**107年「培養數據新視野 翻轉城市競爭力」訓後成果競賽**

**成果報告書**

**主題：登革熱孳生源與不動產價位、都市計畫使用分區之關係**

**Analysis Between Property Value, Urban City Planning and Dengue Fever**

**成員名單及方案執行任務分工表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 一、臺南市政府學員 | | | | |
| 局處(公所)名稱  中文：臺南市南區區公所  英文：South District Office, Tainan City | | | | |
| 編號 | 姓名 | | 負責事項 | |
| 1 | 中文：朱棟  英文：CHU TUNG | | 督導「培養數據新視野 翻轉城市競爭力」提案構想及督導指揮等相關事項 | |
| 2 | 中文：王雪玲  英文：WANG HSUEH LING | | 指導「培養數據新視野 翻轉城市競爭力」提案構想數據收集、資料管理等事項 | |
| 3 | 中文：張育萍  英文：CHANG YU PING | | 執行收集統整「培養數據新視野 翻轉城市競爭力」大數據資料處理事項 | |
| 4 | 中文：黃向慈  英文：HUANG HSIANG TZU | | 處理「培養數據新視野 翻轉城市競爭力」大數據資料整理及其他紀錄事項 | |
| 二、國立成功大學學員 | | | | |
| 編號 | 系所 | 姓名 | | 負責事項 |
| 1 | 中文：電機工程研究所  英文：Electrical Engineering | 中文：林易瑩  英文：LIN YI YING | | 資料處理、報告撰寫 |
| 2 | 中文：電機工程研究所  英文：Electrical Engineering | 中文：鄭宇傑  英文：Zheng Yu Jie | | 資料蒐集、網頁架設 |
| 3 | 中文：資訊工程研究所  英文：Computer Science and Information Engineering | 中文：梁智堯  英文：Liang Zhi Yao | | 資料處理、資料分析 |
| 4 | 中文：資訊工程研究所  英文：Computer Science and Information Engineering | 中文：吳昊  英文：WU HAO | | 資料分析、報告撰寫 |

1. **提案緣由**

民國104年台南市爆發嚴重登革熱疫情，全市共計發生22,754病例，其中在南區(下稱本區)計3,507例。

104年本區疫情分布區域，在田寮、金華、文南、新興……等里之病例數前10名之里別，總計病例數達2,476例，佔本區總案例數的70%，其中大多屬都會區(圖1)。於此，本提案欲探討登革熱疫情的發生與都市、地價、都市使用分區等因素是否具相關性。



圖 1 臺南市南區里別分區，紅圈處為登革熱好發之都市區

1. **提案目的**

本提案之主要目的有三：

1. 分析不動產價位與登革熱患病之間的關係，判斷登革熱疫情是否好發於不動產價位較高的區域。
2. 分析區域都市規畫使用分區與登革熱患病之間的關係，如市場、公園、空地等分布與數量，是否對於登革熱患病數有影響。
3. 分析患病情形較為特殊的區域，找出該區域中可能會影響患病數或鮮為人知的要素。

綜合以上各種分析，將所獲得的結果應用於各階段的防疫作為，例如針對登革熱高風險環境做有效的管理並對防疫資源做更適當的配置，以避免疫情之發生；當有零星病例出現時，以更有效的方式防止疫情擴散。

1. **執行過程**

**資料集**

1. 病患資料

南區區公所及衛生局所提供的病患資料為民國104年臺南市南區登革熱病例，數量總計3516筆(人)，經過前處理後所留用的資料數為3463筆。

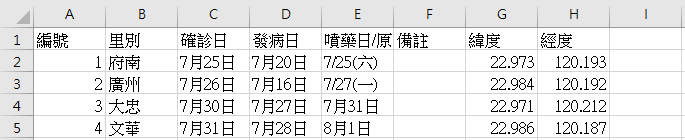
病患資料包含了病患的里別、確診日、發病日、噴藥日以及病患住家的經緯度，由於資料經去識別化，以致經緯度之精細度保留在小數點後三位。資料之範例格式見圖2

圖 2 病患資料範例

1. 地價資料

為了分析地價與登革熱發生之關係，本組一開始使用不動產實價登錄公開資料做查詢，但礙於實價登錄資料為交易紀錄，所以大多數的房價都無從得知，唯一有週期性更新的地產公開資料為地價，所以最後以病患所登記的住家地價作為分析的主要依據。

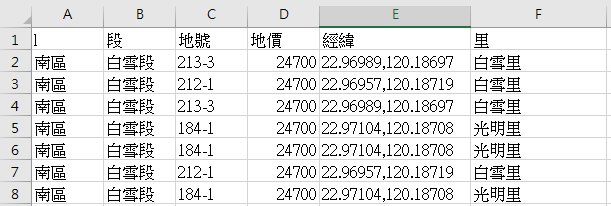
本組將每位病患所登記的住家經緯度至公開地價資料做查詢，由於有些病患的住家地址在去識別化後經緯度精度下降，導致居住在南區與東區邊界的居民有可能被歸類到東區，此部分的資料本組將直接捨棄不用

圖 3 地價資料範例

，清理過後的資料數量如上述為3463筆，圖3為地價資料範例。

1. 都市計畫使用分區資料

本組將南區的都市計畫使用分區以里為單位區分，其都市計畫分區分別有住宅區、商業區、工業區、綠地、學校、公園等等區域，並包含該分區所佔據的面積，於南區39個里當中共分為1381個使用分區，使用分區的範例資料如圖4。

圖 4 南區使用分區資料範例

1. 城市社經資料

除了直接的地價及使用分區外，本組考量了其他見間可能影響的因素如人口、人口比例、平均收入、交通易達性等社會經濟相關資料，都有可能與登革熱的傳播有相關。

人口資料由臺南市開放資料平台抓取，以里區分，其包含了每個里的總人口、男女性人口以及以年齡層分的人口比例，人口資料的範例見圖5。

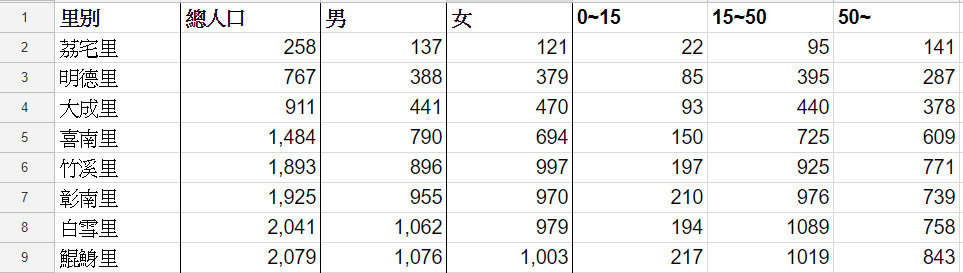


圖 5 人口資料範例

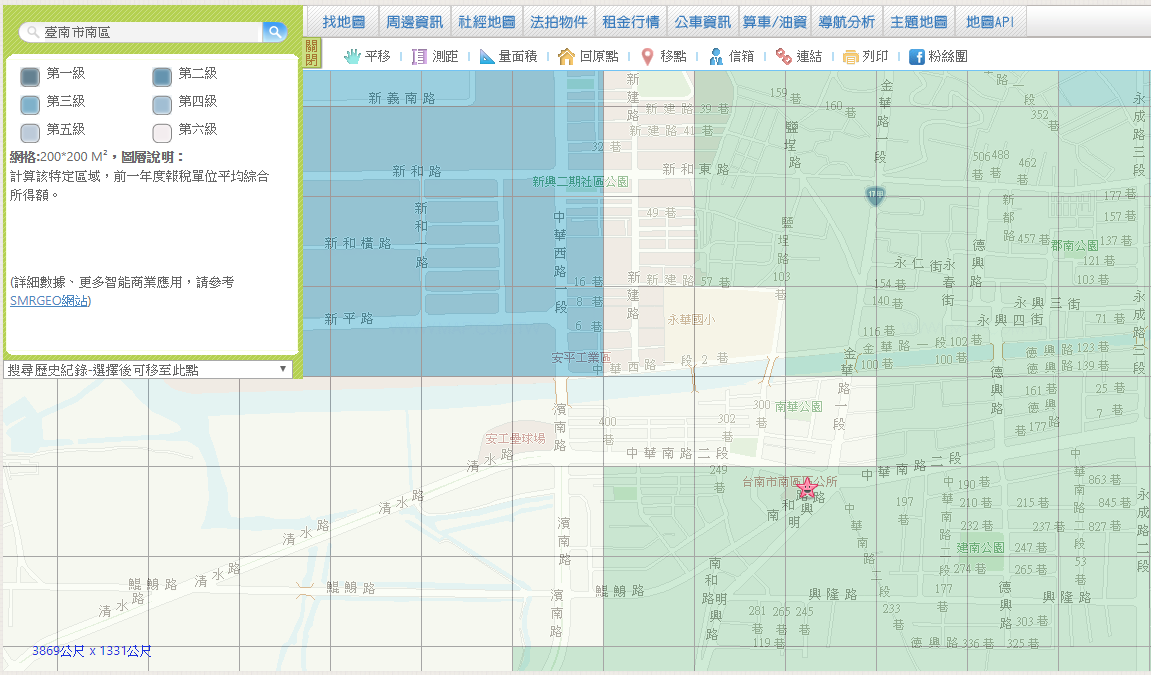
其他社經資料於台灣電子地圖服務(<http://www.map.com.tw/>)透過經緯度抓取，圖6為台灣電子地圖服務所提供的收入資料。

圖 6 台灣電子地圖服務，南區平均綜合所得分布

**得病人數與得病率**

基於每個里的總人口數不同且其分布從300至6000都有，又人口越稠密的區域疾病傳播的通常越快、得病人數也會越多，所以本組於以下的分析使用得病率來代表一個里得病程度的多寡，其計算方式如下。

**地價與得病率分析**

本組將每個里得病人口的地價取平均當作該里的地價，並以此地價對該里的得病率計算Pearson Correlation以及Spearman Correlation，其結果見表格1，可以看出不論在任一種相關係數上，里的平均地價與其得病率呈現中度相關。

|  |  |
| --- | --- |
| Pearson Correlation | Spearman Correlation |
| 0.6082 | 0.6289 |

表格 1 里平均地價、得病率相關係數

本組接著將里以人口區分成0~2千人、2~3千人、3~4千人、4~5千人以及5千以上，並將這些區間的里算出與其平均地價個別的Pearson Correlation，其結果見表格2。可以看出在0~2千人的里中，相關性為負相關，其原因為里人口數過少導致些微的得病數增加就會讓得病率失去相關性；而在3~4千人則呈現較低的相關性，其他區間則呈現高度相關。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0~2千人 | 2~3千人 | 3~4千人 | 4~5千人 | 5千以上 |
| Correlation | -0.2374 | 0.8658 | 0.5679 | 0.8659 | 0.8476 |

表格 2 里人口數分區之里平均地價、得病率相關係數

**收入與得病率分析**

由於平均收入與房地產價格間有間接關係，本組以病患所在里的平均收入對於該里的得病率計算相關係數，其結果見表格3。Spearman Correlation 呈現中度相關而Pearson Correlation呈現低度相關，表示收入的高低排序較收入的值對於得病率更有相關性。

|  |  |
| --- | --- |
| Pearson Correlation | Spearman Correlation |
| 0.3539 | 0.5798 |

表格 3 里平均收入、得病率相關係數

**里人口數與得病率分析**

通常在疾病的傳播都有人口越多則患病率提升的現象，由於病媒蚊會透過叮咬人類將病毒傳播到其他人身上，所以也有此種特性，因此本組以各里人口總數對各里得病人數作圖，試圖找出得病人數較高且異常的里

。

從圖7可以看到，紅線為回歸直線，橘線為本組給定得接受邊界，在此本只探討高於上端橘線，也就是病例數異常高的那些里，分別為新興里、國宅里、田寮里、文華里、新昌里、白雪里以及金華里，這些里的位置都位於南區的北部區域，屬於南區集中的都市區。

本組接著去分析這些里的人口年齡結構是否有特別之處，發現這些里的65歲以上人口比例都落在14%以上(表格4)，但並沒有65歲以上人口越高就越多的趨勢。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 里別 | 新興里 | 國宅里 | 田寮里 | 文華里 | 金華里 |
| 65歲以上人口比率 | 0.1620 | 0.1782 | 0.1619 | 0.1463 | 0.1485 |
| 里別 | 新昌里 | 廣州里 | 白雪里 |  |  |
| 65歲以上人口比率 | 0.1710 | 0.1823 | 0.1607 |  |  |

表格 4 病例數高於邊界之里別65歲以上人口比率表

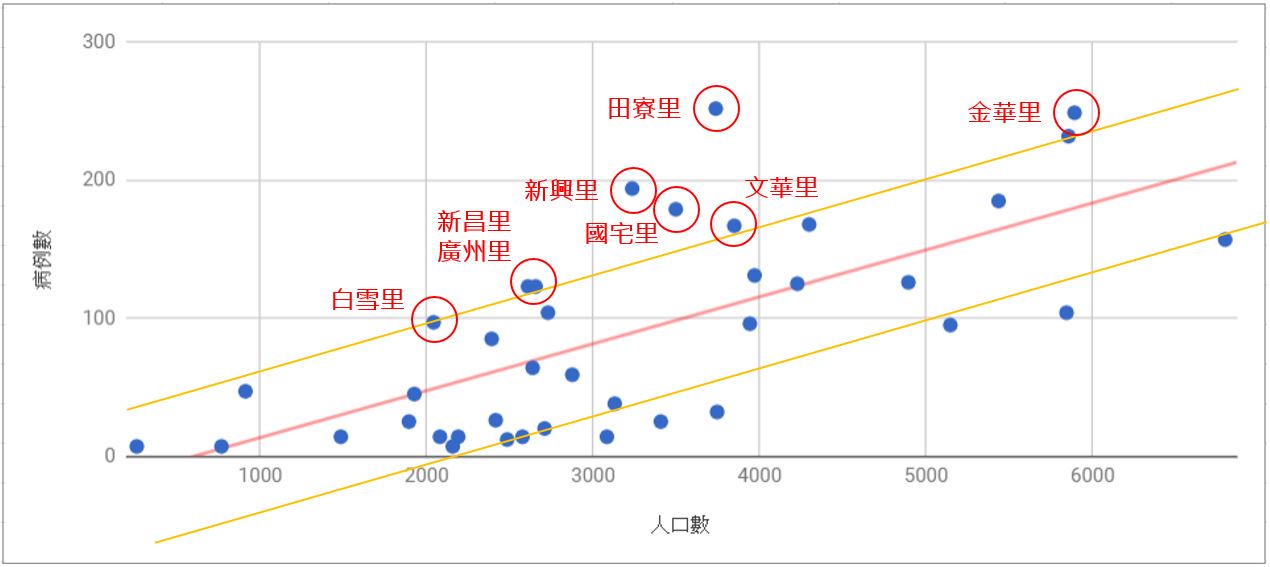


圖 7 各里人口、病例數趨勢圖

**都市使用分區與得病率分析**

南區使用分區資料共分為50種使用分區，我們將每個里中50種分區的數量對該里的得病率算Pearson Correlation，並將相關係數為正數的因子以圓餅圖呈現於圖8，而相關係數為負數的因子以圓餅圖呈現於圖9。

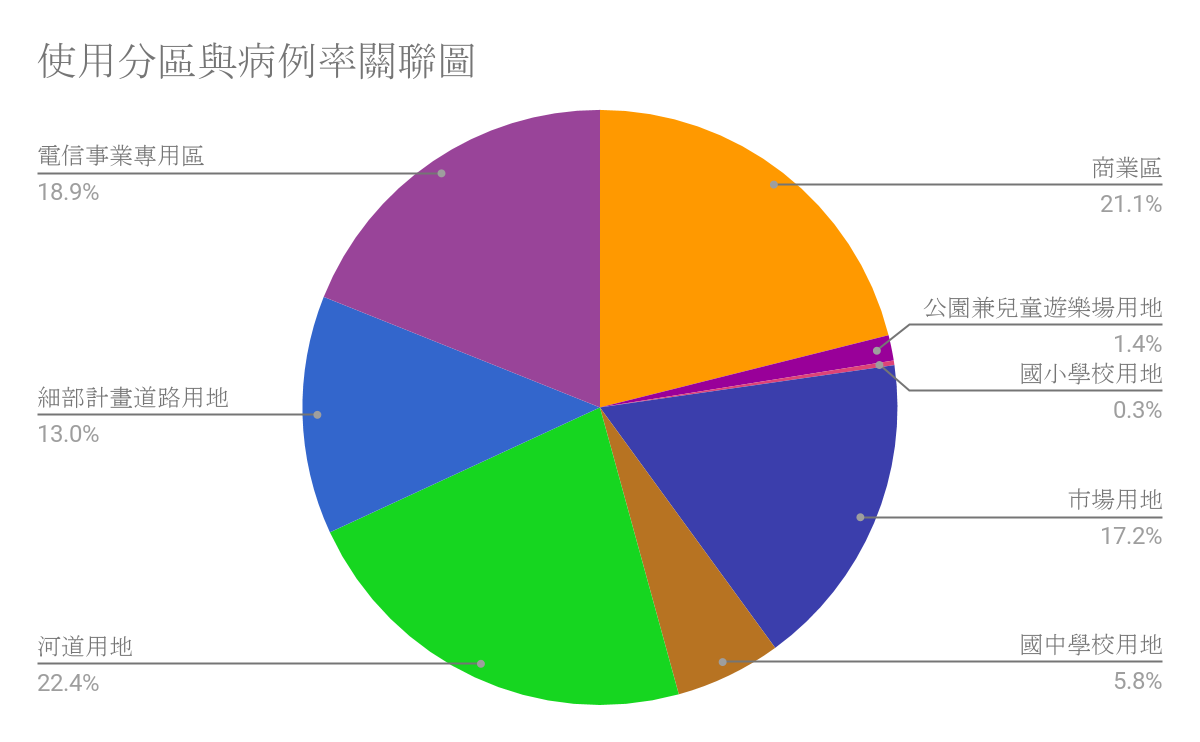


圖 8 都市使用分區正相關因子圓餅圖

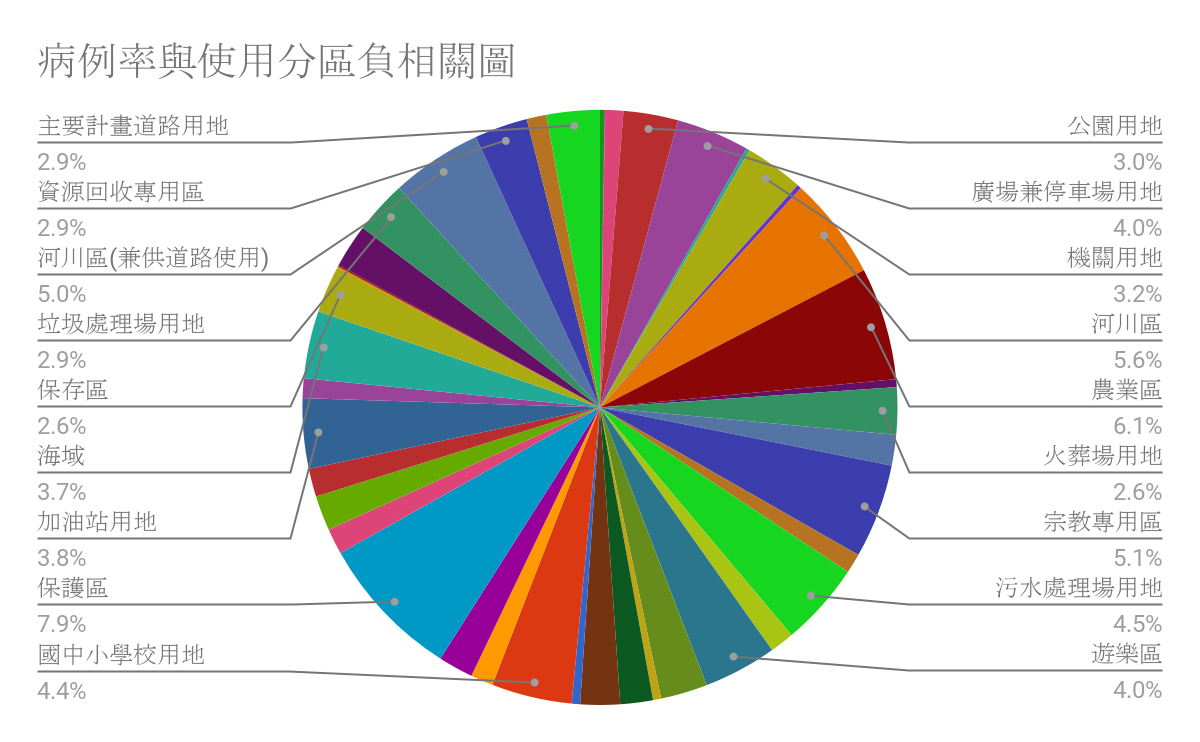


圖 9 都市使用分區負相關因子圓餅圖

為了從繁雜的數據中挑選出有用的分區，於正相關因子本組將商業用地及市場用地合併；於負相關因子本組將綠地、機場用地、廣場用地、公墓用地等合併為空曠地區。此兩種分類之相關係數見表格5，並可以看出商業用地加上市場用地的分群呈現中度正相關，而空曠地點的分群呈現中至高度負相關。

|  |  |  |
| --- | --- | --- |
|  | 商業用地+市場用地 | 空曠地點 |
| Pearson Correlation | 0.4542 | -0.5631 |

表格 5 經篩選之都市計畫分區因子相關係數

**其他社經資料與得病率分析**

為了更有效率的分析其他可能有影響的間接因素，本組將原有的地價、人口、平均所得，結合了白天人口、晚上人口、工作人口、交通網點等因素，以XGBoost分類器作預測，預測的答案為擁有該特徵的病人所在的位置是哪一個里別。

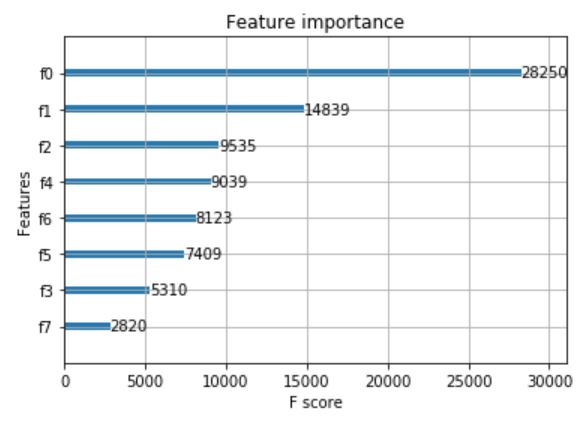
 此分類器在預測表現上達到84%準確度，代表其能有效地將某個病人依照各種因子正確的分到發生的里，本組接著將各個因子計算其重要性，結果見圖10，由前五名相對應為地價、年平均所得、人口、白天人口、交通網點數及工作人口，此結果也間接證明了前述地價與人口的重要性。

圖 10 模型特徵重要性指標

1. **成果應用**

本專案的主旨在於探討登革熱與地價、都市計畫使用分區之相關性，透過直接的地價、都市計畫分區的數量以及間接的收入、人口數等因子，本組證實了：

* 1. 當以里為單位時，該里之患病程度(得病率)與該里之平均地價呈現中至高度正相關。
  2. 都市使用分區如商業區、市場等區域與得病率呈現中度正相關，而空曠區域如綠地、公園等呈現中度負相關。
  3. 高比例65歲以上人口對於南區北部(都會區)的里別是高得病率的指標(在南區其他里別並無發現此種特徵)。
  4. 里平均收入與該里的得病率呈現中至低度正相關。

根據上述之結果，可看出本專案以證實了地價與都市計畫和登革熱之關聯，接著本組對於本專案投入實際應用提出以下建議：

1. 以整個南區而言，重點防疫於地價高的里別，尤其當該里人口高於兩千人時，更為顯著。
2. 南區位於都會區的里，有著高於異常的得病率，其原因之一與該區域中的老年人口數可能有關係，是進一步探討該區域得病率高的方向之一。
3. 重點防疫人口流動大的區域，因人口的流動對於登革熱有其影響，從商業區、市場的中高度正相關，以及XGBoost模型中工作人口、交通網點數的高重要性中可以驗證這點。
4. 空曠區域與得病率有著中至高度負相關，或許在考慮防疫時可以將某些空地排除或減少頻率。