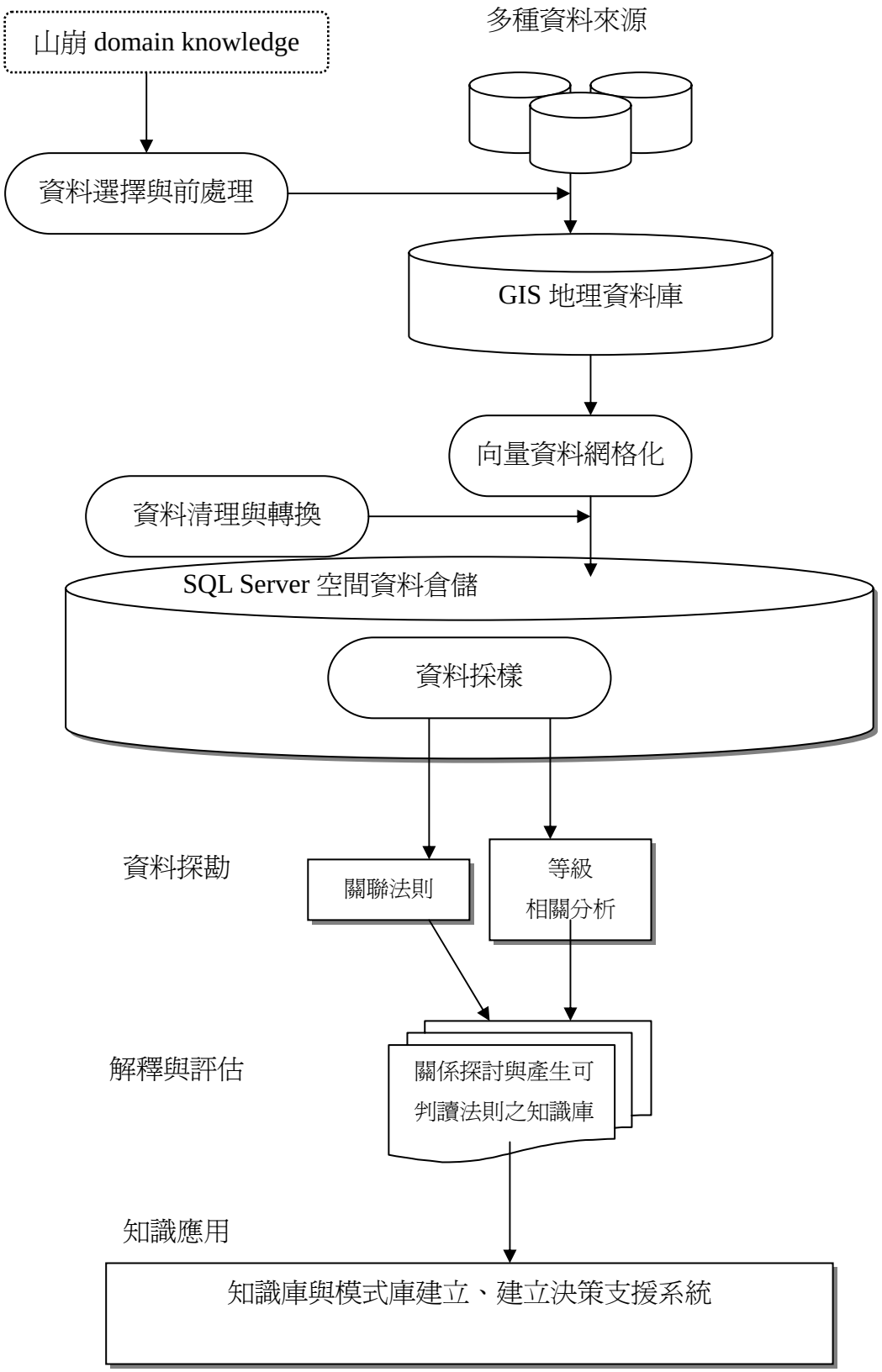

山崩 domain knowledge
多種資料來源
資料選擇與前處理
GIS 地理資料庫
向量資料網格化
資料清理與轉換
SQL Server 空間資料倉儲
資料採樣
資料探勘
關聯法則
等級
相關分析
解釋與評估
關係探討與產生可
判讀法則之知識庫
知識應用
知識庫與模式庫建立、建立決策支援系統

圖 1　研究架構

關聯法則可以說是資料探勘中最有名的一項應用，又稱爲購物籃分析 (Market Basket Analysis)，最早是由 Agrawal *et al.* (1993) 所提出的，並隨後於 1994 提出 Apriori 演算法 (Agrawal and Srikant,1994)。是一套探索變數間關聯的資料探勘技術，常應用於商業行銷與客戶關係管理 (CRM – Customer Relationship Management) 上。其原始構想是用來分析顧客的消費行爲，也就是說當顧客走進賣場推著推車購物時，購物車上總是放著多樣不同的商品，交易結帳後，資料庫內會紀錄著每一顧客交易的明細，關聯法則企圖從消費者的交易紀錄中，找出那些產品總是被顧客同時購買，如此，便可以進行貨架的安排或是行銷策略的制定。最有名的例子，就是美國 Walmart 賣場發現了尿布與啤酒之間的關聯性。另外日常生活中常見的例子，還有亞馬遜電子書局之線上購物，比如當我們挑選了 A 書後，網站會建議顧客，通常購買 A 書的顧客也會購買 B、C 或 D 書，藉以吸引顧客以獲得額外的商機。另外，銀行也可以藉著由探勘消費交易資料庫中的關聯法則來分析消費者的行爲，做爲未來一對一客製化行銷上的參考。

1. 關聯法則之定義：

我們先爲它做一簡單的描述，並以購物交易作例子來說明 (Agrawal *et al.*, 1993; Agrawal and Srikant, 1994; Chen *et al.*, 1996) ：假設 $I = \{i_1, i_2, i_3, \ldots i_m\}$表示商場內所有商品，即項目 (items) 的集合，稱作項目集 (itemsets)，而其中每一筆交易 T 爲部份項目的集合，即 $T \subseteq I$。假設 X 是一個項目集合，若 $X \subseteq T$，我們可以說 T 包含 X。而關聯法則「**購買 X 商品也會同時購買 Y 商品**」的形式如下：

***X → Y* [Support, Confidence]**

其中，

$X \subset I, Y \subset I$ 及 $X \cap Y = \varnothing$(空集合)

Support = Probability $(X \cup Y)$

Confidence = Probability $(X \cup Y)$ */ Probability* (X)

X 和 Y 都是購物項目 (items) 的集合，不限單一商品或項目，還有兩個最重要的指標，信賴度 (confidence) 與支持度 (support)。

信賴度 (confidence)：代表條件機率，即此一規則的準確度有多少，爲購買 X 產品項目的顧客中有也會同時購買 Y 產品項目的比率，即 (the number of transactions containing X and Y) / (the number of transactions containing X)。信賴度越高，則此一法則越有參考價值。

支持度 (support) ：是交易資料庫中，同時包含有 X 與 Y 商品項目，佔全部交易數的百分比，即 (the number of transactions containing X and Y) / (the number of transactions) 的值。信賴度高固然表示規則具有很高的準確度，但是前提是該種交易型態出現的次數夠多，該法則才具有代表性。

進行關聯法則資料探勘前，通常必須先訂出最小的信賴度門檻與最小的支持度，來剔除不合格的規則，才能從大量的候選規則中篩選出合格或有用的法則。它的計算程序很簡單，可以分爲兩個階段： (1) 找出大項目集合 (Large Itemsets) ：首先計算各單一項目在資料庫中出現的次數，判斷其是否大於等於使用者所定義的最小支持度 (minimum support)，以決定出大項目集 Li (Large I-itemsets)。其做法是在每個回合中進行合併及刪除的階段，產生候選項目集合 (candidate itemsets)，接著對每個候選項目集合計算其支持度，將滿足最小支持度的候選項目成爲大項目集合，如此反覆上述步驟，直到無法產生新的候選項目集合爲止。(2) 產生關聯法則：將每個大項目集 (即 X∪Y 的支持度達到最小支持度的集合)，以 X∪Y 的支持度除以 X 的支持度，以計算出 X → Y 的信賴度。若信賴度達到使用者定義之最小信賴度 (minimum confidence)，則 X → Y 這樣的關聯法則成立。本研究以 Apriori 演算法做爲關聯法則探勘上的方法。

2. 多層式關聯法則

資料在概念上有抽象層次高低之分，不同的抽象層次有時可以帶來不同的新發現，Han *et al.* (1999) 曾建議根據資料抽象層度的不同，進行多層次關聯法則探勘。以底下兩個空間敘述爲例：(1) 商店位於忠孝東路上且時間爲七月，則七喜汽水銷量佳，(2) 商店位於信義路

且時間爲八月，則黑松汽水銷量佳。乍看之下似乎沒有什麼關聯，規則性不強，這是因爲資料仍處於低抽象層次上，但當我們把資料層次提高之後，隱含的意義就浮現出來了，忠孝東路、信義路可以視爲東西向幹道，七月、八月可歸納爲夏季，七喜、黑松汽水可以歸納爲碳酸飲料。因此一條新的規律產生如下：當商店位於東西向幹道上且時序處於夏季，則碳酸飲料銷售佳。這樣的隱含規律，在資料庫中處處可見。人腦具有歸納的能力，對於第一、二條描述，可以很快歸納出其中的隱含意義，這是因爲我們對於像汽水爲碳酸飲料這樣的判斷具有先驗的知識 (pre-knowledge) 當基礎，故可以歸納出更高層次的知識，而如何讓程式知道這樣一個知識層次架構，則有賴於專家事先定義好，再交由程式去搜尋。有鑑於此，本研究引入多層式關聯法則的概念，希望藉由對多層知識構念架構找到更多隱含的空間知識。

(三) Spearman 等級相關 (Spearman Rank Correlation)

本研究中除地質分佈因子本質上是類別型資料 (categorical) 外，其它各因子爲配合關聯法則的使用，均已由數值型資料轉換爲類別型資料，且崩塌地是以網格形式來表示，其網格值爲布林型態 (有或無山崩)，崩塌地數目的統計必須做計數 (count)，較不適合使用一般數值型的相關分析，因此藉由 Spearman 等級相關來分析各單一因子變數與山崩之間的關聯性。等級相關爲迴歸與相關的分析方法之一，主要用於當兩變數 X 與 Y 的母體分配未知，爲了解 X 與 Y 之間的相關，將兩樣本資料 X 與 Y 分別依大小排序 (由至大至小或由小至大皆可，其結果相同)，並給予等級 r_x 與 r_y，得到的統計量

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad (1)$$

r_s: 爲關聯係數，n: 爲分區數，d_i: 爲各分區內山崩格數與該分區內屬於某屬性分類之格數排序的差。當 X 與 Y 的等級順序完全相等時 $r_s = 1$，爲正相關，當 X 與 Y 的等級順序完全相反時 $r_s = -1$，爲負相關，當 X 與 Y 不相關時 $r_s = 0$，而得到 $-1 \leq r_s \leq 1$。

案例研究—集集大地震引致山崩地理資料庫

(一) 研究區域資料概述

由於集集大地震所誘發的山崩大多集中在台灣中部區域，故以台灣中部山區爲研究區域 (如圖 3)。參考之前集集大地震山崩研究文獻 (童啓哲，2001；許煜煌 2002；鄒明城、孫志鴻 2004)，選擇了十七個因子圖層進行空間資料庫的建立 (如表 1)，藉以探討這 17 個因子與地震山崩間是否存在某種關聯樣式。而崩塌地的部份，則是以工業技術研究院能源與資源研究所 (2000) 受農委會水土保持局之委託，辦理集集大地震崩塌地調查與治理規劃，在經過 1999 年 4 月 9 日至 7 月 24 日間之災前衛星影像 (SPOT) 與 9 月 27 日災後之衛星影像及航空照片判釋後，所標繪之 21969 筆變異點向量圖層網格化後的資料。在完成以上各圖層之後，由於大部份爲向量式資料，須再以 ArcView 分別將向量式主題圖層予以網格化，爲了配合 40m*40m 的 DTM 大小，因此將每一個圖層的解析度均設爲 40m*40m，並且具有相同的邊界範圍，故每個圖層之相同相對位置的每一網格均具有相同的地理範圍。之後，再以程式將各個圖層相同相對位置的每一網格之資料值結合成一筆一筆的紀錄，每一筆紀錄內之每一個欄位均對應到每一主題圖層相同相對位置的網格值，共計有約 122 萬餘個網格，其中屬於崩塌的網格大約只有約 6 萬個。然後再將這些紀錄建立成爲空間資料倉儲，以提供資料探勘時的資料來源。

(二) 資料化簡與採樣

本研究所面對的資料數量高達 122 萬餘筆，但其中屬於山崩的部份不到 6 萬筆，爲了處理這樣大量的資料，即便是目前的電腦軟硬體技術，處理上亦力有未逮。因此需要採取採樣的技術來簡化資料，但是，山崩資料相對於母體而言過於稀少，若採一般隨機抽樣，可能造

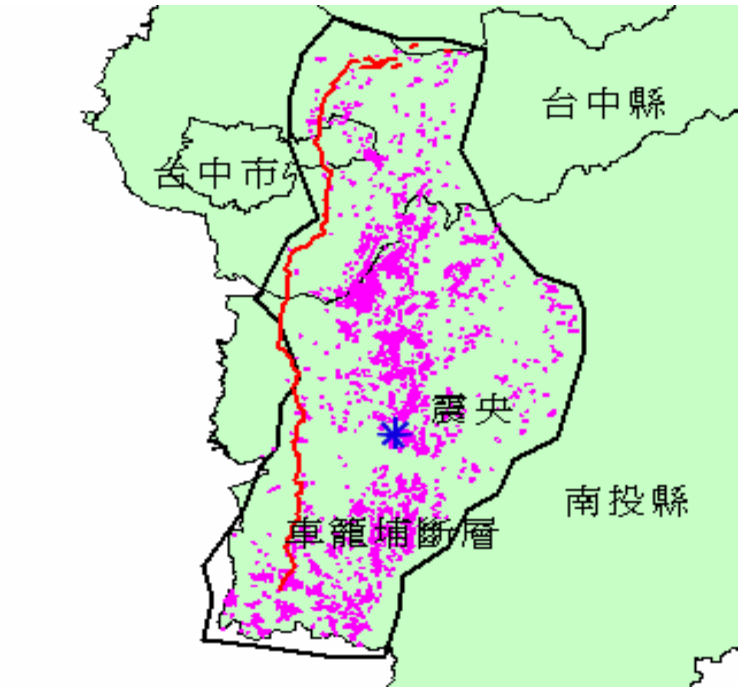


圖 3　研究區域及地震山崩圖

表 1　地震山崩之輸入圖層

原生因子圖層與資料來源	衍生因子圖層
高程 (中央大學太遙中心之 78 年 40M DTM)	九格點之平均坡度
坡向 (由高程資料推演)	九格點之最大最小坡度差
坡度 (由高程資料推演)	九格點之平均坡向
距離車籠埔斷層距離	九格點之最大最小坡向差
距離斷層破碎帶距離	地震強度 (Arias Intensity)
距離道路距離 (由內政部所公佈之省縣鄉道路分佈圖推算)	
距水系距離 (由 88 年水利處所公佈之水系分佈圖推算)	
距震央距離 (氣象局相關紀錄推算)	
地質分佈狀況 (地調所 1/25,000 地質圖)	
垂直向地表加速度 (氣象局相關紀錄推算)	
東西向地表加速度 (氣象局相關紀錄推算)	
南北向地表加速度 (氣象局相關紀錄推算)	

成對於稀少事件的不易掌握。Berry (2000) 建議可以採用超採樣 (Oversampling) 的技術，增加樣本中稀有事件的比率，他認爲維持在 10%~40%的比率，通常可以獲得不錯的結果。故本研究先將資料分成山崩與未山崩二部份，再分別從屬於山崩的六萬筆資料中隨機抽樣 2 萬筆、另從未山崩之 116 萬筆資料中抽樣 4 萬筆混合成 6 萬筆樣本，做爲探勘的資料依據。

(三) 關聯法則之應用

關聯法則最近已實際應用在商業領域，但在空間問題研究上非常少，Koperski and Han (1995) 是最早研究此方面演算法的學者，其它大多爲探討商業與區位或人口統計資料間的關聯法則 (Mennis and Liu, 2003; Kangkachit and Waiyamai, 2002; Tang and McDonald, 2001)，應用的範圍仍不廣。從關聯法則原始設計目的來看，在於探索那些商品有可能同時會被一起購買，或是購買了那些商品後，就有可能購買某種商品。同理，也可以將這樣的觀念應用於空間現象的探討上，將商品項目置換成空間因子，例如，某一空間現象的發生，常伴有那些其他的可能空間現象，或是，當空間環境具有某些環境條件時，會造成某種空間現象的產生，我們可以

把環境視爲一個超大型的賣場，把空間現象看作是對於環境的一個消費，而消費的項目就是各種環境條件，每一地點環境條件皆不同，就相當於每個人的購物習性不同一般，而這些空間現象所構成的空間資料庫，就相當於交易明細紀錄的資料庫，然後再由巨量的空間資料庫中歸納出造成某種現象所隱含的法則。以地震山崩資料庫爲例，地震山崩可能由多種的環境因子所造成，例如地質條件、地震強度、坡度、坡向等，通常是彼此互相關聯，很少是因爲單一因子所造成的，也就是說，某處當它的環境條件符合某些狀況時，則有可能造成山崩的現象，這正是關聯法則可以找出來的規律或法則。

關聯法則應用於空間問題的解決，最重要的關鍵在於資料的轉換與編碼，因爲關聯法則適用的資料形式爲一串項目 (購物清單)，每一個項目包含邏輯值 (購買或未購買)，每一筆紀錄可能都不等長，不完全適用於關聯式資料表，而地震崩塌資料庫則大多爲數值型資料，如何將連續性或類別性的資料轉換成適合關聯法則演算的交易項目 (item) 資料，是本研究所面臨的一大重要課題，誠如 Roiger and Geatz (2002) 所提，資料轉換爲資料探勘過程中最耗成本但也最重要的步驟之一，它關係著接下來資料倉儲的建立與資料探勘演算上所需的基本素材。本研究使用分等的觀念，將數值資料切割爲若干等份，使數值資料轉換爲類別型的資料以減少資料項目的數量 (相當於商品的種類)，分等的方法採 Jenks and Coulson (1963) 所提之自然分等法 (nature break)，經過這樣的轉換後，數值資料轉換爲具有 9 個等級的類別資料 (表 2)，如此，每一筆紀錄的每一欄位屬性值均可視爲某一購物的商品，例如，若原距斷層距離之屬性爲 187 公尺，而分等時以 100 公尺至 200 公尺爲第二級，則產生「DistToFault_2」這樣的類別值來取代原先的數值資料「187」。另外，由於探索的是多層式關聯法則，因此本研究將每一因子再上拉一個層次，並且針對新的一層又新增一個欄位，也就說將原先的 9 個等級縮爲 3 個等級。它的編碼方式 (如表 2) 以之前的例子來說，由於上提一個層，故原先的第二級在上一層將變爲第一級，此時它的編碼值則爲「DistToFault_L1_1」，這裡 1 表示第一級，而 L1 則表示第一層，經過適當的編碼與轉換後，就可以將關連法則的觀念應用於地震山崩的研究上。

(四) Spearman 等級相關分析之應用

Zhao (2000) 曾以等級相關分析搭配地理資訊系統進行研究，其主要的設計在於先找出各多邊形行政區域內的購買顧客數，再搭配各行政區內的人口統計資料如薪資、教育程度等社經因子，探討不同區域間顧客購買汽車數與人口統計社經資料間的關聯性，例如購買車數與收入高低的關聯性。本研究因爲以網格式資料爲探勘的基礎資料，爲了適用於等級相關分析，以及將資料輸進 SQL Server 資料庫中，以便作爲資料倉儲，我們將研究範圍內之資料透過資料庫管理系統予以均等分成 50 區，每一分區具有相同的網格數。做法是由網格圖層將

圖層資料輸出成由左至右每個格點一一對應並且由上而下之逐列 (row) 的網格值資料，接著再將這些資料轉換進入資料庫管理系統中，於轉換的過程中，由系統會爲每一筆紀錄加上一個唯一的識別碼 (unique identifier)，再以此識別碼爲依據將資料區分成 50 等份，因此所得到的每一分區是一個面積相同但非正方形的區域。藉由這樣的分區方式不僅方便作業，更可以獲得較均質的資料，每一分區可涵蓋從平原至山地的範圍，不會僅侷限於全是平原或山地。以地質爲例，不會有某一區只存在一、二種地質條件，缺乏其它的地質條件，造成分析上的不便。使用的資料格式與關聯法則探勘相同，將數值資料轉換爲類別資料，然後針對每一屬性計算每一分區內之山崩網格數，並且加以排序，另外再計算各屬性值在該區內所佔的網格數並加以排序，最後再以公式 (1) 計算崩塌與該屬性質值的關聯係數，來探討環境因子與山崩之間的關聯性。

結果與討論

(一) Spearman 等級相關分析

洪如江 (2000)、廖軒吾 (2000)、童啓哲 (2001)、許煜煌 (2002) 等人曾針對 921 地震崩塌地與周圍環境因子進行探討，許煜煌 (2002) 並歸納出坡度、坡向、地質狀況、與水系的距離、與道路的距離、斷層的影響、地震的影響爲最常被學者所選用的因子。以下針對這幾項因子，個別探討他們與地震崩塌地之間的關係，計算等級相關係數，並繪製成統計圖，對於相關係數超過 0.5 者特別以*號標示。

1. 與地質分佈之關係：

以瑞芳群、三峽群及頭嵙山層具有較高的相關 (圖 4)，其中尤以瑞芳群與地震山崩關係最高，沖積層與台地堆積具有負相關。這樣的結果與之前研究相仿。

2. 與坡度之關係：

山崩普遍分佈於 18°至 65°之間 (圖 5)，之前研究則認爲坡度越大山崩越顯著。

3. 與坡向之關係：

本研究發現在 120°至 160°之間具有最高的相關性 (圖 6)，與童 (2001) 之東南向、許 (2002) 之 90°至 135°相仿，由於與車籠埔斷層及地質構造線走向大致類似，故應與這樣的因素有關。

4. 距震央距離：

以距離震央 20 至 26 公里具有最高的相關性 (圖 7)，與童 (2001)、許 (2002) 與廖 (2000) 則認爲在 15 至 30 公里間顯著性明顯，結論大致相同。

5. 距車籠埔斷層距離：

以距離車籠埔 8 至 16 公里具有最高的相關性 (圖 8)，童 (2001)、許 (2002) 則認爲在 7 至 15 公里間顯著性明顯，結論大致相同。

6. 與高程之關係：

高程在 370 至 732 公尺具有最高的相關性 (圖 9)，童 (2001)、許 (2002) 則認爲高程越高顯著性越明顯。

7. 距水系距離：

以距離河川水系 500 公尺以內具有較高的相關性 (圖 10)，但並不很高，許 (2002) 則認爲在 200 至 400 公尺間顯著性明顯，廖 (2000) 則認爲大多發生在距水系 300 公尺以內，結論大致相同。

8. 距道路距離：

山崩大多發生在距道路 2000 公尺以外 (圖 11)，但關係不明顯，距道路越近不見得有較高的山崩關係，可能與山崩大多發生在較偏遠的山區有關。

9. 與地震強度之關係：

強度在 270gal 以上關係較爲密切 (圖 12)，許 (2002) 則認爲加速度越高顯著性越明顯，廖 (2000) 則認爲大多發生在 250gal 以上。

(二) 關聯法則

關聯法則產生的關鍵主要在於信賴度及支持度的設定，因此在設計上，可以根據對於最小門檻值的設定，得到許多強弱不一的法則，本研究分別針對單層及多層知識結構進行法則探勘，由於法則相當多，以下僅針對規則性最強的法則列表，首先是單層式的知識架構，將

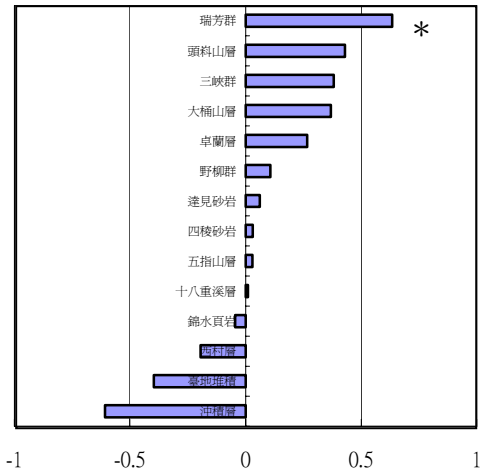


圖 4　地質分佈等級相關分析

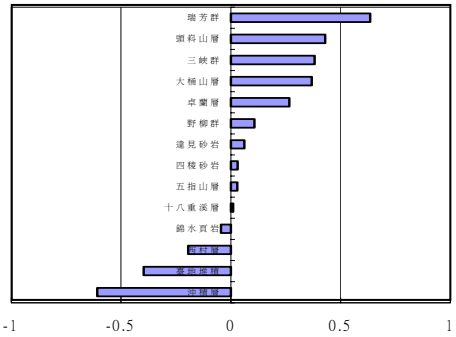


圖 5　坡度等級相關分析

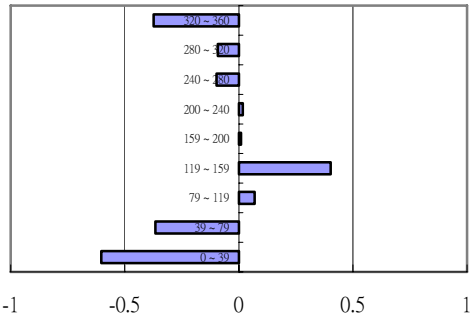


圖 6　坡向等級相關分析

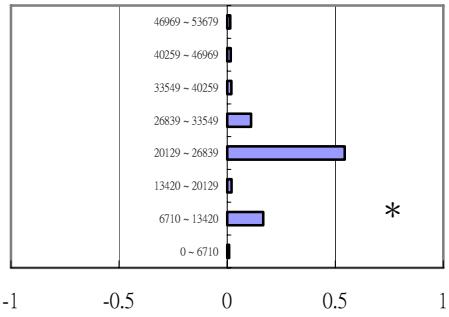


圖 7　震央距等級相關分析

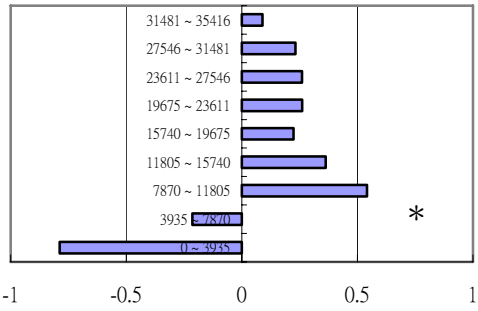


圖 8　斷層距等級相關分析

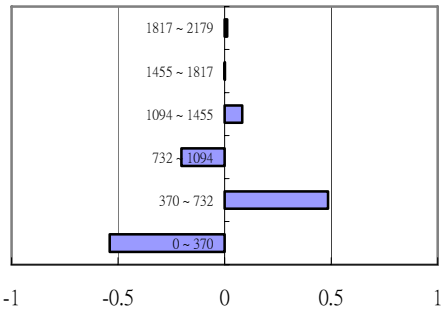


圖 9　高程等級相關分析

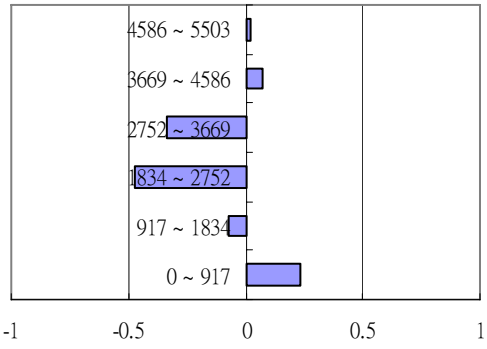


圖 10 水系距等級相關分析

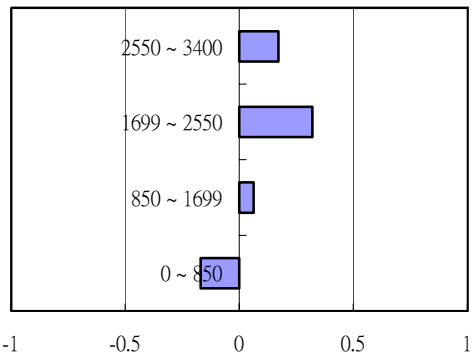


圖 11 道路距等級相關分析

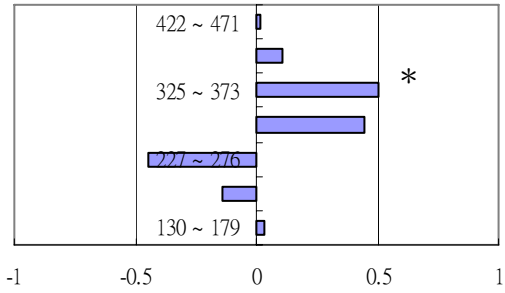


圖 12　垂直地表加速度等級相關分析

支持度設爲 7%信賴度設爲 90%共可獲得 11 條強規律性造成地震山崩的法則 (表 3)，其解讀方式以表 3 項次 6 爲例，代表：

若　(地質分佈 = 頭嵙山層) 且 (370 m <= 高程 <= 732 m) 且 (距水系距離 <= 900 m)　則可能造成山崩
支持度：7.5%, 信賴度：90.1%

對於多層式的知識架構，由於將抽象層次拉高，會獲得更多的法則，而且部份屬性質 (如九格坡度差、九格坡向差) 由於概括化後大部份的值均相同，造成所得的法則不具意義，因此去掉這些屬性並且將支持度設爲 8.8% 信賴度設爲 90%，共可獲得 13 條強規律性造成地震山崩的法則 (表 4)。

從以上二表之強規律性且有用的關聯法則，可以看出一些經常出現的關聯樣式，造成山崩的因素不外乎是某幾樣不斷出現樣式的排列組合構成的法則，而這些構成組合的樣式包含：

1. 地質分佈爲瑞芳群及頭嵙山層最多。
2. 坡度分佈在 30°至 60°之間。
3. 距震央距離在 7 至 14 公里間。
5. 距車籠埔斷層距離在 12 至 16 公里之間。
6. 高程位於 370 至 732 公尺之間。
7. 距道路距離 850 公尺以內。
8. 地震強度則以不論南北、東西、垂直加速度，以強度在 350gal 以上即具有密切關係。

表 2　單層與多層知識架構編碼值對照表

編碼	環境因子	編碼		環境因子	
Level 1, 2	Geology（地質分佈）	Level 1	Level 2	Dist2Faults（距斷層破碎帶距離）單位：公尺	Dist2Rd（道路距）單位：公尺
2	十八重溪層	1	1	1765	850
3	達見砂岩		2	3531	1699
4	西村層；佳陽層		3	5297	2550
5	廬山層；蘇樂層	2	4	7063	3400
6	瑞芳群及其相當地層		5	8829	4250
7	三峽群及其相當地層		6	10595	5100
8	野柳群及其相當地層	3	7	12361	5950
9	四稜砂岩；眉溪砂岩；白冷層		8	14127	6800
10	大桶山層；乾溝層；水長流層		9	15892	7650
11	五指山層；蚊子坑層；粗坑層				
12	錦水頁岩及其相當地層				
13	卓蘭層及其相當地層				
15	頭嵙山層及卑南山礫岩及其相當地層				
16	臺地堆積				
17	沖積層				

編碼		環境因子				
Level 1	Level 2	Slope(坡度) 單位：度	Aspect(坡向) 單位：度	Elevation(高程) 單位：公尺	Dist2Cen(震央距) 單位：公尺	Dist2Fault(斷層距) 單位：公尺
1	1	9.2	39.1	370	6709.8	3935
	2	18.4	79.2	732	13419.6	7870
	3	27.7	119.3	1094	20129	11805
2	4	36.9	159.4	1455	26839	15740
	5	46.1	199.5	1817	33549	19675
	6	55.3	239.6	2179	40259	23611
3	7	64.6	279.7	2540	46969	27546
	8	73.8	319.8	2902	53679	31481
	9	83	360	3264	60388	35416

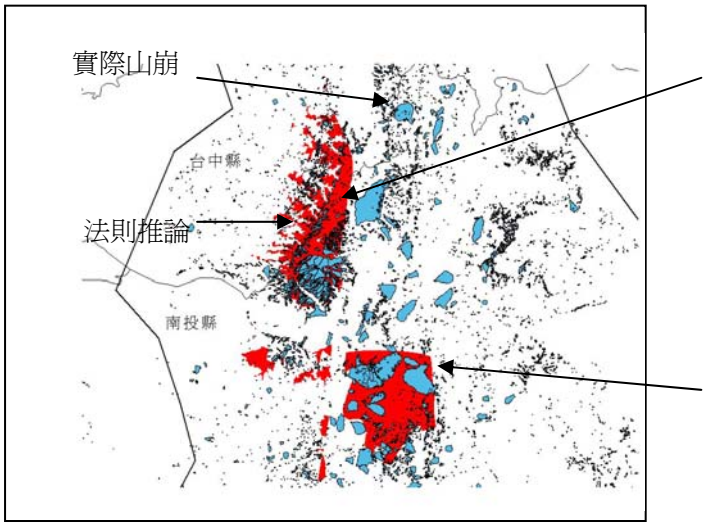
編碼		環境因子			
Level 1	Level 2	Dist2Rv(水系距) 單位：公尺	PGAEW(東西向地表加速度) 單位：gal	PGANS(南北向地表加速度) 單位：gal	PGAVER(垂直地表加速度) 單位：gal
1	1	917	101-200	87-161	81-130
	2	1834	299	234	179
	3	2752	397	308	227
2	4	3669	496	382	276
	5	4586	595	455	325
	6	5503	693	529	373
3	7	6420	792	603	422
	8	7338	891	676	471
	9	8255	989	750	519

表 3　單層知識架構強規律性之地震引致山崩關聯法則

項次	法則代碼表示	支持度	信賴度
1	PgaVer_6，PgaNS_6	7.2%	94.5%
2	PgaVer_6，PgaEW_5	8.7%	90.8%
3	PgaVer_6，DTM_2	7.1%	90.4%
4	Dist2Center_2，Geology_6	7.9%	93.5%
5	Dist2Fault_4，Geology_6，PgaEW_5	7.2%	90.0%
6	Geology_15，DTM_2，Dist2Riv_1	7.5%	90.1%
7	Dist2Center_2，Geology_6，PgaEW_5	7.9%	94.3%
8	Dist2Center_2，Geology_6，Dist2Road_1	7.9%	93.8%
9	Geology_6，PgaEW_5，Dist2Road_1	11.4%	90.0%
10	Dist2Fault_4，Geology_6，PgaEW_5，Dist2Road_1	7.1%	90.1%
11	Dist2Center_2，Geology_6，PgaEW_5，Dist2Road_1	7.9%	94.3%

表 4　多層知識架構強規律性之地震引致山崩關聯法則

項次	法則代碼表示	支持度	信賴度
1	Geology_15，SlpMean_L1_2	8.8%	92.8%
2	Geology_15，SlpMean_L1_2，Dist2Fault_L1_1	8.8%	92.8%
3	Geology_15，SlpMean_L1_2，DTM_L1_1	8.8%	92.8%
4	Geology_6，PgaEW_5，Slope_L1_1	8.8%	90.6%
5	Geology_15, SlpMean_L1_2，Dist2Fault_L1_1，DTML1_1	8.8%	92.8%
6	Geology_6，PgaEW_5，Dist2Cen_L1_1，　Slope_L1_1	8.8%	90.6%
7	Geology_6，PgaEW_5，Slope_L1_1，PgaEW_L1_2	8.8%	90.6%
8	Geology_6，PgaEW_5，Slope_L1_1，DTM_L1_1	8.8%	90.6%
9	Geology_6，PgaEW_5，SlopeL1_1，DTM_L1_1	8.8%	90.6%
10	Geology_6，PgaEW_5，Dist2Cen_L1_1，Slope_L1_1，PgaEW_L1_2	8.8%	90.6%
11	Geology_6，PgaEW_5，Dist2Cen_L1_1，Slope_L1_1，DTM_L1_1	8.8%	90.6%
12	Geology_6，PgaEW_5，Slope_L1_1，PgaEW_L1_2，DTM_L1_1	8.8%	90.6%
13	Geology_6，PgaEW_5，Dist2Cen_L1_1，Slope_L1_1，PgaEW_L1_2，DTM_L1_1	8.8%	91.6%



關聯法則 1：
(地質分佈 = 頭嵙山層)　且
(370 m <= 高程 <= 732 m)　且
(距水系距離 <= 900 m)
支持度：7.5%, 信賴度：90.1%

關聯法則 2：
(地質分佈 = 瑞芳群)　且
(6.7 km <= 距震央距離 <= 13.4 km)　且
(495 gal <= 東西向地表加速度 <= 595 gal)
支持度：7.9%, 信賴度：94.3%

圖 13　部份關聯法則所產生之山崩對應圖

這樣的結果與等級相關分析所獲得的結果相仿，但等級相關分析僅針對單一分子的分析，並未探討多個因子之間彼此的關聯性，透過關聯法則可以發現某些新的知識 (例如在等級相關分析中震央距最高相關者爲 20~26 公里，但在關聯法則中最高信賴度的多重因子組合中卻是 7~14 公里)，歸納出一組組可以理解的法則，這樣的法則不但可以建立成爲專家系統所需的知識庫，另外，也可以藉以產生對應的危險圖，劃定危險區域，

提供決策支援上的參考。圖 13 爲採用了二條規律性極強的法則所繪製出來的對應區域，關聯法則可以掌握重大的崩塌地部份。

結論與建議

(一) 本研究以地理資訊系統做爲資料的前處理工作，整合各種類型的資料，並透過分類編碼的方式將原始數值資料轉換爲類別資料，以便建立資料倉儲提供資料探勘所需的基本資料。本研究以自然分等法 (natural breaks) 做爲分等的方法，部份研究者曾針對數值資料提出其它分等方式 (Wang and Tay, 1998)，不同的分等方式，對於法則的產生將會有一定程度的影響，值得後續的研究。

(二) 等級相關分析可以探討各個單獨因子與山崩之間的相關性，雖僅係針對單一因子，但仍可提供後續資料探勘上的了解與初步假說的建立，做爲起始的探勘工具。

(三) 在關聯法則的分析中，透過最小信賴度及最小支持度的設定，可以定義研究所需的強規律性與有用的法則，找出最可能造成山崩的因子。唯若所訂的法則門檻值較低的話可能產生相當大量的法則，不易了解且某些法則間可能彼此互相矛盾或重複，未來可以考慮再進行 meta mining (由法則中再找法則) 的作業來精練。此外，將抽象層次拉高以取得更多的法則時，必須更小心處理，以避免產生容易誤導的規則。

(四) 經關聯法則分析後，發現瑞芳群及頭料山層的地質條件、高程位於 370 至 732 公尺之間、距震央距離在 7 至 14 公里間、距車籠埔斷層距離在 12 至 16 公里之間、距道路距離 850 公尺以內以及地震強度在 350gal 以上這些條件的組合是最常見於地震山崩中的關聯樣式，經由關聯法則的條件對應可以繪出危險區域提供防災決策上的參考。

(五) 本研究係以集集大地震區域相關地理資料庫做爲測試案例，故所發現之關聯樣式應只較適用於研究區域附近，未來遭受類似斷層錯動引起地震影響下所造成的山崩危險。礙於大範圍資料收集上的困難，目前所使用的因子仍屬有限，且部份資料來源年代較早 (如 DTM、道路與水系圖)，恐怕與當時的現況不甚相符，雖然說服力可能不是十分充足，但這樣的研究方法可以應用於日後其它具豐富地理資料庫的環境議題上。

引用文獻

洪如江、林美聆、陳天健、王國隆 (2000) 921 集集大地震相關的坡地災害、坡地破壞特性、與案例分析，地工技術，8: 17-32。

許煜煌 (2002) 以不安定指數法進行地震引致坡地破壞模式分析，國立台灣大學土木工程研究所碩士論文。

廖軒吾 (2000) 集集地震誘發之山崩，國立中央大學地球物理研究所碩士論文。

童啓哲 (2001) 應用地理資訊系統於地震引致坡地破壞多變量模式分析，國立台灣大學土木工程研究所碩士論文。

鍾新南 (2003) 建立自動化繪圖圖面配置中地圖圖元的關係，國立台灣大學地理環境資源研究所碩士論文。

鄒明城、孫志鴻 (2004) 資料探勘技術在集集大地震引致山崩之研究，國立台灣大學地理學報，36: 117-131。

Agrawal, R., Imielinske, T. and Swami, A. (1993) Mining association rules between sets of items in large database, *Proc. of ACM-SIGMOD 1993 Int. Conference of Management of data*, Washington, D. C, 207-216.

Agrawal, R. and Srikant, R. (1994) Fast algorithms for mining association rules, *Proc. of the 20th International Conference on very large database (VLDB)* , Santiago: Chile, 487-499.

Berry, M. and Linoff, G. S. (1997) *Data Mining Techniques for Marketing, Sales and Customer Support*, New York: John Wiley and Sons.

Chen, M.-S., Han, J. and Yu, P. S. (1996) Data mining: an overview from a database perspective, *IEEE Transactions on Knowledge and Data Engineering*, 8 (6) : 20-35.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds.) (1996) *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press.

Han, J. (1999) Data Mining, *Int. J. Urban and P., Encyclopedia of Distributed Computing*, Kuwer Academic Publisher.

Jenks, G. F. and Coulson, M. R. (1963) Class intervals for statistical maps, *International Yearbook of Cartography*, 3: 119-134.

Kangkachit, T. and Waiyamai, K. (2002) A business-oriented spatial association rule mining system prototype (BoSARM) , *Proc. Information and Computer Engineering Postgraduate Workshop (IECP 2002)* , Thailand.

Keefer, D. K. (1984) Landslides caused by earthquakes, *Geological Society of America Bulletin* , 95: 406-421.

Keissner, C. (1998) Data Mining for Enterprise, *Proc. of the 31st Annual Hawaii International Conference on System Science (HICCS 98)* , 295-304.

Koperski, K. and Han, J. (1995) Discovery of Spatial Association Rules in Geographical Information Database, *Proc. 4th Int. Symposium on Large Spatial Database.*

Koperski, K., Adhihary, J. and Han, J. (1996) Spatial data mining: progress and challenges survey paper, *SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery.*

Lu, W., Han, J. and Ooi, D. C. (1993) Discovery of General Knowledge in Large Spatial Database, *Proc. Far East Workshop on Geographic Information System*, Singapore, 275-289.

Miller, H. J. and Han, J. (eds.) (2001) *Geographic Data Mining and Knowledge Discover*, New York: Taylor and Francis.

Mennis, J. and Liu, J. W. (2003) Mining Association Rules in Spatio-Temporal Data, *Proc. GeoComputation 2003 (GeoComputation CD-ROM)* , Southampton, UK, 8-10 September 2003.

Refice, A. (2002) Probabilistic modeling of uncentainties in earthquake-infuced landslide hazard assesment, *Computer & Geoscience*, 28: 735-749.

Roiger, R. J. and Geatz, M. W. (2002) *Data Mining – A Tutorial-Based Primer*, Addison Wesley.

Tang, H. and McDonald, S. (2001) Spatial Data Mining and University Courses Marketing, *Proc. GeoComputation 2001 (*GeoComputation CD-ROM*)* , Brisbane, Australia, 24 - 26 September 2001.

Wang, K. and Tay, S. H. (1998) Interestingness-based interval merger for numerical association rules, *Proc. of the 4th International Conference on Knowledge Discovery and Data Mining*, 121-127.

Zhao, L. (2000) Integrating rank correlation techniques with GIS for marketing analysis, *Proc. GeoComputation 2000 (*GeoComputation CD-ROM*)* , Greenwich, U. K., 23 - 25 August 2000.

Zhu, A.-X. (1999) A personal construct-based knowledge acquisition process for natural resource mapping, *Int. J. Geographical Information Science*, 13 (2) : 119-141.

94 年 09 月 20 收稿
94 年 10 月 11 修正
94 年 10 月 28 接受