# 結合政府開放資料集於建構開業選址決策支援系統－以店址當地消費能力與鄰近產業特性觀點

# Combing the Government Open Datasets to Build the Location Selection Decision Support System －The Local Consumption Ability and Neighboring Industries Perspectives

杜逸寧、徐維澤、黃祥晉、許明楷、林鈺翔、洪健傑
Yi-Ning Tu, Wei-tse Hsu, Hsiang-Chin Huang, Ming-Kai Hsu, Yu-Hsiang Lin, Jian-Jie Hong

本研究希望結合開放資料集與現今的機器學習演算法建立一個「開業選址決策支援系統」，為了協助房東找到店面適合開的行業；並幫助創業者找到適合開店的位置。本研究為了收集店家資訊串接 3 個跨領域的政府公開資料；接著為了建立商圈資訊計算與公共場所和其他店家的距離；並考慮行業是否有群聚效應使用空間自相關分析 Moran's I；還利用隨機森林重要性找出影響每一個行業的關鍵因素，最後使用 *k*-最近鄰法做為推薦的依據。本研究在 Precision 指標中與其他熱門分類演算法進行比較至少高出 26.8%，足見本研究建立之預測模型相較其他演算法有一定的預測率。

This study hope to combine the open data and current machine learning methods to establish "location selection decision support system", which can provide suggestions to both landlords of the store, and the coming shopkeepers. For collecting store information, this work connects 3 different sources of open datasets. Furthermore, to quantify the surrounding information of a store, this study measures the distance between locations predicted to landmark or to each type of stores. For discriminating whether or not a type of store congregates, this work adopts Moran's I spatial autocorrelation analysis. spatial autocorrelation analysis. This study utilizes the Random Forest Importance to identify the key factors of 30 distinctive types of store, and apply *k*-nearest neighbour for the foundation of recommendation. As the results, this work shows in Precision, the proposed method is at least 26.8% higher than other classification algorithms.

# 一、前言

## 1. 研究背景

本研究欲建立針對多數行業能使用的開業選址決策支援系統。因為傳統的房仲業者以及較為新興的租屋網提供的資訊僅有房屋的訊息，缺乏地理上的商圈資訊。楊宜芬、孫志鴻、榮峻德 (2007) 指出利用地點的空間資料做為選址的考量；黃鵬達 (2011) 認為影響消費者上門的關鍵在於店址是否在人的生活動線上。另外王淑慧、于如陵 (2009) 指出商店在選擇店址時，會選擇與現有同類型商店群聚或遠離，前述研究利用地理資訊系統對店家進行空間分析。而目前選址相關的研究，如許富城 (2006)、廖千慧 (2006) 使用層級分析法 (Analytic Hierarchy Process, AHP) 來做選址因素的探討，而這種方法是以問卷的形式收集專家的意見，故十分仰賴問卷的設計及回收品質。

## 2. 研究動機與目的

故本研究期望延續前述研究建立自動化的店家選址決策支援系統，能夠利用政府各種不同的開放資料，將每個地點周邊居民可能的生活動線，例如：店家、公共場所等納入考量。此外，本研究也利用空間自相關分析 (Moran's I) 來了解不同行業的聚集狀況，並建立一個地區每個地點的商業環境。本研究所提出的雛型系統不僅能推薦房東店面所適合營業的行業，透過分析每個行業適合的商圈環境，推薦適合創業者開業的地點，並提供使用者選址的關鍵因素。同時解決創業者及房東的問題。本研究期望達到以下目的：

1. 結合不同領域多元的資料，將原本散落於各處的政府開放資料集像是：交通、地理、人口等資料，盡可能的整合以分析一個地點的環境特色。
2. 考量每個地點周圍的商圈環境資訊做為決策的依據。
3. 利用空間自相關分析 (Moran's I) 探討每個行業在空間內的群聚效應，區分出適合聚在一起或者適合分散的行業。
4. 找出每個行業的選址關鍵要素並加以量化其權重。

# 二、文獻探討

## 1. 空間分析

本研究透過空間分析中常用的空間加權矩陣以及 Moran's I 方法探討特定區域的產業是否群聚，並運用空間相關的資訊，建立評估是否適合開業的系統。

### (1) 空間加權矩陣

空間加權矩陣是透過空間統計整合空間之間的關係，獲取事物在空間上分布的規則或結構。許智宏 (2005) 認為：「能夠簡化空間單元的實際大小、長度等，接著針對連接的情形評估空間單元在空間中的結構關係，發展出一套可以量化空間元素的數值。」將空間關係透過空間加權矩陣來定義空間的權重。空間加權矩陣可以量化空間數據間存在的距離關係，公式如 (式 1)。

$$W_{ij} = \begin{bmatrix} W_{11} & \cdots & W_{1j} \\ \vdots & \ddots & \vdots \\ W_{i1} & \cdots & W_{ij} \end{bmatrix} \tag{1}$$

其中

$W$：任兩個地點之間的距離

空間加權矩陣的概念如下圖 1，假設有 A、B、C 三個地點，之間的距離分別是 $\overline{AB}$ =100，$\overline{AC}$ =10，$\overline{BC}$ = 1 空間加權矩陣的行和列的變數為地點與其他地點之間距離，左對角線為代表與自己本身的距離而皆為 0。

本研究所使用的空間關係為反距離矩陣。反距離矩陣就是將距離取倒數後的矩陣，左對角線因為考量距離的因素所以定義倒數後仍為 0。以圖 1 為例，經過倒數的加權後如圖 2。

| | A | B | C |
|---|---|---|---|
| A | 0 | 100 | 10 |
| B | 100 | 0 | 1 |
| C | 10 | 1 | 0 |

圖 1. 距離矩陣。

$$\begin{array}{c} \\ A \\ B \\ C \end{array}\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 0 & 0.01 & 0.1 \\ 0.01 & 0 & 1 \\ 0.1 & 1 & 0 \end{bmatrix} \end{array}$$

圖 2. 空間加權矩陣。

### (2) Moran's I

Moran's I 是由 Moran (1950) 提出，是出現最早、應用最廣的一個度量空間自相關的參數。Upton & Fingleton (1985) 定義「空間自相關是地圖資料的空間組織所呈現出來的特質，其特質為土地上的空間所代表數值具有系統性與組織性的分佈」。所謂的系統性與組織性是指兩物體之間存在某種關係而非隨機分佈。朱健銘 (2000) 認為「區域內空間自相關程度高，則相同特質的空間現象將聚集在一起，若自相關程度低，則空間現象可能分散於空間各處。」Anselin (1995) 整理並歸納區域型空間自相關的方法，提出 LISA (Local Indictors of Spatial Association) 解決全域型空間僅能解釋整體空間聚集現象，而將大範圍切成好幾個小區塊討論。因為本研究探討新北市新莊區內以「里」為單位的產業聚集狀況，因此在計算參數上所使用的是全域型空間自相關 Moran's I，如 (式 2)。

$$\text{Moran's I} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{2}$$

其中

$x_i$：表示對於一組第 $i$ 個目標的變數
$x_j$：表示對於另一組第 $j$ 個目標的變數
$\bar{x}$：表示該區域內空間單位平均數
$n$：表示該區域內空間單位的個數
$w_{ij}$：代表該區域空間加權矩陣

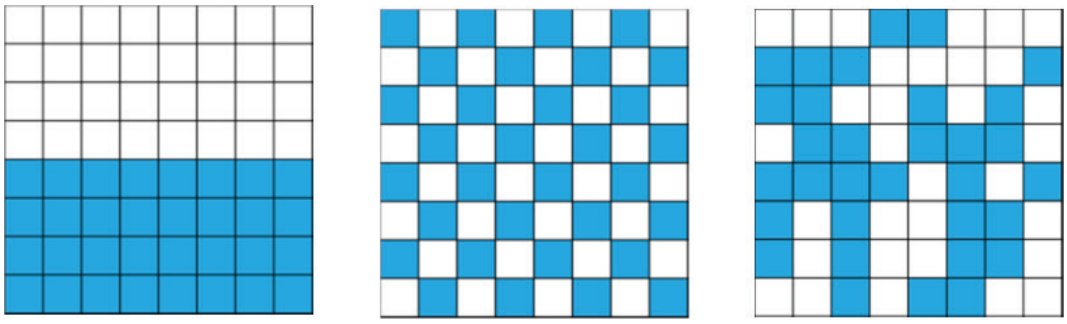
Moran's I 的計算公式，是基於統計學相關係數的共變數推得的，其值範圍介於 –1 到 1 之間，大於 0 為正相關，小於 0 為負相關，從圖 3(a)、3(b)、3(c) 為例，圖 3(a) 表示值越大表示空間分佈的聚集性越大，即空間上有聚集分佈的現象；反之，圖 3(b) 值越小則代表空間分佈聚集性小；圖 3(c) 而當值趨近於 0 時，即代表此時空間分佈呈現隨機分佈的情形。藉由 Moran's I 公式的計算後能得到配適值，由適配值可推得知一個地區的同一店種之間是否存在聚集的關係，並且能夠解釋店種的聚集效應，進而篩選出該店種適合的位置。

## 2. 分類器(Classifier)

本節將介紹 *k*-Nearest Neighbor、LDA、Naïve Bayes 和 Decision Tree 這四種分類演算法並且說明本研究使用這四種分類方法的原因，以及進一步探討這四種分類法的比較。

### (1) *k*-Nearest Neighbor

*k*-Nearest Neighbor又稱 *k*-最近鄰法，其概念根據資料點間的距離，選取 *k* 個相似度最高的資料點



(a) Moran's I = 0.857　(b) Moran's I = –1　(c) Moran's I = 0

圖 3. Moran's I 值圖像意義。

並以眾數決定目標類別。如下圖 4 為 *k*-最近鄰法概念圖，圓形表示為類別一的測試樣本，叉表示為類別二的測試樣本，並以三角形為質心分別以 $k = 1$ 和 $k = 3$ 進行 *k*-最近鄰法的分類法。首先當 $k = 1$ 也就是找距離質心最近的 1 個鄰居是 1 個圓，因此其代表意義為有 1 個類別一 0 個類別二的事件，因此根據 *k*-最近鄰法測試資料會被分為類別一的事件。然而當 $k = 3$ 的時候，距離圓心最近的 3 個鄰居為 1 個圓 2 個叉，其代表的意義為有 2 個類別二和 1 個類別一的事件，因此根據 *k*-最近鄰法測試資料會被分為類別二的事件。

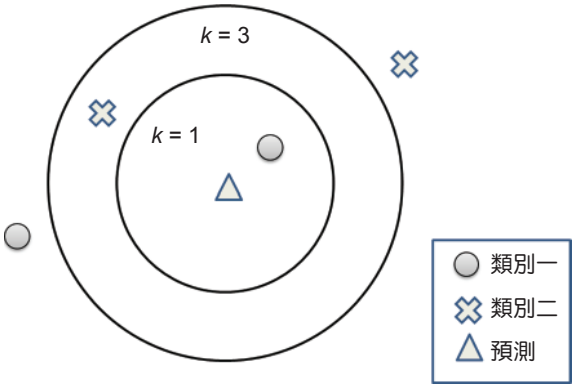



圖 4. *k*-最近鄰法概念圖。

### (2) LDA (Linear discriminant analysis)

Linear discriminant analysis (線性區別分析，簡稱 LDA) 是由 Fisher (1936) 提出之理論，其理論概念是以貝式理論 (Bayes Theorem) 為基礎，且滿足下列兩個條件 (1) 樣本屬於多變量常態分配 (2) N 個目標分類中的變異相同。其公式如 (式 3)。

$$\delta_k(\vec{x}) = -\frac{1}{2}\left(\vec{x} - \overrightarrow{\mu_k}\right)^T \sum\nolimits_k^{-1} \left(\vec{x} - \overrightarrow{\mu_k}\right) - \frac{1}{2}\log\left|\sum\nolimits_k\right| + \log\pi_k \tag{3}$$

其中

$k$：群
$\mu_k$：第 $k$ 群平均數
$\pi_k$：事前機率
$\Sigma_k$：第 $k$ 群的變異

當 $\delta_k(x)$ 最大時，可以找出最佳群數。因為 LDA 是統計驗證常見的方法，因此本研究使用 LDA 於驗證本研究模型的準確度。

### (3) Naïve Bayes

在貝式理論下，在 $y$ 條件下，多變量 ($X_1$, $X_2$, ... , $X_n$) 之間存在獨立關係，由於之間獨立因此將多個變量連乘，其假設表示如 (式 4)。根據 $X_1, X_2, \ldots, X_n$ 前提下，對第 $k$ 個目標變數 $y$ 的分類機率如 (式 5)，因為 (式 5) 要列舉所有可能需要窮盡宇宙時間才能窮舉完，因此根據貝氏理論，若 $X_1, X_2, \ldots, X_n$ 對於 $y$ 式條件式獨立，可以將假設條件配合貝式理論改寫成 (式 6) 簡化式子。

$$P(X_1, X_2, \ldots, X_n | y) = \prod_{i=1}^{n} P(X_i | Y) \tag{4}$$

$$P(Y = y_k | X_1, X_2, \ldots, X_n) = \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\Sigma_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)} \tag{5}$$

$$P(Y = y_k | X_1, X_2, \ldots, X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\Sigma_j P(Y = y_j) \prod_i P(X_i | Y = y_j)} \tag{6}$$

因為貝式理論也是統計驗證常見的方法且本研究期望多方面驗證，因此本研究使用 Naïve Bayes 驗證模型的準確度。

### (4) 決策樹

在資料探勘中，決策樹 (Decision Tree) 是常用於預測模型的演算法，可以將大量的資料進行分析找出資料內具有潛在價值的訊息，且能得知分類或預測正確率。決策樹以樹狀圖為基礎尋找規則，一筆資料從樹的根部開始進行，其中經過樹中的每一個內部節點 (Internal node) 代表測試某個屬性，其屬性下的分支 (Branch) 代表此屬性可能為單值或多值，最後到達的每個葉節點 (Leaf node) 代表從根經過節點、分支的最終的屬性。本研究整理 Berry & Linoff (1997)、Murthy (1998)、Zimenkov (2000) 等學者提出決策樹的優缺點比較。因為決策樹便於使用且能清楚解釋變數正確率，因此使用決策樹作為本研究模型的驗證。

**(5) 隨機森林 (Random Forest)**

Ho (1995) 提出隨機森林 (Random Forest)，隨機森林以決策樹為基礎，將巨量的決策樹組合成決策樹森林。假設資料集內有 $p$ 個特徵變量 (自變數) 表示為 $(x_1, x_2 \ldots, x_p)$，$y$ 為應變量 (依變數)，且應變量 $y$ 有 $N$ 個訓練樣本和 $p$ 個特徵變量。隨機森林採取重複抽樣的方法建構決策樹森林，隨機在訓練樣本中選擇 $N$ 個樣本，其中有些樣本會重複抽取多次，便形成一組新的訓練資料集，並且從這棵樹使 $p$ 個特徵變量中選擇部分變量進行預測，估計其誤差值，因此隨機森林每次建構的決策樹都不同。最後以重複最高次數的類別為預測結果。此外，隨機森林能夠對於變數進行評估重要性並計算出每個變數的權重，因此本研究使用隨機森林演算法進行篩選變數，得出最適變數權重。

## 三、開業選址決策支援系統

本章將逐步說明本研究所使用的方法。首先第一節說明研究流程，第二節介紹如何蒐集資料，並篩選與研究相關的資料，第三節說明資料預處理的做法以及展示如何將地理距離轉為權重變數。而第四節將闡述如何衡量自變數的重要程度並運用在 $k$-最近鄰法上。最後說明本研究所建立的兩個模型是如何進行計算及預測。

### 1. 研究架構與流程圖

本研究的研究架構與流程如圖 5 所示：

跨領域整合：為了提供使用者多元的資訊，本研究將不同領域的資料庫與店家的資料庫加以整合。從政府的開放資料中擷取資料，並將資料進行預處理，以便後續的計算與轉換。

量化商圈資訊：為了量化每個店家在地理上的商圈環境，本研究計算店家與店家之間的距離以及店家與新莊區地標的距離，轉換為遠近程度的變數並增加到本研究的資料庫中，以此當作每個店家周邊的商圈環境。

空間自相關分析：本研究考量到每個行業的分布情況會對開店造成一定程度的影響，在資料預處理後便能直接計算每個行業的 Moran's I 值作為行業分布情況的指標，並轉換為衡量每個行業「與同業遠近程度」變數重要性的權重。

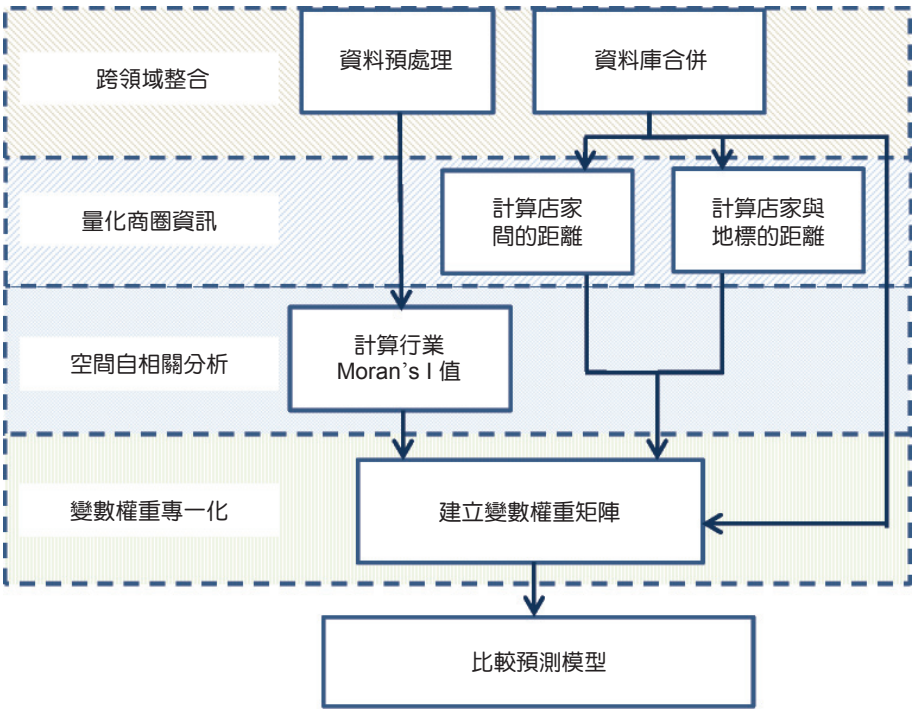



圖 5. 研究架構流程圖。

變數權重專一化：為了找出每個行業的關鍵因素，本研究在建立預測模型前，利用變數篩選演算法對每個變數計算重要程度並轉為權重值，再加上前一個步驟計算「與同業遠近程度」變數的權重值，建立每個行業專屬的變數權重矩陣。

本研究為了預測店家開店成功與否，利用政府公開資料 (Open data) 中的商業清冊得取所有公司的基本資料，資料中的變數包括統一編號、地址、行業、狀態等變數。其中統一編號是為了方便能夠將多個公司相關資料集合併的依據，而地址則是能夠建立空間上絕對位置的變數，行業則是區分店家的營業項目，最後狀態則能夠了解店家目前的營業現況。因本研究欲將資料更加結構化，本研究將原始資料中的狀態轉換為二元型態的資料，店家的狀態假如是歇業會遷出，本研究便將其歸類為失敗，其他則歸類為成功。

## 2. 計算距離與權重

### (1) 計算與公共場所之間的距離

為了達到量化商圈資訊的效益，本研究設法尋求公共場所的位置，並計算公共場所與各店家之間的距離。新北市 IMAP 為整合新北市府各機關開放的空間資料，透過地圖的方式讓使用者能夠快速瀏覽市府的資料與當地即時資訊。因此本研究於新北市 IMAP 中運用網路爬蟲的技術將所公共場所的資料取出，如表 1 所示，資料表中包括：公共場所的絕對位置、公共場所名稱以及其所屬分類。

本研究以歐幾里得距離作為計算公式。有了店家與公共場所之間的距離後，本研究希望對長距離之間的差異進行處罰，因此本研究將所有距離進行倒數，目的是希望將遠距離之間的差異縮小，並放大近距離之間的差異。

## 3. 建立模型

### (1) 自變數標準化

本研究在前一節預處理後的自變數皆為連續型變數，但是每個自變數的本身的變異程度有些不同，為了避免以上的問題，本研究將前一節預處理的自變數進行標準化的動作，公式如下：

$$x_i(\text{normalized}) = \frac{x_i}{\sigma_i} \tag{7}$$

其中

$i$：為第 $i$ 個自變數

$x_i$ (normalized)：為第 $i$ 個自變數標準化後的數值

$x_i$：第 $i$ 個自變數的數值

$\sigma_i$：第 $i$ 個自變數的標準差

### (2) 自變數加權矩陣

本研究認為，對於同一營業項目，預測是否適合營業時，每個自變數的重要程度會有所不同，以飲料業舉例，店家與學校的距離可能會比店家與醫院的距離重要。突顯自變數的重要程度最常見的方法就是給予權重。另外，同一個自變數，對於不同營業項目的重要程度也會不一樣。本研究定義一自變數權重矩陣 $W$，為每個自變數，針對不同營業項目的權重矩陣如式 (8)。

$$W = \begin{bmatrix} W_{11} & \cdots & W_{1j} \\ \vdots & \ddots & \vdots \\ W_{i1} & \cdots & W_{ij} \end{bmatrix} \tag{8}$$

其中

表 1. 公共場所原始範例資料。

| X | Y | 名稱 | 分類 |
|---|---|---|---|
| 292872 | 2768418 | 鴻金寶麻吉影城 | 運動場所 |
| 294866 | 2771188 | 新莊高級中學圖書館 | 圖書館 |
| 297033 | 2771812 | 全國電子化成門市 | 3c 連鎖 |
| 293079 | 2768227 | 農會西盛分部 ATM | atm |

$i$：表示第 $i$ 個自變數
$j$：表示第 $j$ 個營業項目
$w_{ij}$：表示矩陣中第 $i$ 個自變數，第 $j$ 個營業項目的權重值

**(3) 利用 Moran's I 進行加權**

本研究認為，每個行業與同行的地理距離，對於開店的影響是重要的。舉例來說，餐館業常常聚集在一起，相對有群聚的現象；機車行隨處可見，故相對沒有如此明顯的群聚關係，較接近隨機分布；而瓦斯行通常服務的範圍較大，客戶也較固定，幾個街區可能就只有一家。所以瓦斯行較傾向為相互遠離。由此可知對於餐館業及瓦斯行，與同行的距離相對是重要的，前者距離近較好；後者距離遠較好。而機車行與同業的距離，則相對不是很重要。故本研究利用空間自相關分析中的 Moran's I 法，將每個行業現實的分布情形以數值表示。

但僅利用行業的 Moran's I 值還不足以衡量此行業與同業的距離對此行業分布情況的重要性。為了提供判斷重要性的基準，本研究利用 Getis & Ord (1992) 提出將 Moran's I 進行假設檢定的方法，探討每個行業的分布是否是隨機分布。本研究對每個行業進行假設檢定，假設如下。

$H_0$：行業分布狀況是隨機分布
$H_1$：行業分布狀況不是隨機分布

若拒絕 $H_0$，接受 $H_1$，表示此行業在空間中的分布情況傾向是隨機分布，則本研究認為此行業與同業的距離無太大重要性；若拒絕 $H_1$，接受 $H_0$，表示此行業在空間中的分布情況傾向不是隨機分布，則本研究認為此行業與同業的距離具有相當重要性。

為了進行假設檢定，Getis & Ord (1992) 將 Moran's I 值做標準化的處理，算法如 (式 9)、(式 10)、(式 11) 所示。

$$Z_i = \frac{(I_i - E[I_i])}{\sqrt{E[I_i]}} \tag{9}$$

$$E[I_i] = \frac{-1}{n_i} \tag{10}$$

$$V[I_i] = E[I_i^2] - E[I_i]^2 \tag{11}$$

其中

$n_i$：第 $i$ 個行業在特定區域中的總個數
$I_i$：第 $i$ 個行業行業之 Moran's I 值
$E[I_i]$：第 $i$ 個行業 Moran's I 之期望值
$V[I_i]$：第 $i$ 個行業 Moran's I 之變異數
$Z_i$：第 $i$ 個行業 Moran's I 值之 z-score
$i$：第 $i$ 個行業

得到每個行業的檢定結果後，本研究希望針對每個行業與同業距離的重要程度進行修正。因為 Moran's I 的值介於 –1 到 1 之間，並不適合直接當作權重的值，故利用檢定的概念，轉為權重值對與同業的距離變數加權作為修正重要程度的方法，若該行業的權重值較大，代表與同業的距離是重要的，則放大與同業距離變數的重要程度；反之若該行業的權重值較小，代表與同業的距離並不太重要，則縮小與同業距離變數的重要程度。

本研究定義一 Moran's I 權重陣列 $W_I$ 為每個行業與同業距離的權重陣列，公式如 (式 12)，設定其中每個行業的權重值為標準化的 Moran's I 絕對值和信心水準 $Z$ 值的比值，如 (式 13)，比值越大代表此行業傾向不是隨機分布，與同業距離的重要程度高於信心水準；反之，比值越低代表此行業傾向隨機分布，與同業距離的重要程度低於信心水準。

$$W_I = [w_1 \ \ldots \ w_i] \tag{12}$$

$$w_i = \frac{|Z_i|}{Z_0} \tag{13}$$

其中

$w_i$ 為第 $i$ 個行業的 Moran's I 權重陣列值
$Z_0$：使用者自訂的 z-socre
$i$ 為第 $i$ 個行業

## 四、實驗結果

### 1. 跨資料庫整合

本研究利用 Dunne, Lusch & Gable (1995) 所提出的七個變數篩選構面於政府公開資料中蒐集能夠幫助研究的資料集，欲建立一個結合多領域的資料。表 2 為本研究所使用的公開資料集與對應的變數篩選構面。

- 戶籍人口資料：此資料來源為政府開放平台中各村 (里) 戶籍人口統計月報表，資料中包含全台灣各里的戶數、人口數、男女人口數以及各年齡的人口數。表中所包含的各里人口資料符合七個篩選變數構面中的人口因數。
- 教育程度：此資料來源為政府開放平台中各村里教育程度資料，資料中包含全台灣各里的教育程度程度人口數。此資料亦包含人口相關變數，因此符合七個篩選變數構面中的人口因數。
- 綜合所得稅各類所得：此資料來源為政府開放平台中綜合所得稅各類所得金額各縣市鄉鎮村里統計表－縣市別：新北市，資料中包含新北市各里的各類稅收。此資料符合七個篩選變數構面中的成本因素。
- 商業清冊：此資料來源為政府開放平台中商業設立、變更、歇業登記清冊 (月份)，資料中包含全台灣公司的統一編號、名稱、地址、類別以及狀態。此資料能夠幫助本研究計算店家與店家之間的距離關係，因此符合七個篩選變數構面中的競爭店因素。
- 新北市 IMAP：此資料為新北市政府整合各機關開放的空間資料，本研究使用網路爬蟲的技術，將新北市 IMAP 中屬於新莊區的公共場所資料抓下來。5 個變數中包含公共場所的絕對位置 (X、Y 座標)、名稱以及所屬公共場所類別 (兩個)。

蒐集五個政府公開資料集後，為了結合多領域多資料集的資訊，本研究將五個資料集合併，同時考慮到位置的背景因素與外部因素。在考慮背景因素與外部因素之前，為了建立新莊區所有公司資料，本研究將三個設立、變更、歇業的商業清冊合併，如圖 6 所示，因為三個商業清冊皆有統一編號，因此將三個商業清冊依據統一編號合併為公司資料。

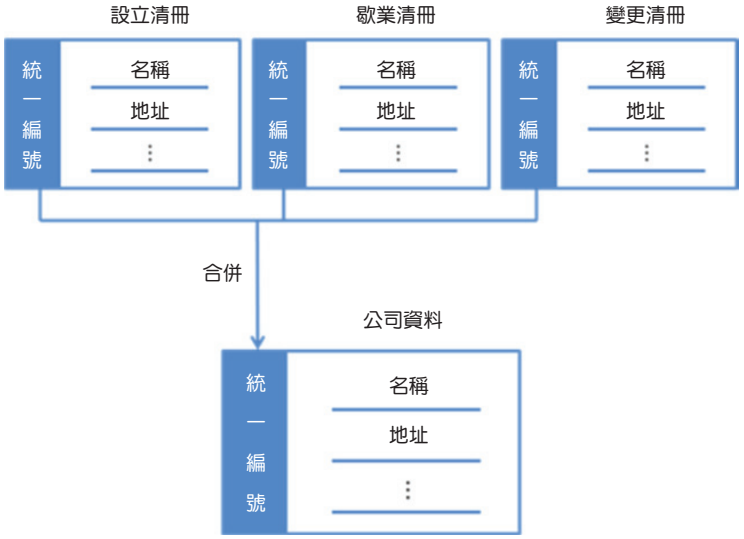



圖 6. 商業清冊合併。

背景因素：本研究將里的三個相關資料集 (人口資料、教育程度、各類稅收) 以變數里合併成里資料。並使用「里別」作為合併資料的欄位，這時已考慮一家店本身背景的資訊，如圖 7 所示。因為政府開放資料集中並沒有直接明確的資料表或者變數去直接反應「主要流動人潮花費水平的概念」。

表 2. 變數選擇對應資料來源圖。

| 資料集 | 所屬構面 | 變數個數 | 原始筆數 | 新莊區筆數 |
|---|---|---|---|---|
| 戶籍人口 | 人口因素 | 209 | 7851 | 84 |
| 教育程度 | 人口因素 | 50 | 7851 | 84 |
| 綜合所得稅／各類所得 | 成本因素 | 19 | 1092 | 84 |
| 商業清冊 | 競爭店因素 | 5 | 30017 | 10277 |
| 新北市 IMAP | 交通因素<br>地點特徵因素 | 5 | 少筆數 | 4289 |

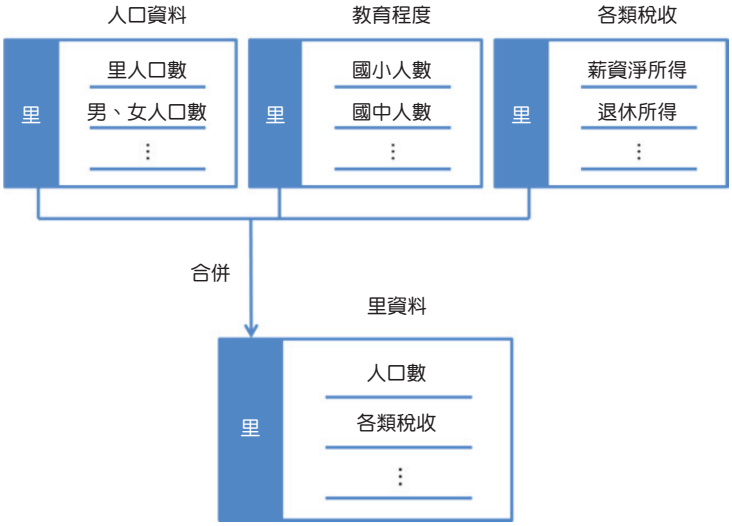



圖 7. 背景因素合併。

因此本研究試圖以背景因素當中以里為單位串接其「人口資料、教育程度、各類稅收」反應當地消費能力。

外部因素：也就是每個店家附近的商圈分布。如圖 8 所示，為了量化商圈的資訊，本研究將公共場所的資料納入資料表中，計算所有公共場所與每一家店之間的距離關係，本研究更考慮公司資料表中的交互關係，並計算店家與店家之間的距離關係。

圖 9 所示，最後將五個資料集合併後進行資料預處理 (刪除不必要變數、整併相似變數、轉換變數)，便可整理為本研究分析用的跨領域資料表。

## 2. 模型結果比較

本節將比較各個步驟對預測模型準確率的差異，本研究設定若預測某一地點適合的營業項目曾經在此地點開店，則判定預測正確；反之，若預測某一地點適合的營業項目不曾在此地點開店，則判定為預測錯誤。根據過去的文獻，每一家店是否持續經營的背後關鍵因素非常複雜，有本身專業技術層面、經營管理或資金的問題。但這些資料並無法在政府開放資料及中揭露出來。政府開放資料集僅能揭露為客觀環境因素的部分。因此本研究認為是否推薦成功的方式為曾經有業主願意投入大量資金於同一地址同一產業，對於預測將近三十種的產業

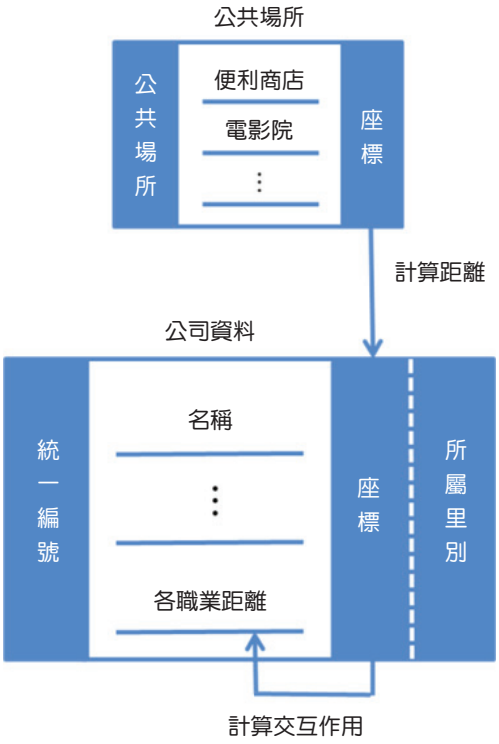



圖 8. 外部因素整併。

的選擇而言相對嚴格，此外本研究使用各種模型與毫無環境資訊而言對比的盲猜作為比較。實驗架構如圖 10 所示，(1) 比較自變數是否加權，(2) 相似度是否加權，(3) Moran's I 是否加權，(4) 跟其他演算法建立的模型進行比較。

### (1) 自變數加權比較

為了放大重要自變數的差異，縮小不重要自變數的差異。本研究使用 Information Gain，Chi-square 以及 Random Forest 三種篩選變數的方法，能在預測某一營業項目時，給予每一個自變數相對應的權重。使較重要的自變數有較高的權重，較不重要的變數則有較低的權重。

圖 11 為三種變數篩選方法及自變數未加權的情況下，三大類自變數權重比例之比較，其中地標因素和店家因素合稱外部因素。其中外部因素皆為超過一半的比例，顯示外部環境與店家的距離為非常重要的因素。

取得三種變數權重後，本研究將加權後的變數針對營業項目建立預測模型，方法為相似度未加權

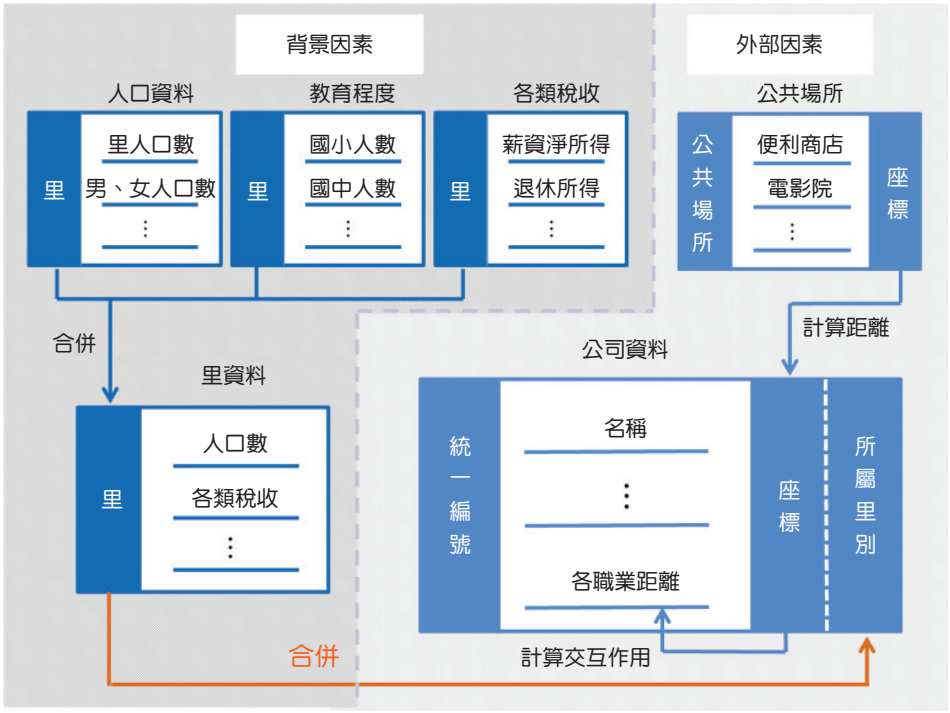



圖 9. 跨領域資料集合併圖。

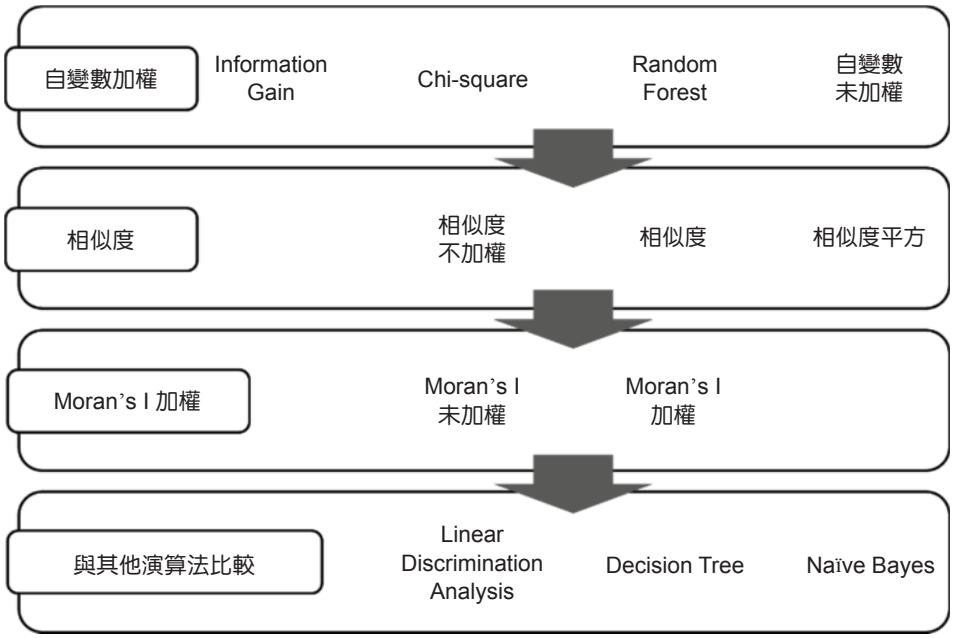



圖 10. 實驗架構圖。

*k*-最近鄰法，可以得到四種不同的模型。而圖中的隨機預測為本研究設定隨機預測開店成功的機率，作為四種模型比較的基準線，如 (式 14) 所示，每個行業隨機預測為成功的機率，各自乘上本身行業占全部行業總數的比例，最後加總得到加權平均後的隨機預測機率。

$$p\,(\text{隨機預測}) = \sum_i^k p_i \times w_i \tag{14}$$

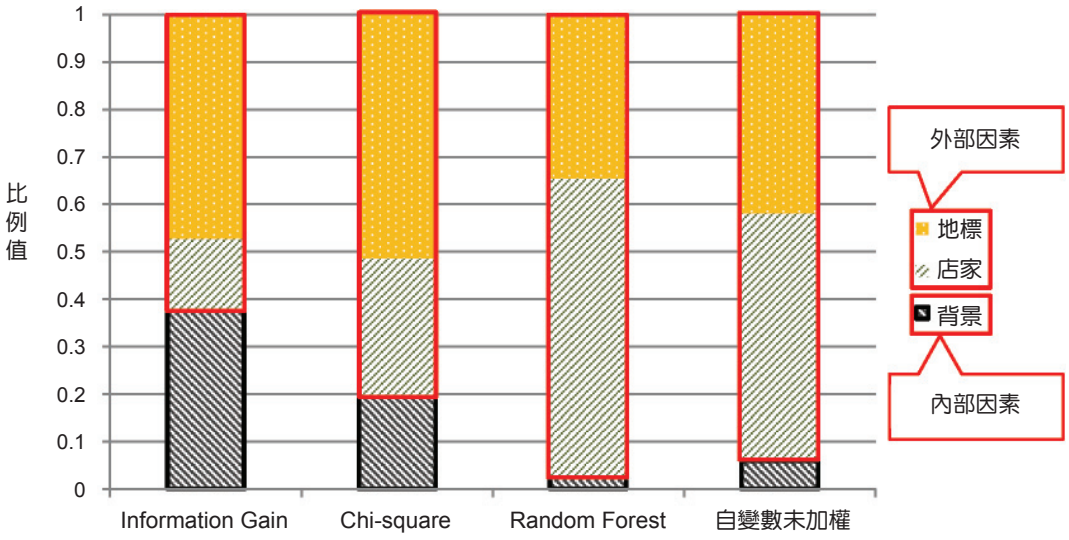



圖 11. 篩選變數方法比較圖。

其中：

$k$：行業總數
$i$：第 $i$ 個行業
$p_i$：第 $i$ 個行業隨機預測為成功的機率
$w_i$：第 $i$ 個行業占所有行業的比例

圖 12 結果顯示 Information Gain、Chi-square 及無變數加權三種模型無太大差異，僅優於隨機預測；Random Forest 的模型在所有的 $k$ 值都有所有模型中最好的預測結果。故本研究選用 Random Forest 為變數篩選及加權的方法。值得注意的是，所有的模型隨著 $k$ 值的增加，F-mesare 的值有呈現下降的趨勢，表示 $k$ 值越小，預測的越準確。代表變數最相似的資料點的行業，即為預測的結果。

**(2) 相似度加權比較**

在計算 $k$-最近鄰時，本研究認為與目標點越相似的資料，其對於判斷適合營業項目的重要程度越大。故本研究以相似度作為權重，對營業項目進行加權。

本研究分別建立了相似度不加權、相似度加權、相似度平方加權的預測模型。相似度加權、相似度平方加權的公式如 (式 15)、(式 16) 所示：

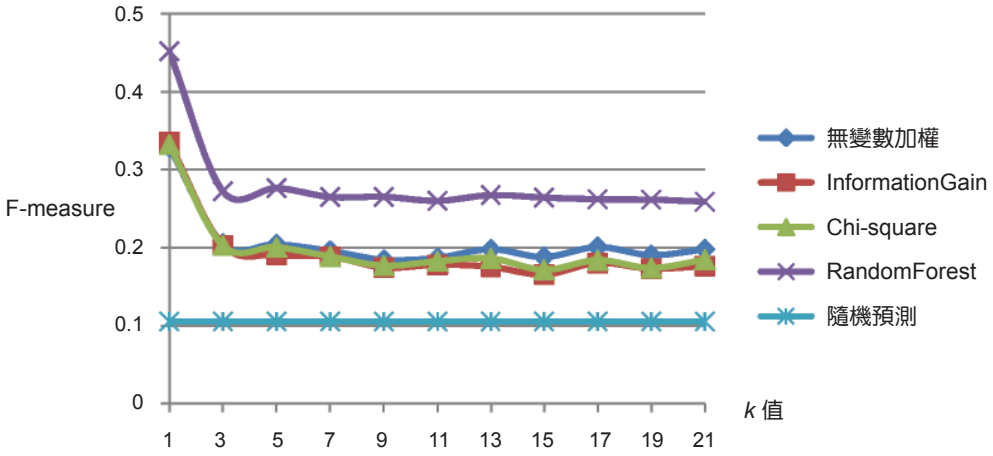



圖 12. 各權重模型 F-measure 比較 ($k$ = 1~21)。

相似度加權公式：$$\frac{1}{\sqrt{\sum_{j=1}^{p}(X_j - x_{ij})^2}} \quad (15)$$

相似度平方加權公式：$$\left(\frac{1}{\sqrt{\sum_{j=1}^{p}(X_j - x_{ij})^2}}\right)^2 \quad (16)$$

其中

$X_j$：輸入資料點的第 $j$ 個變數
$x_{ij}$：第 $i$ 筆樣本的第 $j$ 個變數
$p$：變數總個數

圖 13 及圖 14 分別為 $k$ 值為 1 到 21，間隔為 2 的情況下，三種模型的 Precision 和 F-measure 的比較。一樣是相似度平方加權有最好的預測結果，相似度加權次之，無相似度加權的模型僅優於隨機預測的結果。而在相似度平方加權且 $k = 1$ 的模型中，Precision 達到 0.537，F-measure 達到 0.452，是當中預測結果最好的模型。故本研究選擇使用相似度平方加權的方法。與前一小節的結果類似，所有的模型隨著 $k$ 值的增加，Precision 及 F-measure 的值都呈現下降的趨勢。

### (3) Moran's I 加權

圖 15 為本研究針對 30 個行業項目加入 Moran's I 加權 (Moran's I Weighted) 與未加權 (Original) 的 Precision 值比較圖，圖中 Moran's I 加權大部分都比未加權的 Precision 值來的好，唯有便利商店業比較差。因此本研究針對 Moran's I 加權與未加權對行業項目個數進行加權平均，圖 16 整理了不加權、隨機森林加權、以及隨機森林和 Moran's I 皆加權的 Precision、Recall、F-measure

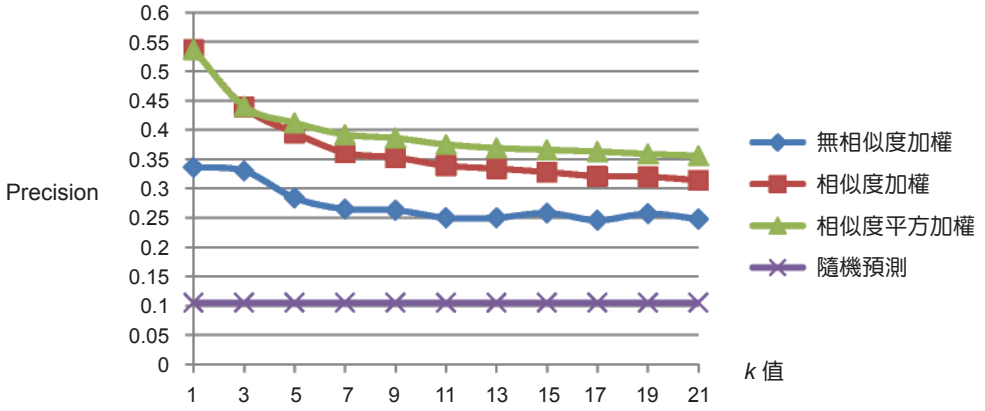



圖 13. 各模型 Precision 比較 ($k$ = 1~21)。

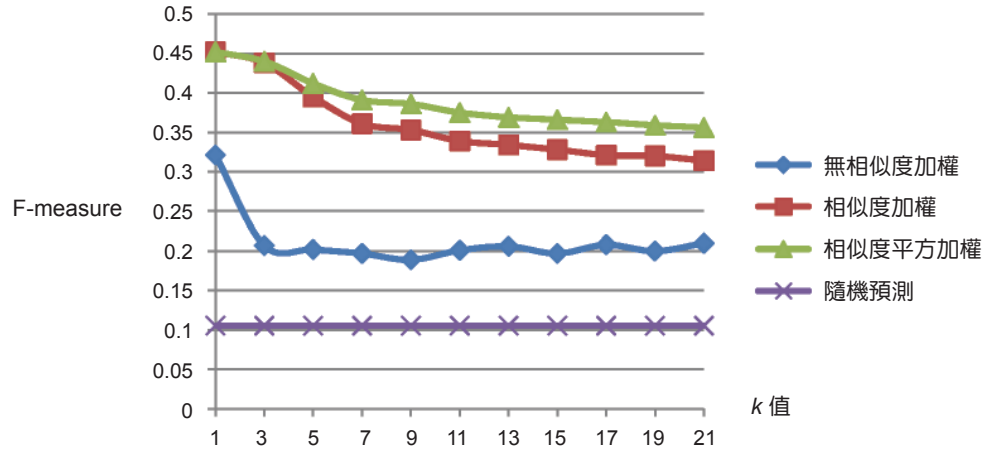



圖 14. 各模型 F-measure比較 ($k$ = 1~21)。

比較圖，發現若直接使用隨機森林重要性加權，其 Precision 雖然較高，但是 Recall 和 F-measure 皆是下降的，但如果加入 Moran's I 加權，其結果 Precision， Recall 和 F-measure 均是三種方法之中最好的。

### 3. 演算法比較

依據前一節所得到的結果發現在使用 Moran's I 加權與相似度平方加權可以得到最佳的 Precision 與 F-measure 值，接著本研究針對不同的演算法進行計算並比較在 Precision、Recall、F-measure 下的指標。

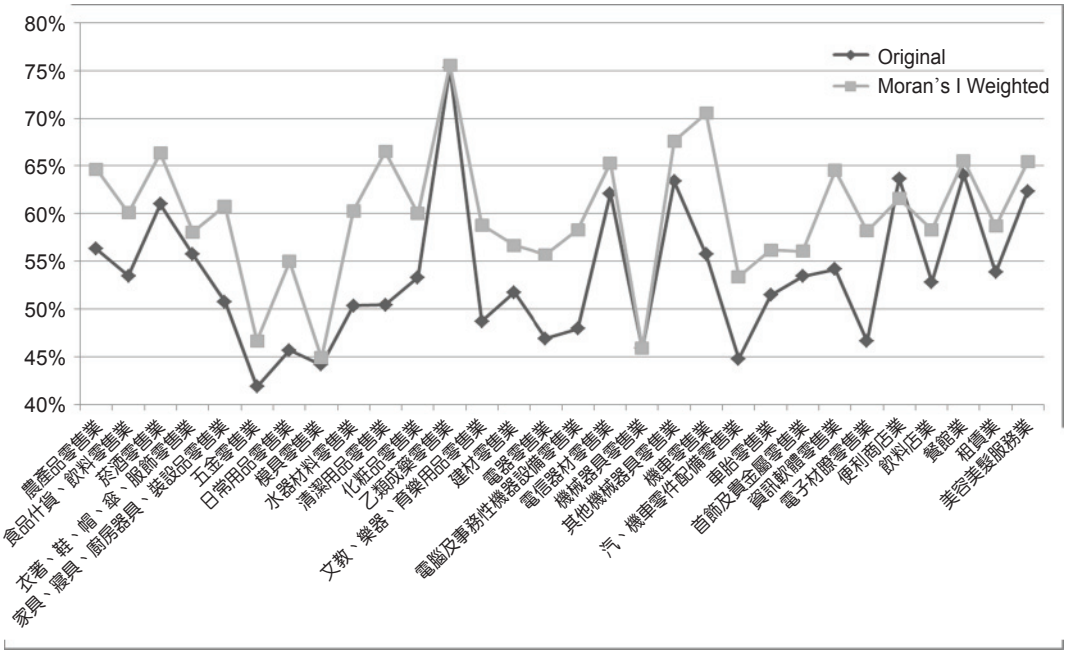



圖 15. Moran's I 加權與否 Precision 各行業項目比較圖。

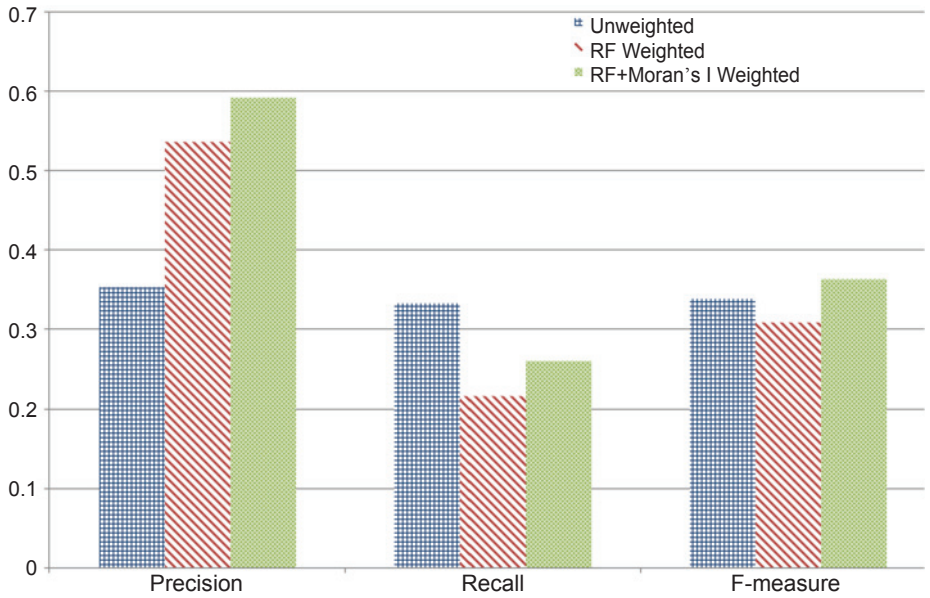



圖 16. 隨機森林 (RF) 及 Moran's I 加權之 Precision、Recall、F-measure 比較圖。

本研究針對資料集使用線性判別分析 (Linear Discrimination Analysis)、決策樹 (Decision Tree) 與單純貝氏分類器 (Naïve Bayes) 三種不同得演算法與 Moran's I 加權 (Moran's I Weighted) 以及隨機預測 (Random Prediction) 進行 Precision、Recall、F-measure 的比較。

於本研究中，Precision 值代表在演算法的預測次數中，預測到成功開店的比例，而 Recall 值代表在所有開店成功的個數中，演算法預測到成功開店的比例。所以假如演算法的預測次數較多，預測到實際開店成功的次數也會較多，Recall 值也會較高。

圖 17 為不同演算法 Precision、Recall、F-measure 的比較圖，圖中的數值是以三十個營業項目依據資料筆數進行加權平均後的 Precision、Recall、F-measure。從圖中可見本研究在使用任何演算法計算後的結果皆比隨機預測好，雖然 Naïve Bayes 在 Recall 中有最佳結果，而在 Moran's I 加權中 Precision 與 F-measure 有最佳結果。

圖 18 為各演算法的預測次數圖，由圖中可發現 Naïve Bayes 預測為適合開店的結果遠高於其他演算法，因此有較高的 Recall 值，因為此問題 Naïve Bayes 的 Precision 值沒有很高。因為本研究欲了解在所有預測中預測為成功開店的機率有多高，因此本研究應著重 Precision 值中的表現，可發現在即使在沒有任何加權的情況下，*k*-NN 在 Precision 和 F-measure 已經有最佳的表現，證明是最適合解決此類問題之演算法。除此之外，本研究加上了隨機森林重要性以及 Moran's I，更改進了預測的結果，如圖 19 所示

## 五、研究結論

本研究希望能建立開業決策支援系統，提供開店選址的建議，同時解決房東與業主兩方使用者的問題。

本系統達到以下目標：

**(1) 多領域整合**

本研究從政府的開放資料中量化每個行業在地理上的各種性質，並將七個不同面向的資料庫加以串聯，形成一個包含店家、地標和背景三大因素，共 250 個變數的資料庫，再利用統計工具進行分析。本研究整合地理、統計以及零售管理三大領域，以更加全面的觀點提供使用者在營業管理上可靠的建議。

**(2) 量化商圈資訊**

本研究將店家與店家之間，店家與公共場所之間的交互關係加以量化，也就是建立店家的所屬的商圈網絡資料，讓使用者了解自己想開的店附近的商圈性質。

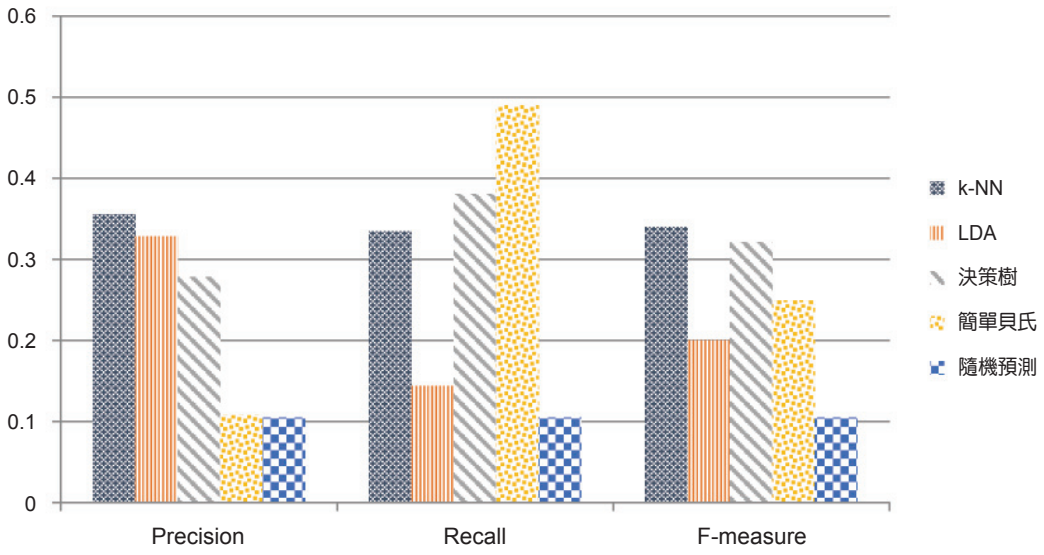



圖 17. 不同演算法 Precision、Recall、F-measure 比較圖。

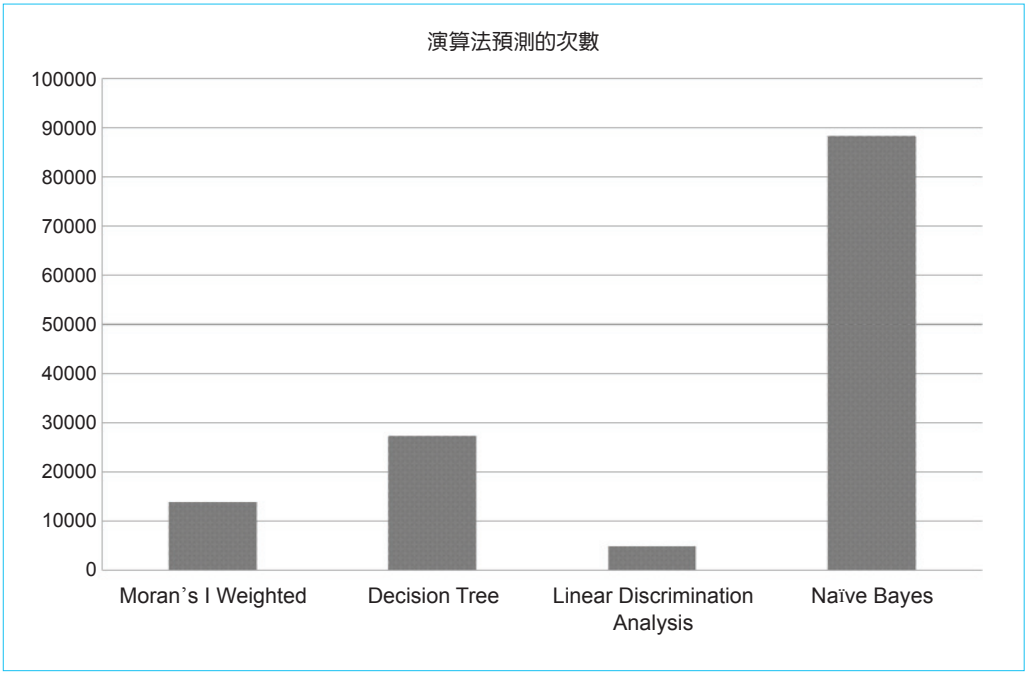



圖 18. 各演算法預測次數圖。

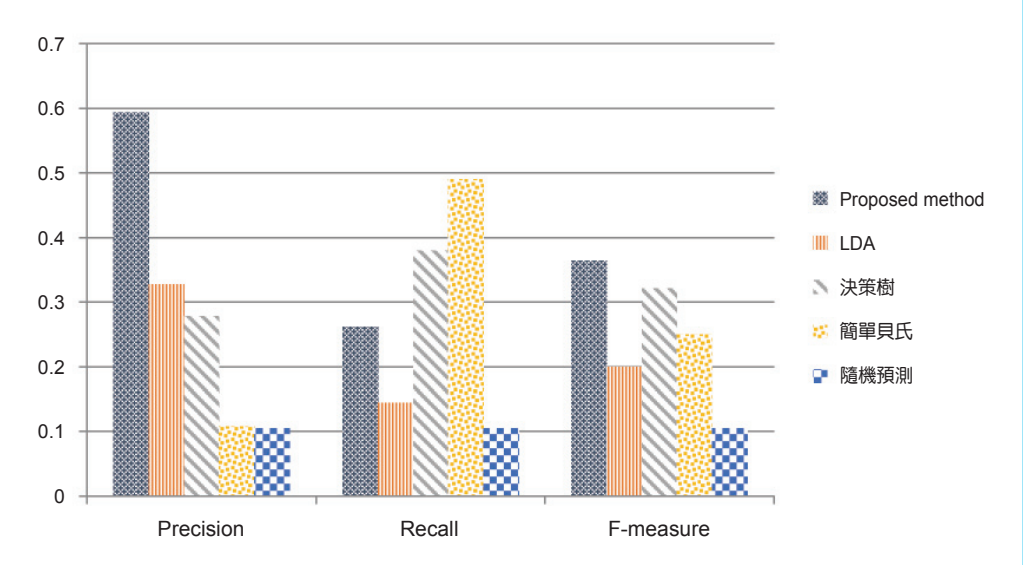



圖 19. 提案方法與其他演算法 Precision、Recall、F-measure比較圖。

**(3) 變數權重專一化**

本研究透過資料探勘中的區別分析，找出每個行業開店最關鍵的要素並放大這些要素的重要性，不僅能提高預測的準確度，更能讓使用者了解自己想開的店應當關注的核心要點。本研究比較出使用 Random Forest 篩選變數及加權的模型，其 Precision 達到 0.536，比隨機預測的 0.105 高出 43.1%。

**(4) 空間自相關分析**

本研究使用空間自相關分析中的 Moran's I 法將每個行業在空間上的整體分布情況量化，作為每

個行業空間上飽和度的指標。本研究使用 Moran's I 加權後，其 F-measure 比起 LDA 模型高出 16.4%，比 Decision Tree 模型高出 4.3%，比 Naïve Bayes 模型高出 11.4%。證明本研究的模型較其他傳統演算法優秀。

**(5) 本決策支援系統提供之輸出資訊**

本系統可以分為房東輸出頁面與創業者輸出頁面。其中，創業者輸出頁面可分為三個部分：推薦的店面、商圈性質及租屋情況。第一部分：推薦店面結果顯示使用者在選擇該產業且該地點中，系統所推薦給他們的店面位置，如圖 20 所示。第二部分：商圈性質顯示了使用者在選擇該地點中，附近方圓五百公尺內的行業種類多寡的泡泡圖，如圖 21 所示。第三部分:租屋情況顯示了在使用者所選該地點附近的租金行情，如圖 22 所示。房東輸出頁面也分為三個部分：推薦的營業項目、商圈性質及租屋情況。第一部分：推薦營業項目結果顯示使用者選擇該地區中，系統所推薦這店面所適合的營業項目，如圖 23 所示。第二部分及第三部分呈現方式則與創業者介面相同。

**(6) 與現存之選址系統之比較**

現存提供開業業主使用之相關系統有 Smart SAS 與經濟地理資訊系統 (GIS)，但僅能提供該地理環境之相關政府開放資料集的交叉分析。意即提供使用者單就地理為址所能得到的資料集串聯，讓業主可以根據自身需求決定開店位置。但當業主本身為新進入者且沒有足夠的背景知識時，本研究所提供之系統可以協助不論業者或者是房東，提供其機器學習後最為配適之媒合產業或開業地點。此外，本研究可以同時針對 30 種行業針對某依行政地區做選址的推薦預測。在政府開放資料集之前，選址系統大多仰賴專家學者以問卷訪談方式進行。現今許多房仲業者根據自身委託之案件可以提供地理資訊系統讓使用者選擇，但受限於受委託之案件


圖 20. 創業者輸出頁面－推薦店面。


圖 21. 創業者輸出頁面－商圈性質。


圖 22. 創業者輸出頁面－租屋情況圖。


圖 23. 房東輸出頁面－推薦的營業項目。

數，本研究可以針對所有於政府登記有案的地址皆作為預測的範疇，使得開店的決策可以是由出租方或承租方互相考慮而非單向受限選擇。

最後，許多新聞報導宣稱使用大數據作為選址之決策依據，但沒有公布其研究方法以及蒐集之變數，本研究透過蒐集串接六大資料集 (戶籍人口、教育程度、綜合所得稅與各類所得、商業清冊、新北市 IMAP) 處理新莊區資料筆數近 10277 曾經開業的店家，並使用多種機器學習演算法進行權重篩選與正確率比較，相信能針對使用者需求提供決策支援。

本研究之限制為：本研究在設計時以新莊區為例，不論是否為山區部分或者市區部分，只要登記有案的店址，本研究都會將其作為預測的資料點。結果顯示，資料仍趨向和其他業者鄰近比較容易有開業的機會。意即絕大多數的產業目前開業的決定仍傾向加入原有已經產生商圈的附近。但這樣的方法的確無法判斷是尚未開發的市場還是不需要此種服務。本研究之方法只能就目前資料的現況作為描述和預測，無法判斷其背後的因素。

此外，一個店家的汰換率高或低因素相當複雜，本研究相信商圈競爭激烈時反而是一個問題。但本研究所採用之資料集中涵蓋新莊區也有輔大學校商圈的特性。同一店址短期間不斷汰換店家可能主要的原因並非只是競爭激烈。承租該店址店家本身的經營條件，技術能力，價格商譽等等可能更是重要因素，但受限資料的蒐集無法窺見店家本身經營條件的資訊，只能透過店址之間地理位址以及本身環境的背景資料 (人口資料、教育程度、各類稅收) 去做分析。

## 參考文獻


1. 王淑慧, 于如陵, 遠東學報, **26** (4), 541 (2009).
2. 朱健銘 (民 88), 土地利用空間型態之研究 (未出版之碩士論文), 國立台灣大學, 台北市.
3. 許智宏 (民 94), 都市混和土地使用型態及其影響因素之研究－以台南市為例 (未出版之碩士論文), 國立成功大學, 台南市.
4. 許富城 (民 94), 國際速食連鎖店選址之研究 (未出版之碩士論文), 國立臺北科技大學, 台北市.
5. 黃鵬達 (民 99), 選址作業模式之研究－以連鎖早餐店為例 (未出版之碩士論文), 國立台北大學, 新北市.
6. 楊宜芬, 孫志鴻, 榮峻德, 中國地理學會會刊, **38**, 45 (2007).
7. 廖千慧 (民 94), 零售連鎖業店址選擇因素之研究: 以連鎖超級市場為例 (未出版之碩士論文), 中國科技大學, 新竹縣.
8. Anselin, L., Spatial Econometrics: *Methods and Model*, Vol. 4, Berlin: Springer Science & Business Media, (1988).
9. Anselin, L., *Journal of Geographical analysis*, **27** (2), 93 (1995).
10. Anselin, L., *Journal of Housing Research*, **9** (1), 113 (1998).
11. Berry, M. J., & Linoff, G., *Data mining techniques: for marketing, sales, and customer support*, New York: Wiley, (1997).
12. Dunne, P., Lusch, R., & Gable, M., Retailing Cincinnati, OH: SouthWestern Publishing Co, (1995).
13. Getis, A., & Ord, J. K., *Geographical analysis*, **24** (3), 189 (1992).
14. Ho, T. K., "Random decision forests", *Proceedings of 3rd International Conference on Document Analysis and Recognition*, August 14-16 (1995).
15. Moran, P. A. P., *Biometrika*, **37**, 17 (1950).
16. Murthy, S. K., *Data mining and knowledge discovery*, **2** (4), 345 (1998).
17. Upton, G., & Fingleton, B., *Spatial data analysis by example. Volume 1: Point pattern and quantitative data*, New York: Wiley, (1985).
18. Zimenkov, A., Tree classifiers, Department of Information Technology, Lappeenranta University of Technology, (2000).


---


杜逸寧女士為國立政治大學資訊管理研究所博士，現為輔仁大學統計資訊學系副教授。

Yi-Ning Tu received her Ph.D. in management information system from National Chengchi University. She is currently an associate professor in the Department of Statistics and Information Science at Fu Jen Catholic University.



徐維澤先生現為國立臺北大學統計所碩士生。

Wei-tse Hsu is currently a M.S. student in the Department of Statistics at National Taipei University.



黃祥晉先生為私立輔仁大學統計資訊學系學士。

Hsiang-Chin Huang received his B.S. in statistics and information science from Fu Jen Catholic University.

許明楷先生為私立輔仁大學統計資訊學系學士。

Ming-Kai Hsu received his B.S. in statistics and information science from Fu Jen Catholic University.

林鈺翔先生為私立輔仁大學統計資訊學系學士。

Yu-Hsiang Lin received his B.S. in statistics and information science from Fu Jen Catholic University.

洪健傑先生為私立輔仁大學統計資訊學系學士。

Jian-Jie Hong received his B.S. in statistics and information science from Fu Jen Catholic University.