

# 應用電信數據於新冠肺炎疫情熱區預測—以臺北市、新北市為例

周敬棋<sup>1</sup>，陳正誠<sup>2</sup>，陳潤秋<sup>3,4</sup>，詹大千<sup>5,6\*</sup>

## 摘要

本研究運用電信信令數據，進行傳染病熱區預警的建構，針對臺北市、新北市共 1488 個村里在 2021 年 5 月間的新冠肺炎(COVID-19)疫情進行實證分析，分別以雙北市村里之靜態人口統計資料與電信數據資料建構村里人口流動網絡，以網絡空間中的結構等位特性估計人流足跡相似度，結合確診者的空間分布，預測未來各村里的確診者數量是否高於風險門檻值，並以 ROC 曲線(Receiver Operating Characteristic Curve)觀測衡量預測效果，同時利用 Logistic regression、Poisson regression 建構基於村里間足跡相似度的感染區預測模型。

分析結果顯示，基於結構等位所估計之村里足跡相似度對 COVID-19 高風險區預測有幫助，且電信信令數據之平均預測水準較靜態人口資料佳，說明電信信令數據可提升動態人流預測的應用價值，協助決策單位在醫療資源有限的前提下，將防疫量能分配至最高感染風險區域，進而達到降低傳染風險與減災的效果。

**關鍵字：**電信數據、社會網絡分析、結構等位、傳染病預測、COVID-19

## 壹、緣起

在面對傳染病疫情時，由於資源有限，如何將資源分配於未來高風險感染區至關重要。在政策實務上，常以行政區的人口數、累計確診數，或是與感染區的空間鄰近性，劃列高風險感染區以加強防疫量能的投入，如新北市在 2021 年 6 月初，劃列 199 個重熱里作為防疫工作的重點區域。呼吸道傳染病的傳播，人與人間的接觸是必要的條件，有別於耗時耗力的疫情調查，針對的是防止點感染源的擴散，若當大規模的社區感染時，難以針對所有的感染個案或疑似個案進行大規模疫調，此時若能以人流資料預測傳染病的走勢，將可以有助於防疫資源的部署。

人流數據又可分為靜態人口統計資料、普查資料與電信信令數據資料三種。靜態人口統計資料會基於一些人口吸引的假設，進行區域間人流關係的建構，使用的模型如重

<sup>1</sup> 國立臺灣大學地理環境資源學系

<sup>2</sup> 臺北市政府衛生局

<sup>3</sup> 新北市政府衛生局

<sup>4</sup> 國立陽明交通大學生物醫學影像暨放射科學系

<sup>5</sup> 中央研究院人文社會科學研究中心

<sup>6</sup> 國立陽明交通大學公共衛生研究所

\* 通訊作者 tachien@gate.sinica.edu.tw

力模型(gravity model) (Barthélemy, 2011)或是輻射模型(radiation model) (Simini et al., 2012)；普查資料則是以大規模的面訪調查民眾的居住地與上班、上學的行政區建構不同地區之間日間與夜間流動的關係(Barbosa et al., 2018)；電信信令數據可根據電信信令訊號，定義人群在地區之間的停留時間與旅次關係。隨著科技革新與計算效能的進步，電信信令數據的應用日漸蓬勃，國外也有學者將其用以輔助傳染病的預測(Barbosa et al., 2018)，因此本研究將著重探討人流資料之特性與應用於預測新冠肺炎高風險區在雙北市的實例。

## 貳、課題

本研究使用遠傳電信信令數據資料與靜態人口資料模擬人流，並且探討其中的差異與可能的成因，透過建構村里尺度的人口動態網絡，以社會網絡分析法建立模型，旨在尋得各時期的村里相對風險，並比較不同統計模型的優劣，以作為熱區劃列與防疫資源分配的參考依據。

## 參、研究方法及步驟

### 一、研究資料

#### (一) 電信信令數據資料

電信信令數據資料作為建構人口移動網絡使用，本研究以2021年5月10日、5月16日與5月20日之遠傳電信信令數據資料估計人流，該公司的市佔率約為24.6%。電信使用者所在的信令訊號網格會被紀錄，網格內面積佔比最大之村里即為電信使用者所在之區域，並將用戶過去一個月夜間22:00至03:00停留時間總和最長之村里認定為用戶居住村里；用戶在各村里的停留時間若超過十五分鐘將會被紀錄。電信使用者依年齡分組為18-21歲、22-29歲、30-59歲與60歲以上四個族群進行加總計算。此外，電信業者為保護個人隱私，村里的各年齡層居住人口數少於五人時將無法取得該村里之居住人口數資訊，故電信人口過少之數據將不納入分析。

#### (二) 靜態人口資料

為了衡量電信信令數據的預測力，本研究亦使用靜態人口資料，透過重力模型模擬人流作為參考基準，所使用的資料為內政部2021年2月所統計之各村里分齡戶籍人口數，並使用與電信數據相同的年齡分組。

#### (三) 村里確診數資料

使用臺北市與新北市衛生局統計之各村里在2021年5月8日至5月27日之逐日確診數資料，因涉及確診者隱私，村里確診數未以年齡分組。

### 二、研究方法

#### (一) 建構地理空間網絡

使用電信信令數據資料，可透過居住村里與停留村里的關係，以村里作為節點(nodes)，連結(edges)的方向為居住村里(起點，source)指向停留村里(終點，Target)，

以居住村里人口在其他村里的停留時間比例作為衡量人流大小的連結權重（Weight），建構以年齡分組的雙北市人流有向網絡。

除了電信信令數據外，亦使用靜態人口資料，以重力模型模擬人流量的多寡，作為與電信信令數據比較的基準。在重力模型方面假定村里間之人流多寡與兩村里之人口數乘積成正比、兩村里之距離平方值成反比：

$$P_{ij}^{Gravity} = \frac{POP_i \times POP_j}{d_{ij}^2} \quad Normalized P_{ij}^{Gravity} = \frac{P_{ij}^{Gravity}}{\sum_{j \in J} P_{ij}^{Gravity}}$$

其中  $P_{ij}^{Gravity}$  表示重力模型間的停留時間比例，被視為村里 i 之居民停留於村里 j 之機率； $POP_i$ 、 $POP_j$  分別表示村里 i、村里 j 之人口； $d_{ij}$  表示兩村里地理中心點之直線距離，最後將該數值進行正規化處理，使每個起點村里（source）在所有村里的停留時間比例之和為 1。

## （二）網絡中心性與確診數熱區分析

在傳染病數理模型中，接觸行為是造成疾病傳染的主要因素，我們可以透過計算村里間的往返停留關係描繪傳染病模型中的接觸過程。在人流流入量較大的地區，如商業區、重要交通轉乘站等所在村里，成為了雙北市足跡最為密集之處。為了衡量地區作為終點村里所吸納的人流量，本研究計算各村里的加權度中心性（weighted in-degree）觀測空間中的人流熱區，並透過 Bivariate Local Morans' I 檢視人流所集中之熱區與確診數的熱區是否呈現正相關。

## （三）結構等位、村里足跡相似度與相對風險計算

計算加權度中心性可得知雙北市中較為危險、人流較密集或來源複雜的區域。然而，中心性指標對於預測確診數不一定會有良好的準確度與參考價值，因確診數是以確診者的居住地作為統計，而非感染地，在分配防疫資源與篩檢量能時也應當參考確診者的居住地，因此，本研究以節點在地理空間網絡中的位置（network position）估計確診者數量高低的相對風險。

若兩村里擁有類似的人流流向，則可推測兩村里的居民有較高的機率互相接觸。換而言之，地理空間網絡中的兩節點共享相同的終點，且停留時間比例的分配也大致相同，當一個節點成為了感染區時，與其足跡相似的另一個節點也會暴露於較高的風險之中。

以結構等位（Structural Equivalence）的方式即可計算兩村里之足跡相似度：

$$Equivalence_{ij} = \sum_{k \in K} |P_{ik} - P_{jk}| \quad Similarity_{ij} = 1 - \frac{Equivalence_{ij}}{\max(Equivalence_{ij})}$$

其中  $Equivalence_{ij}$  表示村里  $i$  與村里  $j$  之結構等位性，由此兩村里與所有村里的停留時間比例差距總和計算而得， $K$  表示村里的集合。而  $Similarity_{ij}$  則是將  $Equivalence_{ij}$  進行標準化的處理，將結構等位特性轉換為相對的足跡相似度，數值介於 0 與 1 之間， $Similarity$  為 1 表示兩村里之足跡相似度是所有村里中最高的；0 則代表兩村里的足跡無任何重疊現象。

若進一步結合逐日確診數資料，可利用足跡相似度推估村里未來成為感染區的風險。假定相似度為 1 之村里代表兩村里足跡高度重疊，因此當其中一處受感染時，另一個村里將承受相同的風險；因此若  $a$ 、 $b$  兩村里之相似度為  $S_{ab}$ ，且  $a$  處新增  $C$  例確診個案( $C_a$ )，則另一個村里將承  $C_a \times S_{ab}$  之風險。

$$Risk_j = \sum_{i \in I} C_i^{Xdays} \times Similarity_{ij}$$

其中， $I$  表示村里之集合、 $C_i^{3days}$  表示近  $X$  日之村里  $i$  之累積確診者數量、

$Similarity_{ij}$  表示村里  $i$  與  $j$  之足跡相似度，以此方式計算雙北市每個村里之相對風險高低，即可針對雙北市村里進行風險估計與確診個案數預測。

由於新冠肺炎的潛伏期為 5 日、受感染後至可傳染期需 3 日、可傳染期為 7 日，無症狀之可傳染期 2 日與出現症狀之第 1 日有較高機率外出傳播病毒，在出現症狀後即因為隔離等因素停止外出，因此本研究在估計風險時以  $X=3$  進行計算，並預測未來 7 日各村里之確診者數量的相對風險。舉例而言，針對 5 月 11 日至 17 日之村里相對風險，我們使用 5 月 10 日之人流足跡相似度與 8 日、9 日、10 日之村里確診數資料進行風險估計。

#### (四) 預測與迴歸分析

考量到防疫資源有限，在分配時大都以地區疫情嚴重與否作為優先順序上的參考，本研究計算相對風險，以決定未來 7 日是否需被劃列為高風險感染區。若假設欲劃列的高風險感染區數量為  $N$ ，則依據風險選出較高的  $N$  個村里劃列為高風險感染區，並觀測所有村里是否達到確診數門檻值  $T$ 。若一村里被劃列為高風險感染區，且未來 7 日村里累積確診數大於  $T$  值，則分類正確且為真陽性 (TP)，小於  $T$  值則為分類錯誤且為偽陽性 (FP)；若一村里未被劃列為高風險感染區，且未來 7 日村里累積確診數小於  $T$  值，則分類正確且為真陰性 (TN)，大於  $T$  值則為分類錯誤且為偽陽性 (FN)。

在給定  $T$  值代入不同  $N$  值作為防疫量能的閾值，以 ROC 曲線與曲線下面積 (AUC) 比較下列三種不同的風險預測方式：以過去三日各村里累積確診數多寡計算風險、以重力模型模擬之人流所計算之風險、以電信信令數據所計算之風險。

迴歸分析部分則使用 Logistic regression 與 Poisson regression 計算不同模型的勝算比 (odds ratio)，其中 Logistic regression 評估特定確診門檻值 ( $T$ ) 之狀況下，是否高

於 T 值 (y) 與風險 (x) 的關係；Poisson regression 則評估確診數量 (y) 與風險 (x) 之關係。

## 肆、主要結果

### (一) 建構地理空間網絡

分別以重力模型模擬與電信信令數據之人流建構以年齡分組方式呈現地理空間網絡如圖 1 所示。根據連結的人流量大小與歐幾里得距離 (Euclidean distance, 以下以距離簡稱之)，可觀察到網絡特性有所不同，在電信信令數據之人流網絡中，距離較長之連結也有可能擁有較高之停留時間比例；而重力模型模擬之資料則是呈現對數關係穩定隨距離遞減的趨勢。

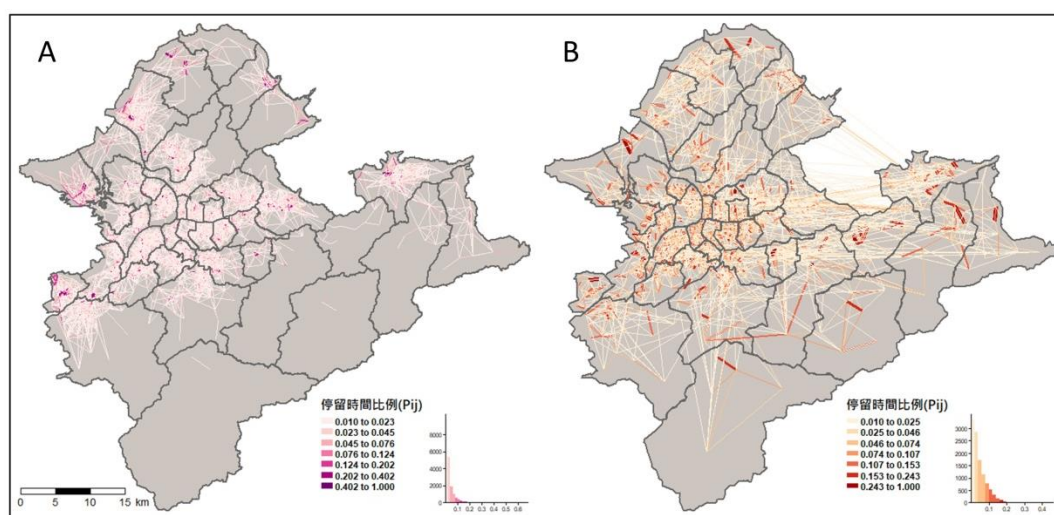


圖 1 2021 年 5 月 10 日不同資料所建構之 30-59 歲人流網絡

A.重力模型模擬之人流；B.電信信令數據人流資料

若使用電信信令數據觀察年齡的組間差異，也可發現較為年輕的族群流動力較佳，連結之地理距離較遠；而老年人口之流動力較弱，隨距離的遞減率也較大，符合一般大眾對於人口流動的認知；然而，重力模型所得之結果在年齡間並沒有顯著的差異。

### (二) 網絡中心性與確診數的熱區分析

將兩種不同資料來源的網絡分別進行加權度中心性分析後，可發現電信信令數據之人流集中處主要位於雙北市之市中心，或是位於都會中交通網絡較為密集的地區，然而重力模型模擬之人流網絡中，雙北市住宅區村里通常擁有較高的加權度中心性，位於市中心的村里數值反而較低，因為在使用重力模型模擬人流時並未考慮土地使用與交通網絡之型態，因此所得之人流中心大都與現實之中所認知之熱鬧區域不同。

以雙變數局部空間自相關分析 (Bivariate Local Morans'I, BiLISA)，觀察加權度中心



性與確診數之變化(圖 2)，結果顯示確診數熱區與人流中心性熱區並無明顯的相關性，在有大量人流湧入的地區，未必會在未來有較高的確診數，因此在進行高風險感染區的劃列時，人流流入量不一定會是好的指標。

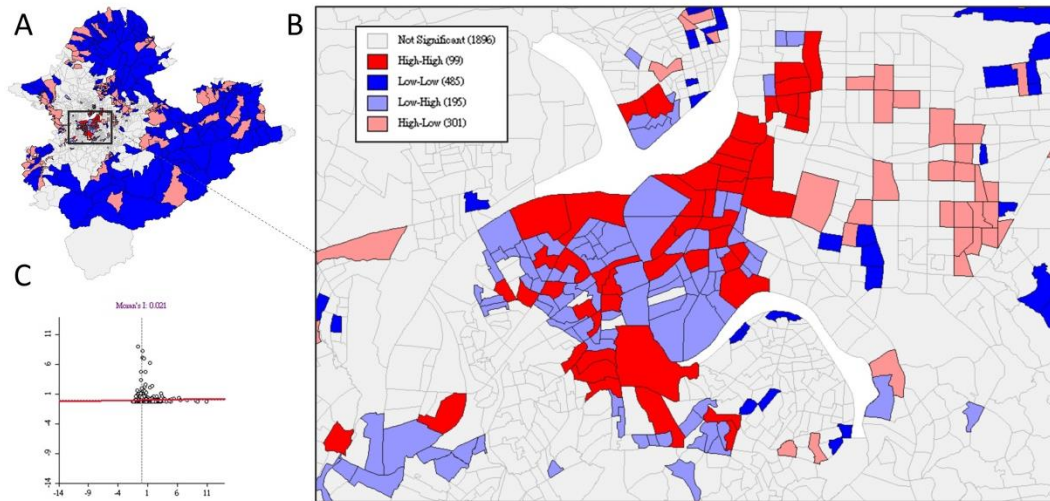


圖 2 2021 年 5 月 10 日電信信令數據之人流度中心性與累積確診數(5 月 11 日至 5 月 17 日)之雙變數局部空間自相關分析

A. 研究區全結果；B. 萬華、板橋地區之放大圖；C. Moran's I 散佈圖

### (三) 結構等位、村里足跡相似度與相對風險計算

根據村里足跡相似度與村里間距離之關係(圖 3)，可觀察到足跡相似度擁有高度的空間聚集現象。電信信令數據而言，距離較為相近的兩村里，也有可能出現較低的足跡相似度，必須依居民特性、兩村里在產業、交通上的依賴度而定；然而重力模型所得之結果卻說明距離較近的兩村里擁有較高之足跡相似度，凸顯現了靜態資料的缺點。

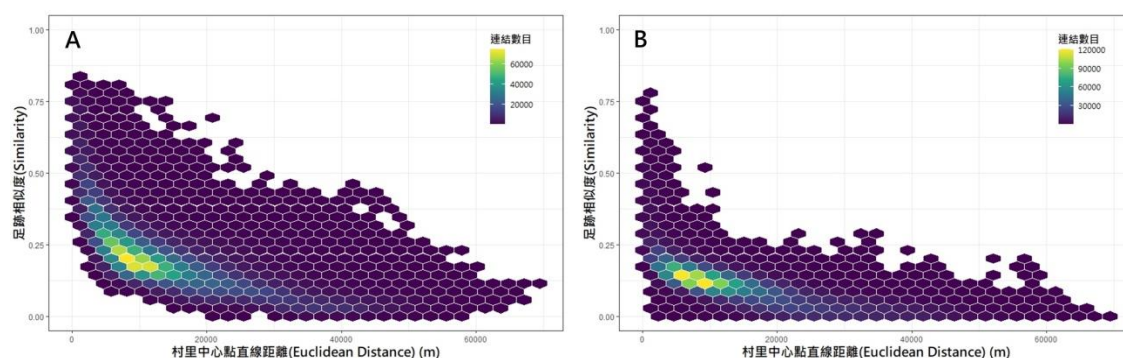


圖 3 村里連結距離與足跡相似度關係圖

A. 重力模型模擬所得；B. 電信信令數據人流資料

### (四) 預測與迴歸分析

基於三個不同的相對風險與不同確診數門檻  $T=1$ 、 $T=5$ 、 $T=10$ ，以電信信令數據、重力模型等兩種不同模型劃列高風險感染區之 ROC 曲線，部分結果如圖 4 所示。三種  $T$

值的情況下，電信信令數據資料的分類最佳，其次為重力模型。在不同年齡層方面，以 60 歲以上族群進行預測有較高的 AUC，推斷是因該波流行中確診人口以高齡人口為主所致。

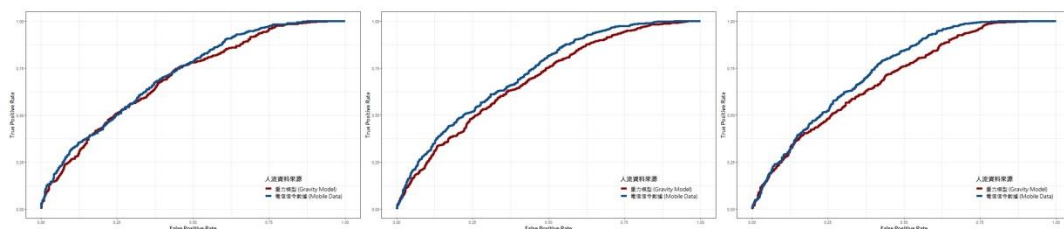


圖 5 2021 年 5 月 10 日、16 日、20 日兩資料類型之 ROC 曲線比較  
(皆使用 30-59 歲組之資料， $T = 1$ )

迴歸分析方面，無論是 Logistic regression 或是 Poisson regression 的結果皆顯示電信數據資料擁有較高的勝算比，說明了本研究中以電信信令數據估計風險的方法，能夠有效反映對未來 7 日各村里的確診狀況，且數值相較其他傳統方式來得敏感。

## 肆、結論

電信信令數據資料具備靜態人口資料難以捕捉的特性，如距離對於移動行為的影響等。在建構地理空間網絡後，除了可以使用國內外文獻中常見的社會網絡分析、空間分析等方法分析人流資料，亦可探討節點在網絡的位置與結構等位性，引入傳染病模型中「接觸」的概念，估計區域與區域間的足跡相似度、了解區域間的風險依存關係，並結合逐日病例數資料計算風險，在預測未來感染區上的表現能夠較傳統的方法來得佳。

面對未來的傳染病時，如何在資源有限的情況下將醫療量能分配至最易受感染之區域，是至關重要的課題。隨著科技日新月異的發展，資料的蒐集與取得也更加便捷，可多加利用電信信令數據資料進行分析，如此可更有效率地分配資源，同時將災害最小化，進而達到減低傳染病災害、增進社會福祉的效果。

## 伍、致謝

感謝台灣大學流行病學與預防醫學研究所林先和教授提供指導。也感謝衛生福利部疾病管制署「防疫新生活健康行為監測研究計畫」提供研究經費購買去識別化電信人流數據。感謝遠傳電信公司在電信數據處理上的技術諮詢服務。

## 陸、參考文獻

1. Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., & Tomasini, M. (2018). Human mobility: Models and applications. *Physics reports*, 734, 1-74.
2. Barthélemy, M. (2011). Spatial networks. *Physics reports*, 499(1-3), 1-101.
3. Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American journal of Sociology*, 92(6), 1287-1335.
4. He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., &

- Tan, X. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature medicine*, 26(5), 672-675.
5. Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9), 577-582.
  6. Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1), 49-80.
  7. Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392), 96-100.
  8. Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., & Rothe, C. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature*, 581(7809), 465-469.
  9. Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K., & Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*, 112(38), 11887-11892.
  10. Zipf, G. K. (1946). The  $P^1 P^2/D$  hypothesis: on the intercity movement of persons. *American sociological review*, 11(6), 677-686.