

地理加權迴歸在視覺化分析之探討

余清祥^{1†} 梁穎誼² 郭柔芸¹

¹ 國立政治大學統計學系

² 逢甲大學風險管理與保險學系

摘 要

近年大數據蓬勃發展，統計分析的應用更為廣泛，各領域資料以不同型態出現，資料視覺化（Data Visualization）成為探索性資料分析（Exploratory Data Analysis）的核心。視覺化對於空間資料尤為重要，藉由圖表等工具可有效地呈現資料主要特性，包括空間異質性（Spatial Inhomogeneity）、空間自相關（Spatial Autocorrelation），做為後續研究進行的依據。地理加權迴歸（GWR，Geographically Weighted Regression）可視為空間資料的迴歸分析，描述目標變數與解釋變數間的局部關係，用於展示變數關係隨地理位置的變化。本文探討地理加權迴歸的適用時機，透過電腦模擬說明 GWR 的限制及可能問題，測試修正方法是否有效，同時提出這個方法的使用建議。

關鍵詞：地理加權迴歸、空間統計、視覺化、探索性資料分析、電腦模擬。

JEL classification: C21, C61.

[†]通訊作者: 余清祥

E-mail: csyue@nccu.edu.tw

1. 前言

IBM (International Business Machines Corporation) 在 2010 年提出大數據 (Big Data) 這個概念，促成了統計方法及機器學習的發展，將資料分析廣泛應用於各專業領域。過去資料分析通常限於有確切定義及欄位的結構資料 (Structured Data)，由於紀錄及儲存技術的演進，對於無特定模式的非結構資料 (Unstructured Data) 之分析需求愈發殷切，包括如何選取合適的資料結構化 (Structurization) 方法。除了資料類型的多元化，如何以視覺化 (Visualization) 呈現資料特性也成為近年統計分析的重要訴求，由北美洲規模最大的聯合統計年會 (Joint Statistical Meetings, JSM)，在 2019 年將考普斯獎 (The Committee of Presidents of Statistical Societies Awards, COPSS) 頒給 Hadley Wickham 博士可見一斑。COPSS 堪稱統計學界的最高榮譽，Wickham 博士因為在資料視覺化的貢獻而獲獎¹，包括 R 軟體的 ggplot2 模組，他現在是 Rstudio 的首席統計學家。

視覺化分析工具的選擇取決於問題定義、資料屬性等因素，其中空間 (Spatial) 類型的資料尤為棘手，部分原因與人類眼睛較易於辨識二維圖形有關，使得三維以上的圖像工具並不普及。在此以臺灣鄉鎮市區 (Township) 層級的資料為例，說明視覺化空間資料時遭遇的挑戰。如果目標在於探索各鄉鎮市區哪些特性 (即解釋變數) 與該地區特定死因死亡率 (即目標變數) 有關，像是大腸癌、慢性阻塞性肺病 (Chronic Obstructive Pulmonary Disease, COPD) 等對國人健康造成重大威脅的死因，若能找到相關風險因子可有效提供疾病預防及治療的參考 (李宗儒等人, 2021; Chan *et al.*, 2014)。上述資料分析可歸類為迴歸分析，檢視解釋變數與目標變數是否相關，以及解釋變數間是否存在共線性，如果變數均為空間資料，變數間的關係將為三維以上的圖形，不容易透過一般散佈圖 (Scatterplot) 之類圖形目視判斷變數間的關係。

由於不易呈現空間資料中變數間的關係，多數研究僅呈現單一變數的空間特性，如圖 1-1 為 1999-2007 年臺灣鄉鎮市區層級的男性慢性阻塞性肺病 (標準化) 死亡率，資料明顯屬於空間異質性 (Spatial Inhomogeneity)，亦即可能存在死亡率較高的熱區 (Hot spot)，或是各鄉鎮市區死亡率存在空間相關 (Spatial Autocorrelation)。圖 1-1 的視覺化訊息，可視為空間資料的探索性資料分析 (Exploratory Data Analysis, EDA) 的一種，有助於挖掘出變數的重要特性。然而，空間資料的雙 (及多) 變數分

¹Wickham was awarded the international COPSS Presidents' Award in 2019 for "influential work in statistical computing, visualisation, graphics, and data analysis" including "making statistical thinking and computing accessible to a large audience". 來源: https://en.wikipedia.org/wiki/Hadley_Wickham

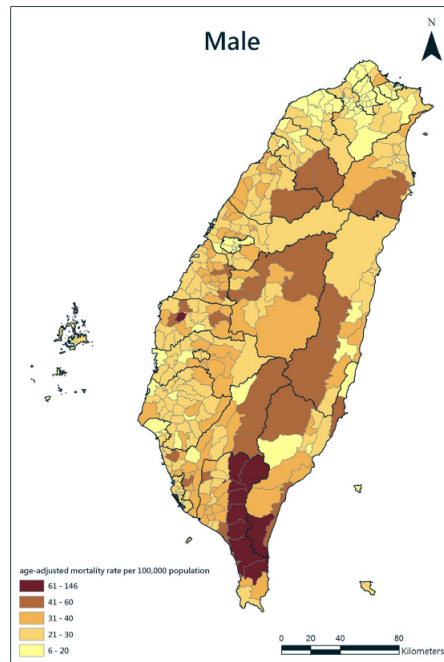


圖 1-1: 1999-2007 年臺灣男性慢性阻塞性肺病死亡率。
來源：Figure 2 ([Chan et al., 2014](#))

析較少以視覺化方式呈現，大多透過空間相關係數（Spatial Correlation），或是空間版本的迴歸分析描述兩者的關聯。

地理加權迴歸（GWR, Geographically Weighted Regression）是常見的空間迴歸模型之一（[Fotheringham et al., 2002](#)），用於描述某個目標變數與解釋變數間的關係，可視為常見迴歸分析的空間版本。常見的迴歸分析多半會透過圖形檢視模型是否合宜，像是殘差（Residual）與解釋變數的散佈圖，可用於檢視誤差是否具有固定變異數、判斷模型缺適性（Lack-of-fit）等的參考。但地理加權迴歸的相關研究大多僅呈現解釋變數是否顯著、判定係數（ R^2 ）之類的訊息，或是類似圖 1-1 顯示各解釋變數的空間特性，很少提及解釋變數、目標變數的空間關係，包括如何藉由視覺化工具檢視迴歸模型合宜性，是否需要進一步調整模型（[Edayu and Syerrina, 2018](#); [Matthews and Yang, 2012](#)）。

我們認為空間分析應加入視覺化判斷，協助指引適當的分析方向及工具，以建構合乎需求的模型，因此本文以地理加權模型為研究對象，以圖像顯示此方法的估計值檢查模型的適確性。本文將使用電腦模擬，除了檢視目標變數的估計結果，像是迴歸參數曲面的視覺化估計值，同時也計算變數參數估計值的誤差，作為評估地理加權模

型的依據，包括是否可將估計結果用於資料視覺化。另外，根據我們先前的研究經驗（[Leong and Yue, 2017](#)），地理加權模型的目標估計值有不夠平滑之虞，我們也引進條件地理加權迴歸模型（Conditional GWR, CGWR），調整原始 GWR 估計值的震盪，使其更合適用於圖像化呈現。以下將先介紹地理加權迴歸的原理，之後在第三節比較 GWR 和 CGWR 的估計結果，第四節評估兩種模型用於實證分析的差異，第五節再提供 GWR 的使用建議與感想。

2. 地理加權迴歸

本節先介紹地理加權迴歸，包括想法、模型、估計方式等。地理加權迴歸（以下簡稱 GWR）由 [Brunsdon 等人 \(1996\)](#) 提出，目的在於描述目標變數與解釋變數間的關係，可視為一般迴歸分析的空間推廣，但 GWR 允許模型參數隨地點改變，能夠用於描述區域特性，因此估計方法較為複雜。GWR 的模型可由下式表示：

$$Y_i(u_i, v_i) = \beta_0(u_i, v_i) + \sum_{j=1}^k \beta_j(u_i, v_i) X_j(u_i, v_i) + \varepsilon_i(u_i, v_i) \quad (1)$$

其中 $X_j(u_i, v_i)$ 、 $\beta_j(u_i, v_i)$ 分別為地區 (u_i, v_i) 第 j 個解釋變數及其參數值， $\varepsilon_i(u_i, v_i)$ 為誤差項 ($i = 1, 2, \dots, n$; $j = 0, 1, \dots, k$)。誤差項一般上假設為常態分配，即 $\varepsilon_i(u_i, v_i) \sim N(0, \sigma^2)$ 。然而這種區域資料 (Areal data) 的估計與一般迴歸不同，因為各地只有一個觀察值、參數個數卻不只一個，衍生出自由度不足的問題。因此 GWR 運用地理學第一定律 (Tobler's first law of geography)：「所有事物都與其他事物相關，但是近處的事物比遠處的事物更相關」([Tobler, 1970](#))，亦即考量鄰近地區資料，再給予不同權重作為參數估計的依據。通常距離愈遠、權重愈小，這也與地理模型常見的距離遞減效應 (Distance decay) 相似。GWR 的參數估計可由下式獲得：

$$\hat{\beta}(u_i, v_i) = (\mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(u_i, v_i) \mathbf{Y} \quad (2)$$

其中 $\hat{\beta}(u_i, v_i) = (\hat{\beta}_0, \dots, \hat{\beta}_k)^T$ 為長度為 k 的參數向量之估計值，而權重矩陣 $\mathbf{W}(u_i, v_i)$ 通常是對角矩陣：

$$\mathbf{W}(u_i, v_i) = \begin{pmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{pmatrix} \quad (3)$$

w_{i,i^*} 表示其他地區 i^* 於第 i 地區的權重。 w_{i,i^*} 可透過核函數 (Kernel functions) 估算，一般上常見的核函數為高斯距離遞減函數 (Gaussian distance-decay)，可表示為：

$$w_{i,i^*} = \exp \left\{ -\frac{1}{2} \left(\frac{d(i,i^*)}{h} \right)^2 \right\} \quad (\text{高斯距離遞減函數}) \quad (4)$$

上式中 $d(i,i^*)$ 可視為 i 與 i^* 兩地區之間的距離，其定義可為歐氏直線距離或其他定義（如行車距離），但符合上述地理學第一定律的假設，也需符合地理區域間的實際需求。例如：臺灣本島的東西方向的幅原較短，但由於中央山脈的阻隔，必須考慮交通旅行等實務限制，不應直接套入直線距離。

套用核函數估計參數時，有兩個因素影響估計結果：核函數的選擇、核函數的環寬 (bandwidth)，也就是 (4) 式的 h ，其目的在於調整權重的距離衰退速度。相對於核函數的選擇，環寬對估計值的影響較為明顯，讀者可參考估計密度函數 (Density estimation) 的討論 (Silverman, 1985)。環寬與參數估計值的平滑度 (smoothness) 相關，若環寬趨近於無窮大，則每個地區的權重相等，其估計值與簡單線性迴歸一樣；若環寬較小，代表只以鄰近觀察值進行估計。換句話說，環寬的選擇和資料使用個數有關，必須在整體趨勢或局部特色間取捨，由於 GWR 為偏向於局部模型，通常環寬也不會選擇太大的數值。

GWR 是以資料驅動 (data-driven) 的思維設定環寬，先選擇某個目標函數，像是估計誤差之類的適合度指標 (goodness of fit)，透過最小化目標函數找出最佳環寬值。一般迴歸也是採取相似作法，通常以最小平方誤差找出參數估計值，GWR 多半也是以平方誤差為標的，藉由最小平方交叉驗證法 (least square cross-validation) 求取模型參數 (Brunsdon et al., 1998)。GWR 使用交叉驗證的原因在於避免過度配適 (over-fitting)，如果依照最小平方方法通常代表使用鄰近少數觀察值估計參數，亦即環寬相對比較小，如此得出的估計值大抵與觀察值接近（以及有較大判定係數 R^2 ），多半只能呈現出局部特色。反觀交叉驗證則不同，其想法類似加入了「預測」的視角，避免整體模型只能適用於某些地區，以致產生過度配適。最小平方交叉驗證法公式如下：

$$CVSS(h) = \sum_i \{y_i - \check{y}_i(h)\}^2 \quad (5)$$

其中 $\check{y}_i(h)$ 為去除掉第 i 樣本後，在環寬設定為 h 之下估計值。此外，GWR 有時會加入其他判斷條件，仿造迴歸等分析定義 AIC (Akaike Information Criterion) 值作為環寬選擇的輔助標準 (Fotheringham et al., 2002)。

另一方面，GWR 也可以透過廣義相加模型 (Generalized Additive Model, GAM) 表示 (Fotheringham *et al.*, 2017; Leong and Yue, 2017; Yu *et al.*, 2020)：

$$\mathbf{y} = \sum_{j=1}^k \mathbf{f}_j + \boldsymbol{\varepsilon} \quad (6)$$

廣義相加模型中的目標變數 $\mathbf{y}$ 、模型函數 $\mathbf{f}_j$ 、誤差 $\boldsymbol{\varepsilon}$ 皆為向量： $\mathbf{y} = (Y(u_1, v_1), \dots, Y(u_n, v_n))^T$ ； $\mathbf{f}_j$ 為個別自變數 j 所對應的平滑函數，其形式可表為 $\mathbf{f}_j = (f_j(\beta_j(u_1, v_1), X_j(u_1, v_1)), \dots, f_j(\beta_j(u_n, v_n), X_j(u_n, v_n)))^T$ ，也就是說目標變數可由數個平滑函數相加而成，使用者可根據需求選擇較為合宜的平滑函數；誤差 $\boldsymbol{\varepsilon} = (\varepsilon(u_1, v_1), \dots, \varepsilon(u_n, v_n))^T$ 。在地理加權迴歸的模型下， $\mathbf{f}_j$ 為地理加權的迴歸估計式，本文模擬及實證分析採用單變量在 (2) 式的函數。事實上，地理迴歸的估計式除了採用本文介紹的核函數以外，也有採用了平方呈反比的距離遞減函數 (Bi-square)。從文獻的結果與筆者經驗歸納，核函數的選擇所分析的結果差異不會太大。相對而言，環寬的選取影響較大，這也是 GWR 模型在使用前，都建議經過前述交叉驗證步驟以選擇較好的環寬。近年有不少 GWR 衍生的修正模型以廣義可加模型為基礎，包括 Multiscale-GWR (MGWR) (Fotheringham *et al.*, 2017)、Conditional-GWR (CGWR) (Leong and Yue, 2017)，透過調整各個 $\mathbf{f}_j$ 的環寬（平滑程度）達到更好的估計效果。GAM 的參數估計一般採用疊代方式估計平滑函數 $\mathbf{f}_j$ ，像是 MGWR 與 CGWR 都採用類似作法，不過 MGWR 使用 Back-fitting 估計，而 CGWR 則是以 Jacobi 疊代求解。兩種估計方法的差異在於更新疊代的速度，Jacobi 式的疊代法使用上一回的估計值繼續估算，Back-fitting 在疊代過程中即刻更新估計值。本文將使用 Jacobi 法為基礎的 CGWR，與傳統 GWR 比較用於資料視覺化之優劣。CGWR 的估計步驟如下 (Leong and Yue, 2017)：

第一步：設定所有的 $\mathbf{f}_j$ 的初始值。常見的設定為 0，或表示為 $\mathbf{f}_j\{0\} = \mathbf{0}$ 。

第二步：對於第 p 個變數（截距項可視為單位向量的變數），令 $\mathbf{y}^* = \mathbf{y} - \sum_{j=1, j \neq p}^k \mathbf{f}_j\{l-1\}$ 。接著使用 $\mathbf{y}^*$ 對 x_p 進行 GWR 模型估計，得出新的 $\mathbf{f}_p\{l\}$ 的值， l 為疊代的次數。接著，將上述步驟對所有 k 個變數進行一遍。在每個 $\mathbf{f}_p\{l\}$ 的估計，則可以重新根據一般 GWR 的模型估計步驟給予不同的環寬。

第三步：重複第二步直到每項平滑函數 $\mathbf{f}_p\{l\}$ 收斂。

根據以上 CGWR 的估計流程，我們也發現在疊代過程中如果使用移動平均，即

$f_p^* \{l\} = (f_p^* \{l\} + f_p^* \{l-1\})/2$ ，收斂速度會更快。另外，上述估計流程也有數個地方可以略為調整以提高效率。一、Fotheringham *et al.* (2017) 提出以 GWR 當成初始值會收斂的更快；二、疊代過程可使用 GWR 的最佳環寬進行疊代，如此可免去重新校對環寬。我們將在本文比較 GWR 以及 CGWR。為了與 CGWR 原先設定一致，本文設定初始值為 0。第三節將以電腦模擬比較這三種模型的差異。

3. 模擬研究

本節的模擬研究安排如下，首先呈現在不同解釋變數的曲面函數設定下，GWR 的估計結果，主要以曲面的估計值、目標變數的估計誤差為評斷依據；另外，我們也將比較 GWR 與 CGWR 的估計差異。本節的模擬設定將參考 Leong and Yue (2017) 的曲面設定，同時為了簡化討論，本文僅考慮簡單線性迴歸的假設，亦即 $Y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)X_1(u_i, v_i)$ ，僅有一個解釋變數 X_1 ，考量截距、斜率兩個參數曲面，並以視覺化呈現估計結果。

模擬的空間環境為 10×10 的格子狀 (Regular lattice grid)，因此每筆模擬資料為 100 筆樣本。其模型設定如下。

$$Y_i(u_i, v_i) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)X_1(u_i, v_i) + \varepsilon_i(u_i, v_i),$$

$$\varepsilon_i(u_i, v_i) \sim N(\mu = 0, \sigma = 0.25)$$

其中 $X_1(u_i, v_i) \sim U(0.5 - d, 0.5 + d)$ 服從連續均勻分配， d 為自變數的範圍，其設定為 $d \in \{0.01, 0.02, \dots, 0.19, 0.20, 0.3, 0.4, 0.5\}$ ； $X_1(u_i, v_i)$ 與 $\varepsilon_i(u_i, v_i)$ 相互獨立。一般簡單線性迴歸中，自變數 X_1 係數估計值之標準誤 $Var(\hat{\beta}_1)$ 也與 X_1 的變動幅度有關，當自變數的標準差越大，則標準誤則越小；這種現象可能在 GWR 也會出現，本文將觀察不同 d 下的 GWR 估計結果。模擬過程在產生新的資料時 $X_1(u_i, v_i)$ 會重新抽樣，這種設定可稱為隨機效應模型 (Random effect model)，上述兩點與 Leong and Yue (2017) 不同。另外，因為核函數選擇對於估計結果影響較小，因此與多數研究類似，本文採用高斯距離遞減函數。

圖 3-1 為本文設定的兩個空間曲面，分別為線性（標示為 sp1）與二次拋物曲面（標示為 sp2），這兩個曲面在 GWR 實證分析中經常出現。由於 sp2（非線性）曲面較 sp1（線性）曲面複雜，預期 GWR 等模型的估計效果可能較不理想。另外，為了簡化本文模擬研究的設定，GWR 截距、斜率參數曲面的組合僅考慮以下兩種：情境一為

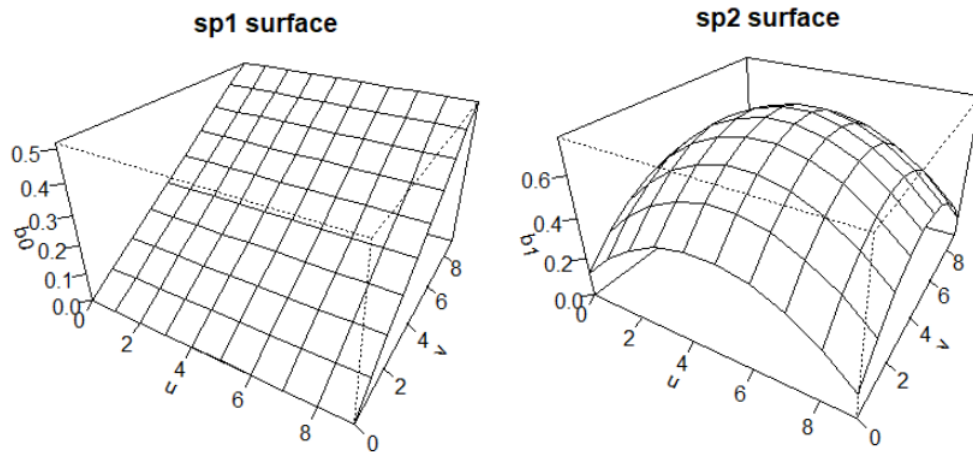


圖 3-1: 本文模擬使用的兩個迴歸係數空間曲面 (sp1、sp2)。

β_0 設定 sp1, β_1 設定 sp1; 情境二為 β_0 設定 sp1, β_1 設定 sp2。因為情境一的設定較為單純, 預期估計結果會優於情境二, 而截距和斜率曲面有不同假設, 其間可能產生交互作用。

上述的 d 值與解釋變數的變化程度有關, 進而影響 GWR 的參數曲面估計值, 我們將嘗試不同 d 值觀察影響程度。另外, 本文使用 R 軟體撰寫估計程式, 因為 GWR 的計算時間較長, 每個情境均考慮 500 次電腦模擬, 每次模擬採用了平均均方差 (Mean square error, MSE) 檢視目標變數、參數曲面等估計值。此外, 由於本文以空間資料的探索性分析為研究目標, 在此也會檢視 GWR 的參數估計值曲面, 評估視覺化呈現的可能性, 比較參數曲面估計值的平均值與理論值之間的差異。

3.1 GWR 與 CGWR 係數曲面估計表現

首先, 我們先檢視 GWR 的係數曲面估計。圖 3-2 為 $d = 0.01, 0.03, 0.05$ 在情境一的平均估計曲面, 明顯可見 GWR 估計值與理論值有不小差異, 顯示 GWR 無法獲得可靠的參數估計值, 無法以視覺化反映真實資訊。在不同 d 值之下, 不論 β_0 或 β_1 參數估計值都和線性曲面 (sp1) 不同, 但估計結果隨著 d 值增大而改善, 在 $d = 0.05$ 時與理論值很接近。情境二的估計結果亦類似 (圖 3-3), 參數估計值在 $d = 0.01$ 的設定下有相當大的偏誤 (bias), 隨著 d 值增大而改善。我們也可以透過 MSE 誤差顯示 d 值在兩種情境的參數估計值 (圖 3-4, $0.01 < d < 0.5$), 因為誤差數值差異頗大, 圖中的 MSE 經過對數調整。模擬的 MSE 誤差結果顯示 GWR 在 d 值較小時比較不穩定, 隨著 d 值上升漸趨穩定, 顯示使用 GWR 時需考量適用範圍。

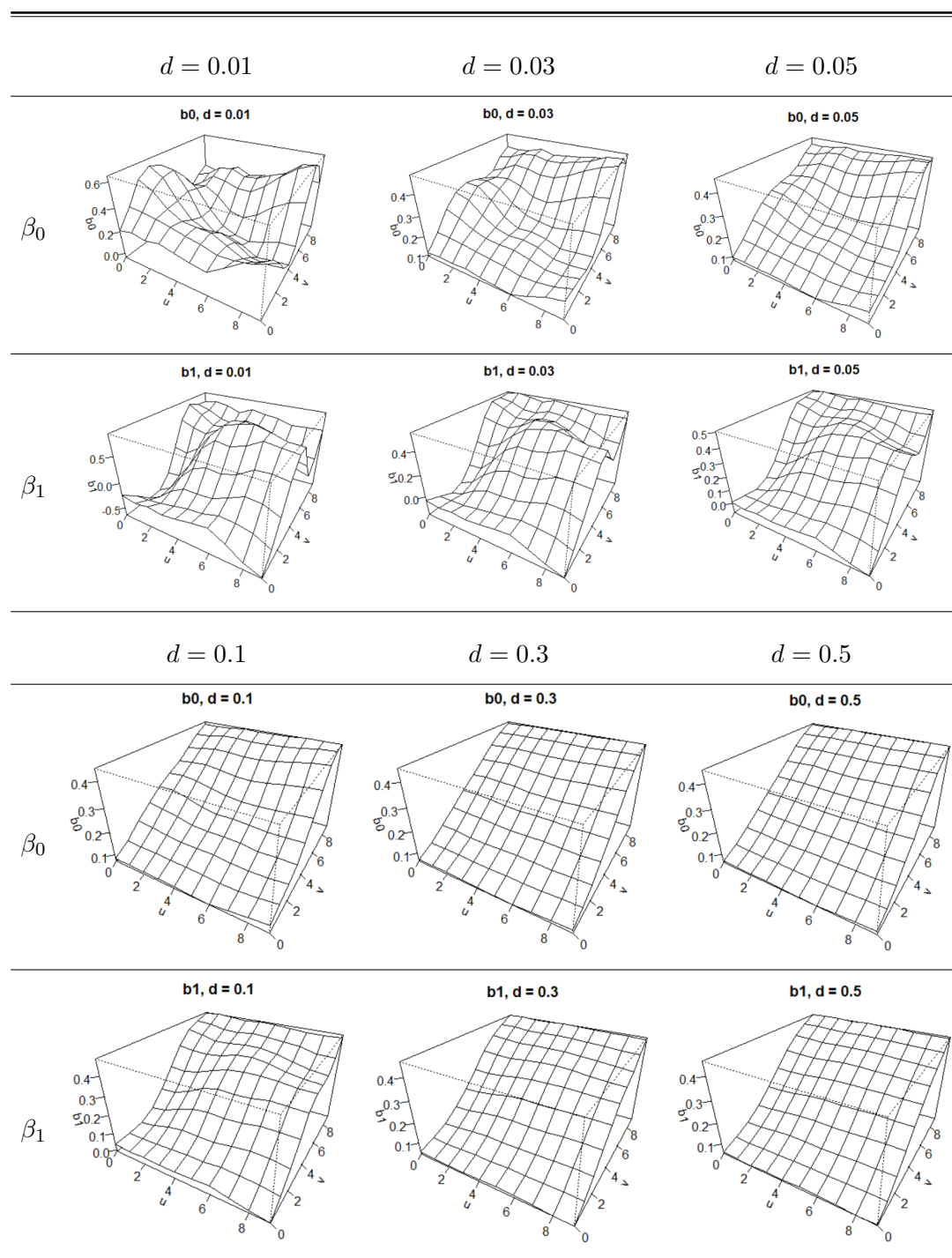


圖 3-2: GWR 於情境一的係數估計平均曲面。

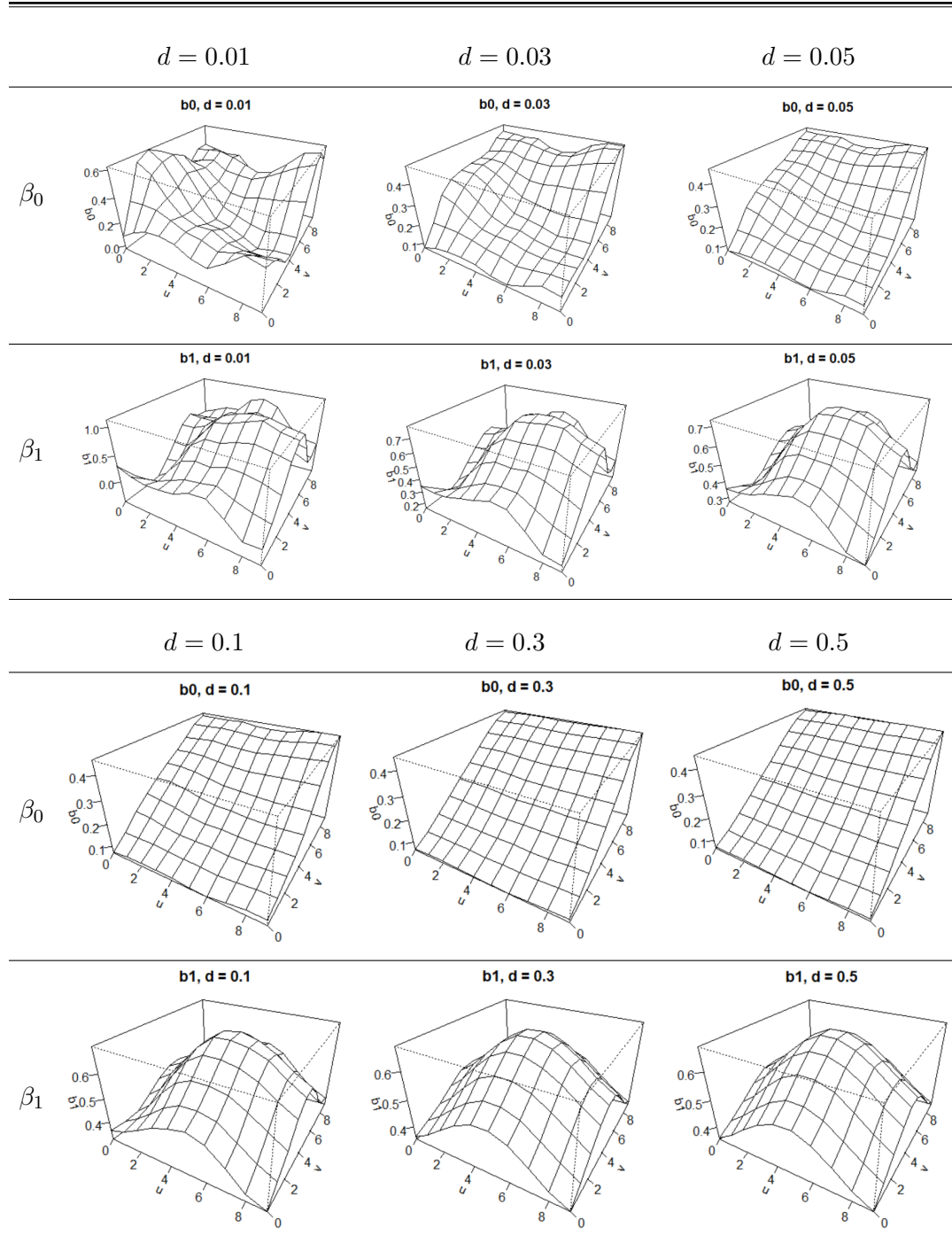


圖 3-3: GWR 於情境二的係數估計平均曲面。

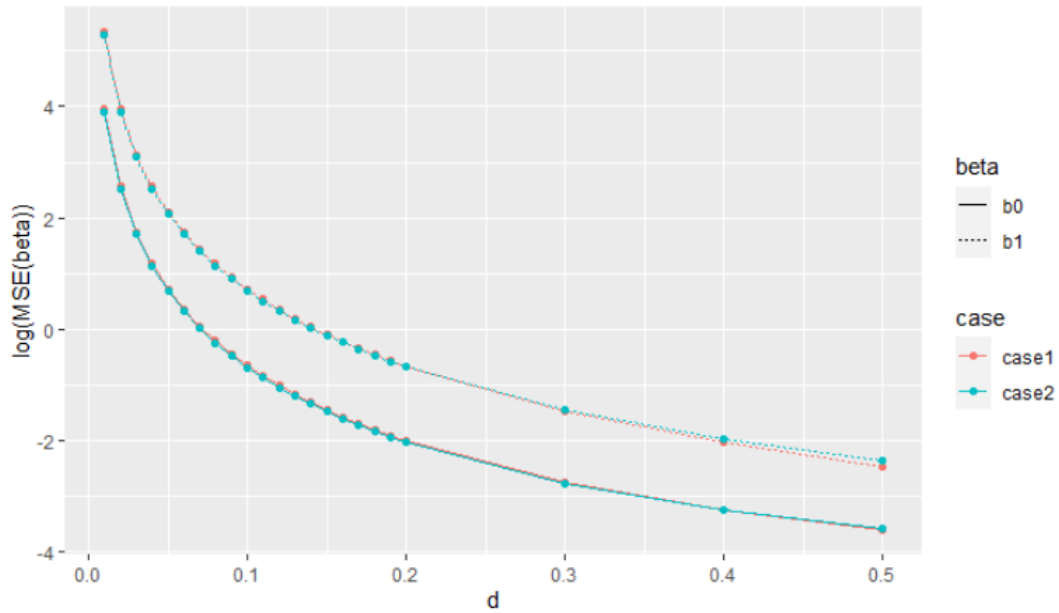


圖 3-4: GWR 估計曲面平均 MSE。

接著，我們進一步評估 CGWR 的估計結果，圖 3-5 與圖 3-6 分別是 CGWR 在情境一、情境二的平均參數估計曲面，CGWR 在 $d = 0.01, 0.03, 0.05$ 的結果都與理論曲面非常接近，優於 GWR 的估計。例如 $d = 0.01$ 時，由 GWR 的參數估計值看不出線性或二次拋物線，但藉由 CGWR 仍可大略看出真實曲面的形狀，亦即可透過 CGWR 作為視覺化的依據。GWR 和 CGWR 的差異也可透過誤差比值描述，亦即定義平均 MSE 比值 $avg(MSE_{CGWR}(\beta_j))/avg(MSE_{GWR}(\beta_j))$ ，比值愈小、CGWR 的優勢愈明顯。圖 3-7 為 CGWR 與 GWR 的誤差比值，總體而言， β_0 與 β_1 的平均 MSE 比值皆小於 1，尤其是 d 小於 0.1 的情境；隨著 d 上昇，GWR 的曲面估計效果也變得穩定。

此外，我們也以幾次 $d = 0.05$ 之下情境二（sp1、sp2 混合曲面）的模擬為例，說明 GWR 和 CGWR 參數曲面的估計差異（圖 3-8），CGWR 的曲面估計值明顯較佳。從圖中可見 GWR 所估計的曲面相當不平整，在第二個模擬中看不出有線性曲面；而 β_1 的 GWR 估計曲面也有相同的現象，甚至有些模擬結果（第三次模擬）與理論圖形相反， β_1 曲面估計值的邊緣高於中心。另一方面，CGWR 的估計大略可以反映真實曲面形狀，其中 β_0 估計曲面可反映原本的線性形狀； β_1 估計值稍微遜色，似乎還看不出拋物線的形狀，但邊緣估計值仍然低於中心，保留了二次曲面的部分特徵。這些模擬結果對資料視覺化非常重要，顯示 CGWR 的估計曲面可用於空間迴歸的探索性資料分析。

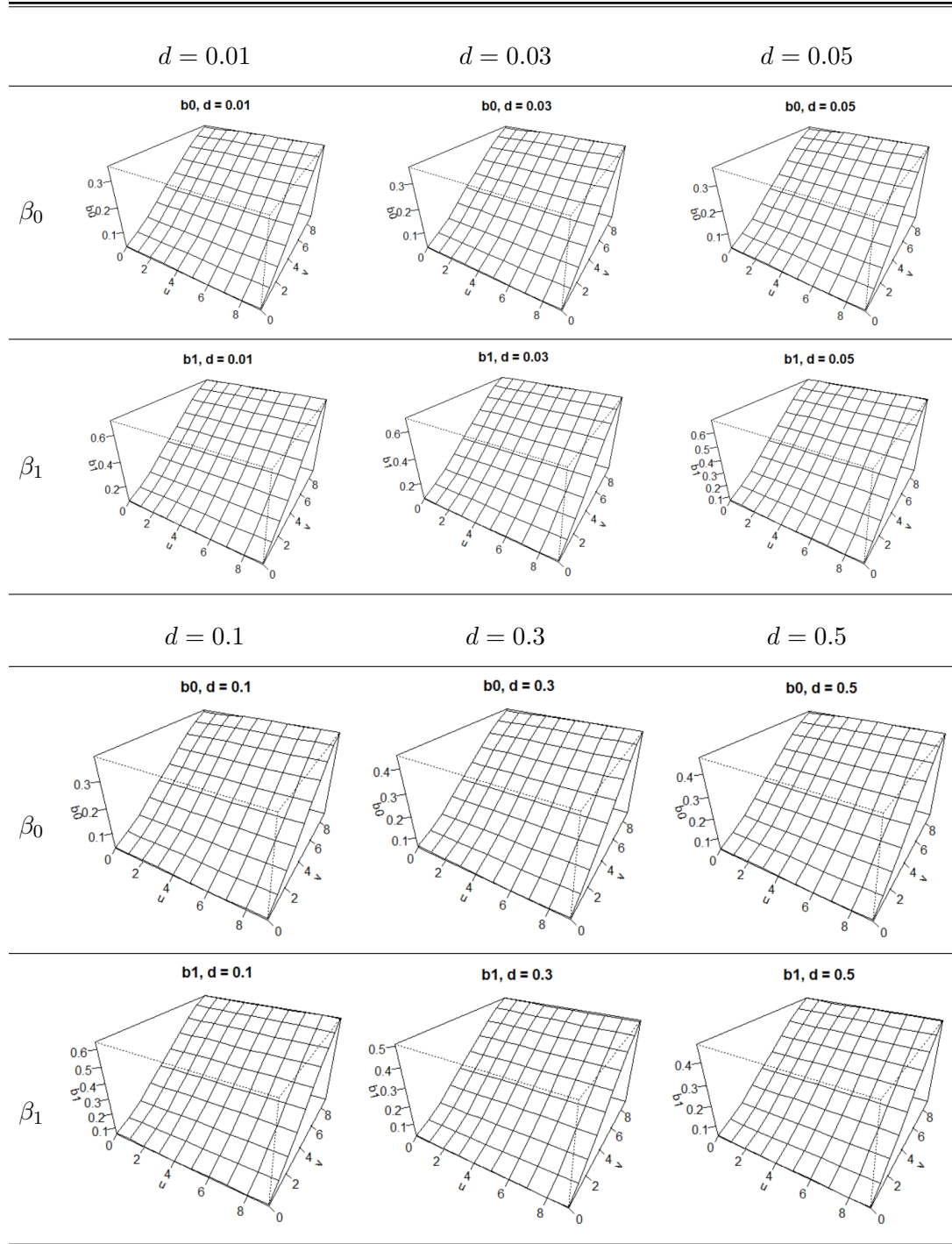


圖 3-5: CGWR 於情境一的係數估計平均曲面。

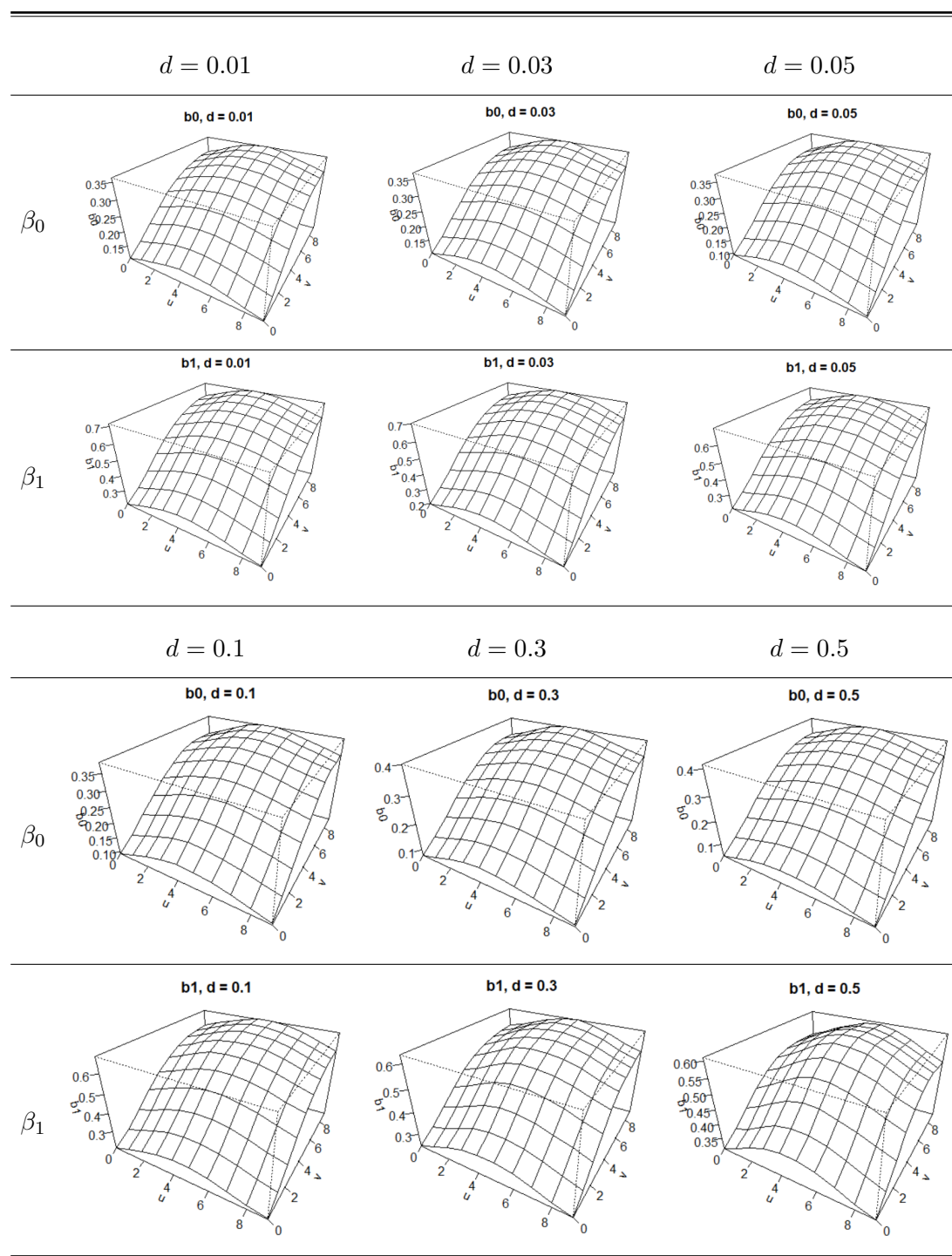


圖 3-6: CGWR 於情境二的係數估計平均曲面。

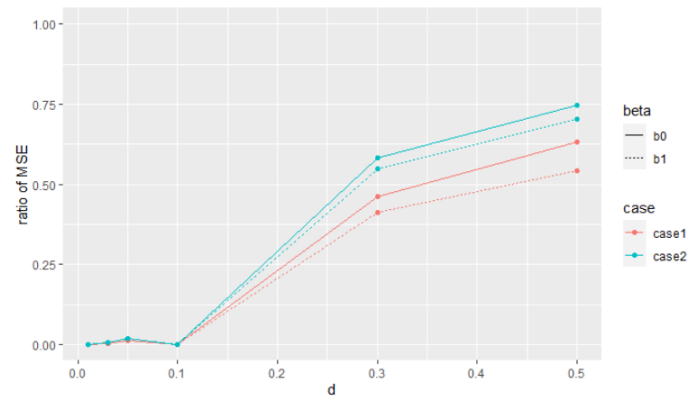


圖 3-7: GWR 與 CGWR 係數估計曲面的平均 MSE 比值。

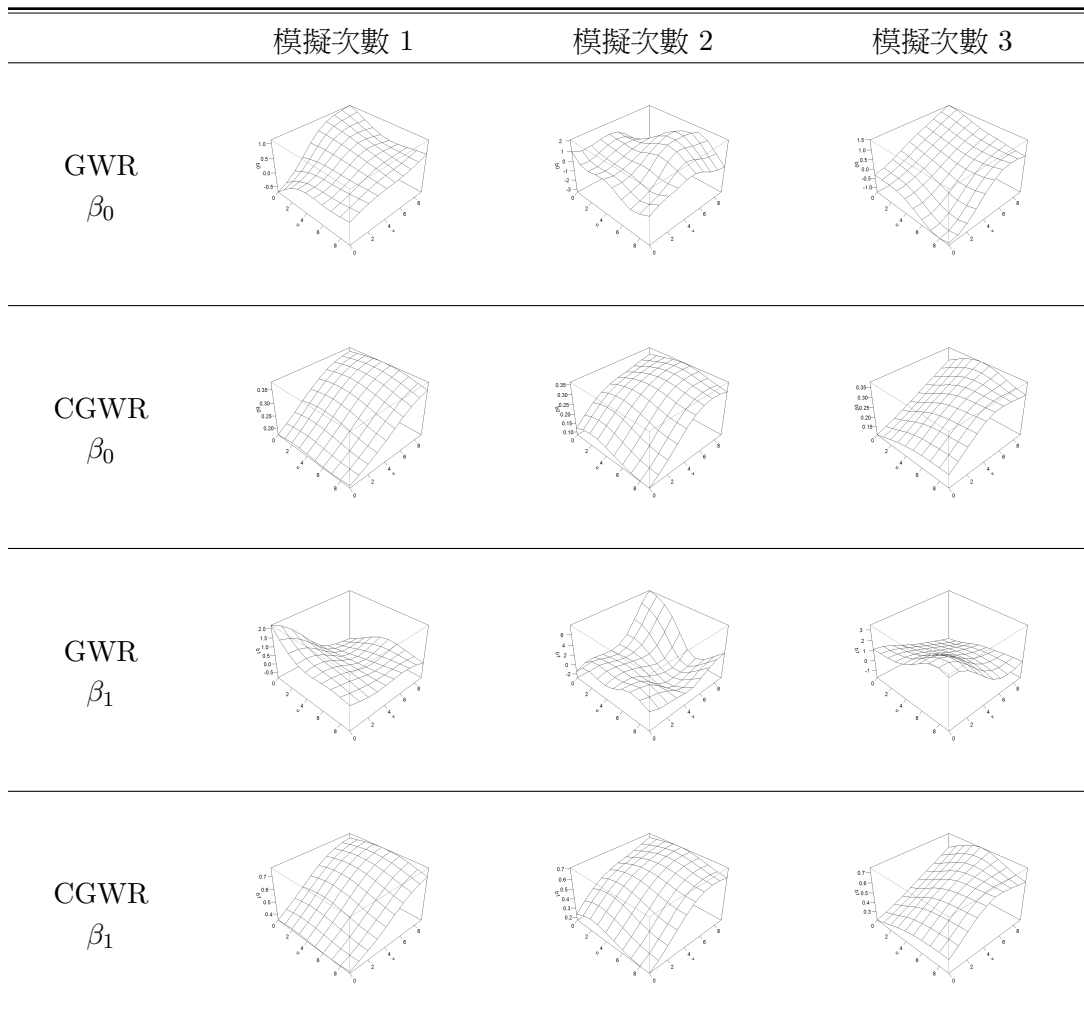


圖 3-8: GWR 與 CGWR 部份模擬資料之估計曲面。

3.2 GWR 與 CGWR 的目標變數之估計結果

接下來，我們分析目標變數 Y 的估計結果，除了沿用上述視覺化方式呈現曲面估計值，我們也以「虛擬 R^2 」(pseudo- R^2) 作為配適度指標 (goodness-of-fit)，評估 GWR 和 CGWR 的整體估計結果。虛擬 R^2 的定義與迴歸分析類似，衡量估計值 $\hat{y}$ 與理論值 $E(Y)$ 的差異，我們考慮兩種類型的虛擬 R^2 ，分別測量估計值在固定效果 (fixed-effect) 或隨機效果 (random-effect) 的差異。固定效果的虛擬 R^2 為 $cor(\hat{y}, E(Y|x))$ ；而隨機效果的虛擬 R^2 為 $cor(\hat{y}, E(Y))$ ；若虛擬 R^2 愈接近 1、估計值 $\hat{y}$ 的配適結果愈好，反之亦然²。由於每次模擬都能計算 $cor(\hat{y}, E(Y|x))$ 與 $cor(\hat{y}, E(Y))$ ，在此呈現虛擬 R^2 將以 500 次模擬的平均數。

圖 3-9 為 GWR 在不同 d 值下的目標變數的平均估計曲面。與前一小節 β_0 與 β_1 在的估計結果不同，估計值 $\hat{y}$ 相對穩定許多，即便在 $d = 0.01$ 的平均估計值也接近真實的反應曲面，亦即 GWR 的目標變數估計值遠較參數曲面估計值可靠，無論是線性或是二次曲面。CGWR 的 $\hat{y}$ 估計值結果也優於截距和斜率的估計值，參數曲面大致與理論值非常接近 (圖 3-10)。估計值的 R^2 配適度結果也類似，圖 3-11 為固定效果的虛擬 R^2 ，CGWR 的估計結果優於 GWR，但 d 值影響相對較小，GWR 在兩種曲面情境都有不錯的估計結果，不過 GWR 和 CGWR 在情境二的估計結果略遜於情境一。值得注意的是，CGWR 似乎有過度平滑的傾向 (over-smoothing)，亦即當 X 變異較小時 CGWR 的平滑化特性使估計較為準確，但當 X 變異變大時 CGWR 的平滑化反而成為劣勢，造成配適度下降。儘管如此，從圖 3-11 可以得知 CGWR 整體而言依然穩定。

相比之下，隨機效果的 R^2 配適度結果則略微不同，CGWR 的估計結果仍舊優於 GWR 較佳，但隨著 d 值卻有不一樣的變化趨勢 (圖 3-12)。當解釋變數 X 的產生波動時，其數值會與 $E(X)$ 有差異，因此 d 值愈大，GWR 與 CGWR 的目標變數震盪也愈大，使得估計值配適度愈小；平均而言，隨機效果的估計效果明顯不如固定效果，其中曲面較為複雜的情境二更易受到解釋變數波動帶來的影響， R^2 配適度隨 d 值的下降幅度更大。

²由於在本文設定的模擬資料為隨機效果模型。因此
 $E(Y_i(u_i, v_i)) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)E(X_i(u_i, v_i))$

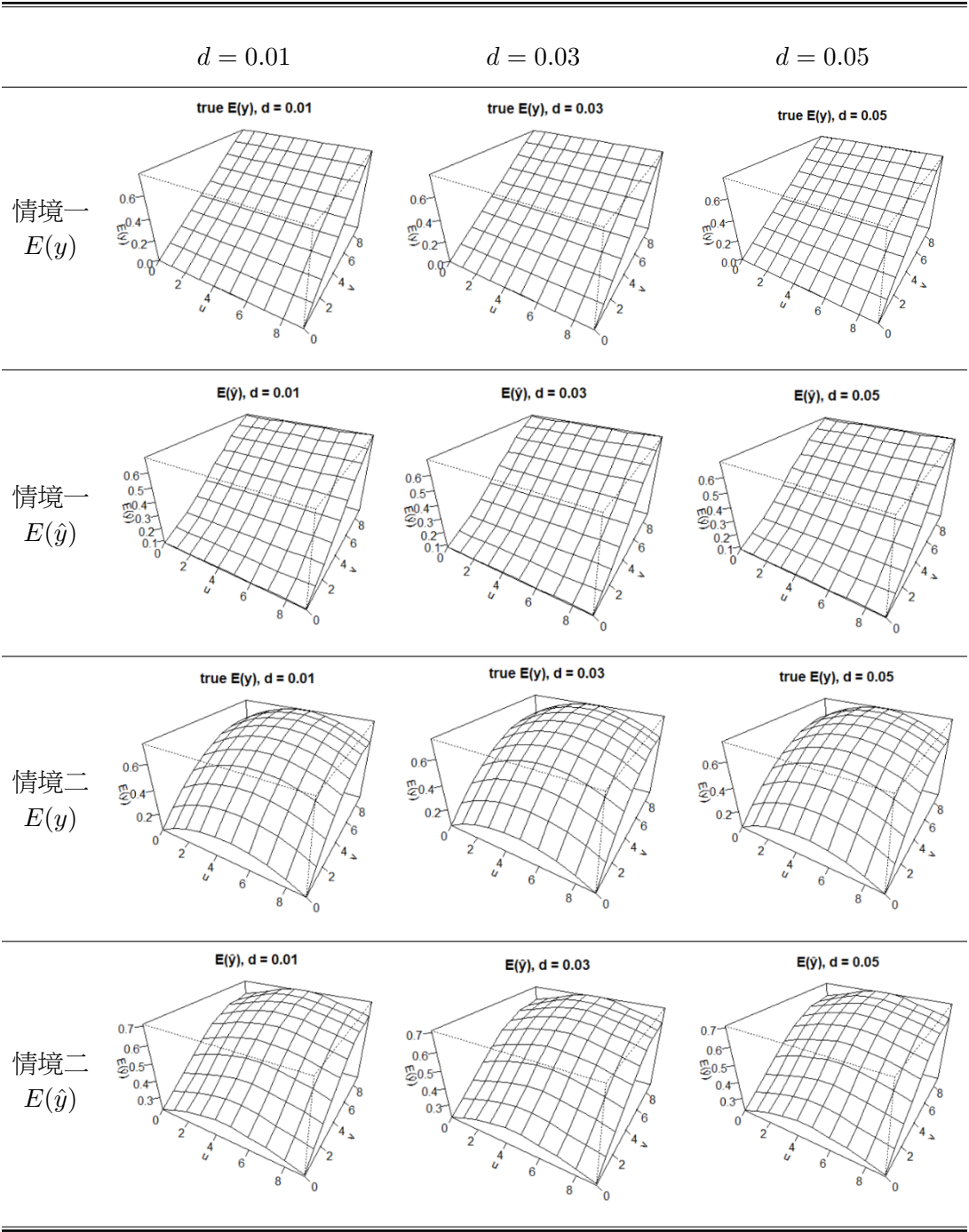


圖 3-9: GWR 的期望值估計曲面。

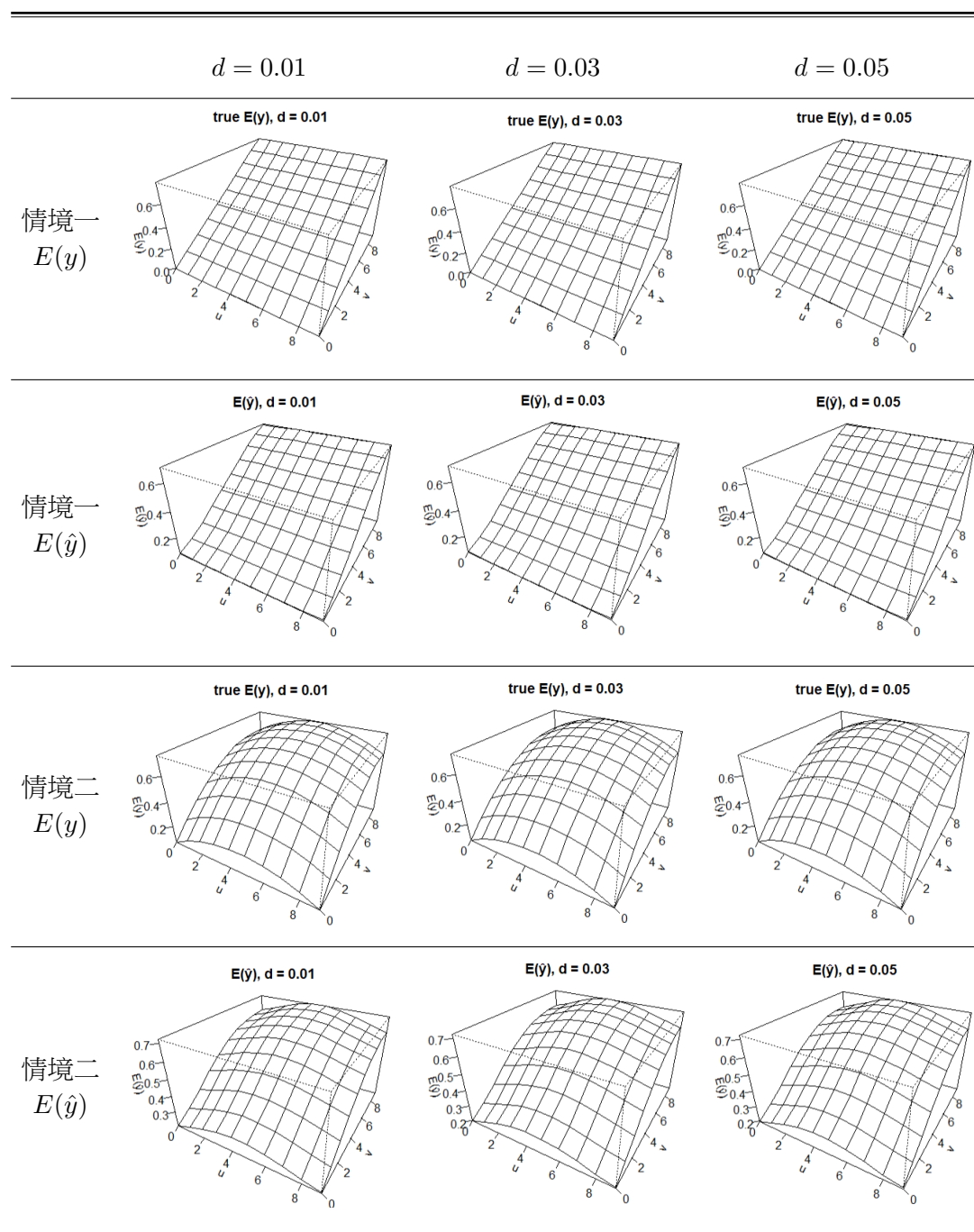


圖 3-10: CGWR 的期望值估計曲面。

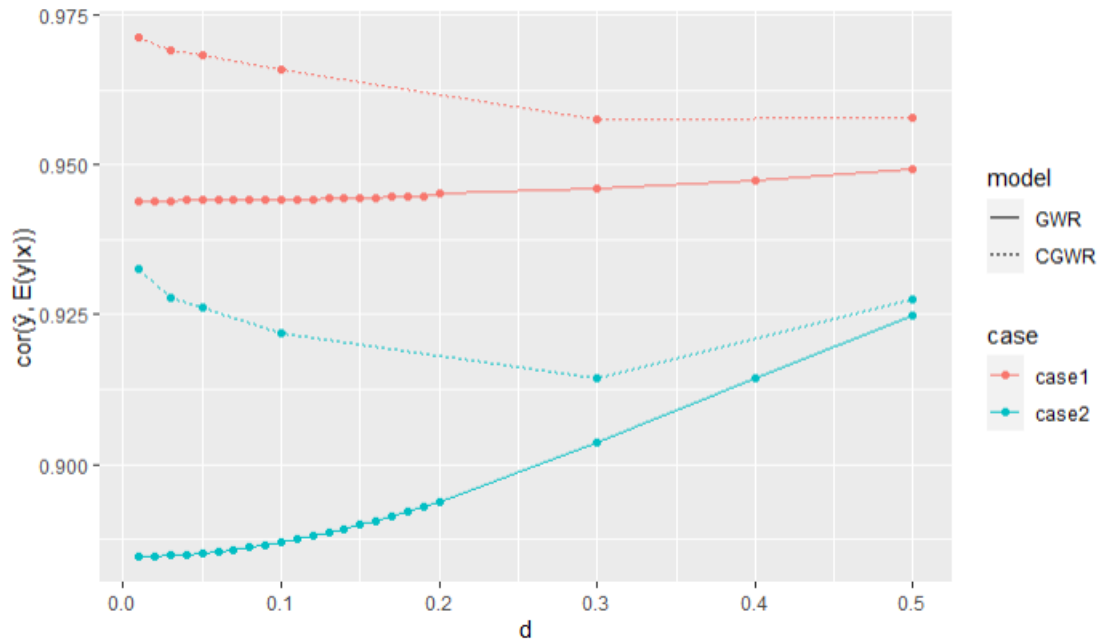


圖 3-11: GWR 與 CGWR 之固定效果虛擬 R^2 。

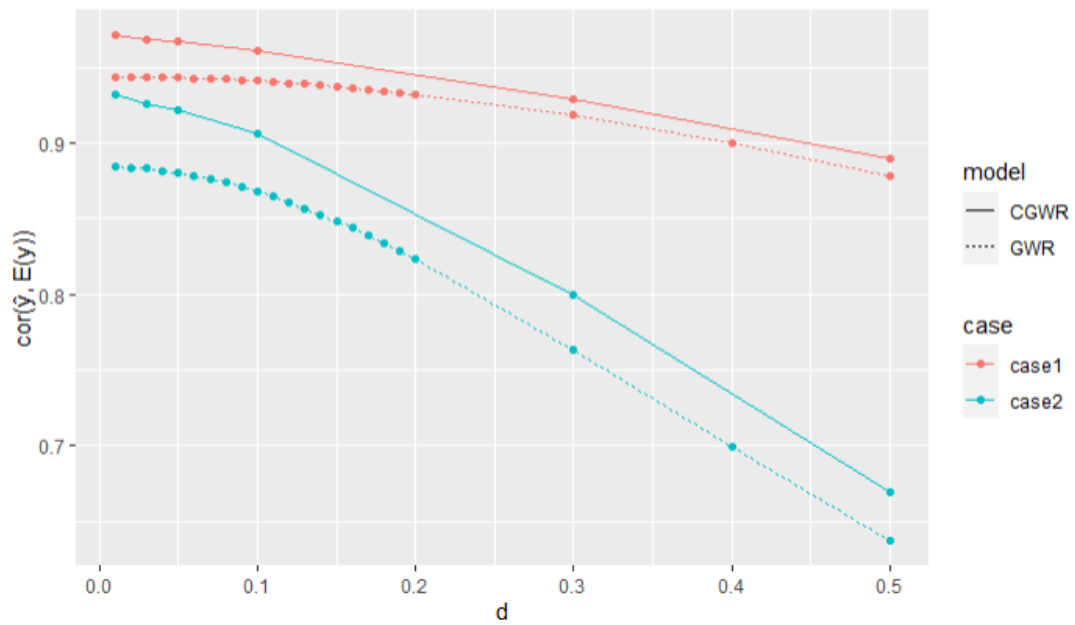


圖 3-12: GWR 與 CGWR 之隨機效果虛擬 R^2 。

4. 實證分析

本節將以 2005 年臺灣男性為研究對象，研究區域為臺灣本島 349 個鄉鎮市區，探討有哪些因素與各地區死亡率 (Y) 有關，我們使用 GWR 與 CGWR 進行局部空間關係建模。因為死亡率與年齡有關，我們選擇間接標準化調整法 (Indirect standardization)³，亦即以全國男性各年齡人口為標準母體，再套用標準死亡比 (Standardized Mortality Ratio, SMR) 至各鄉鎮市區的男性死亡率。除了目標變數死亡率外，在此選擇兩個解釋變數：各地區人口數 (X_1)、平均門診次數 (X_2)，這兩個變數同樣經過 SMR 的標準化處理。由於發現原始資料呈現右偏，因此本分析將死亡率、人口數、平均門診次數皆使用自然對數轉換，使其分佈較為對稱，更加符合 GWR 的模型假設。

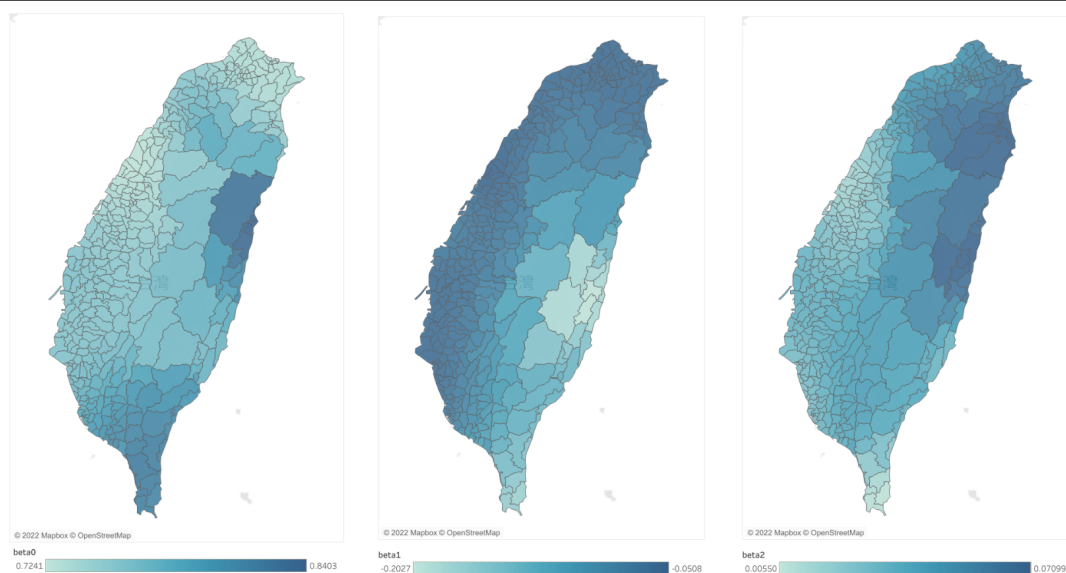
圖 4-1 為 GWR 與 CGWR 的 β_0 、 β_1 、 β_2 三個參數之曲面估計值，兩者的估計值頗為類似，僅在 β_0 有較大差異，因為前一節 CGWR 的參數曲面估計值較為準確，在此根據 CGWR 估計值詮釋結果代表的意義。人口數 (X_1) 的係數由西向東漸層加強 (負數越大)，人口數的係數多半小於 0；平均門診次數 (X_2) 的係數由西南向東北漸層增加，平均門診次數大於 0。綜合這些資訊可推得東部地區的死亡率，與人口數、平均門診次數的關係都較強，也就是東部人口數越多的地方、死亡率改善影響越為明顯；對於西部而言，這種效應並比較不明顯。換言之，東部鄉鎮市區的醫療資源較少，顯示臺灣各地醫療可近性 (medical accessibility) 存在地區不平等。GWR 與 CGWR 的 β_0 截距在南部（屏東縣附近）都有較高的標準化死亡率，但 CGWR 死亡率較高的鄉鎮市區幾乎涵蓋所有山地偏鄉，這個結果與山地原住民死亡率高於平地原住民相當一致。

GWR 與 CGWR 的 R^2 配適相關係數 $cor(\hat{y}, y)$ 分別為 0.841 與 0.941，意謂 CGWR 有較佳的估計結果。另外，圖 4-2 的標準化殘差分析圖顯示，GWR 模型的殘差值中有較多異常值，CGWR 的異常值明顯較少。GWR 標準化殘差甚至有四個觀察值 Z 值接近或超過 4，由於資料觀察值僅有 349 個（鄉鎮市區），GWR 殘差並不服從常態分配，顯示 GWR 估計值可能較不穩定；CGWR 的標準化殘差的離群值則較少。總括來說，如果配合 CGWR 的估計結果，更能以地理加權迴歸呈現視覺化的意義。

根據 Wheeler (2010) 的建議，GWR 模型的係數估計可使用平行坐標圖 (parallel coordinates plot) 進一步呈現視覺化訊息。圖 4-3 為 GWR 與 CGWR 的係數平行坐標

³死亡率與人口數來自內政部人口統計資料，平均門診次數則計算自全民健保資料庫，資料處理與擷取過程可參照 涂明憲 (2020)。

GWR



CGWR

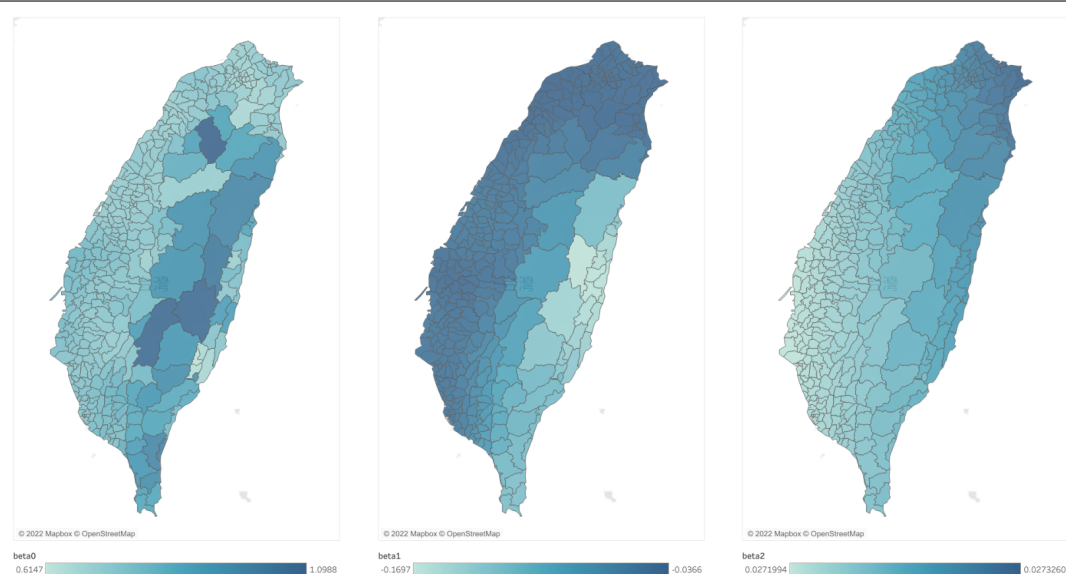


圖 4-1: 臺灣 SMR 之 GWR 與 CGWR 係數估計曲面。

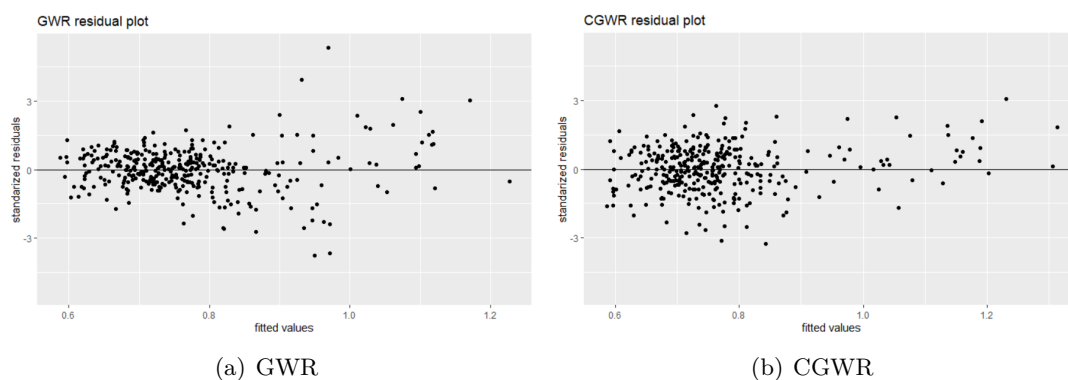


圖 4-2: 臺灣鄉鎮市區男性死亡率的標準化殘差分析

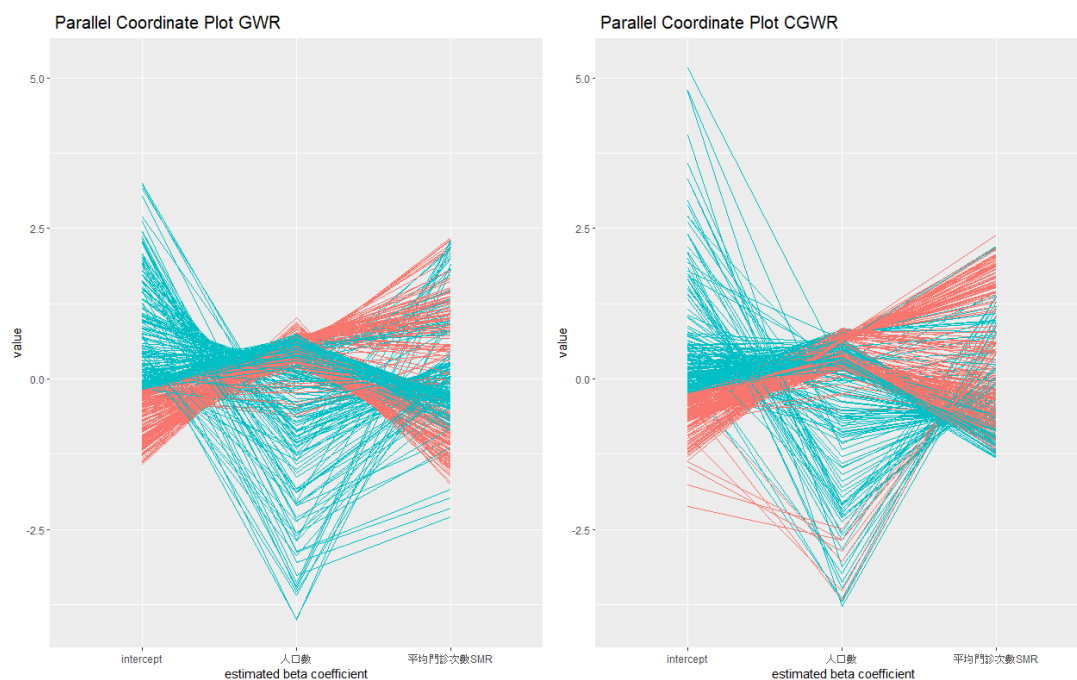


圖 4-3: GWR 與 CGWR 的係數平行坐標圖。

圖⁴，我們使用截距項的中位數作為區分，以藍色/紅色分別標識高/低死亡率地區，可以看出高死亡率地區的死亡率與人口數關係較為明顯。若配合圖 4-1 的參數曲面的估計值，則可推得死亡率愈高（截距越大）的地方，人口數所引起的負向效應愈發明顯（迴歸係數為負值），也與看診次數呈現正向效應（迴歸係數為正值）；相反地，死亡率較低的地區中，人口數所引起的負向效應則不明顯。綜合上述推論，可用於推測臺灣

⁴其係數為了便於呈現，故經由標準化處理。

死亡率較高地區，其醫療資源有不足之虞。這些空間視覺化顯示的結果，可作為政府設置醫療院所，以及分配有限公共資源的參考。

5. 結論與建議

資料視覺化在大數據時代非常重要，藉由設計適當的圖像工具，可提高分析龐雜資料的效率，尤其是帶有地區訊息的空間資料。Tukey (1977) 曾經提到 "The greatest value of a picture is when it forces us to notice what we never expected to see."，亦即圖像工具往往帶給我們不同視角的資訊。有別於一般迴歸模型，基於空間異質性衍生的 GWR 模型蘊含龐大資訊，包括（目標變數及迴歸係數）估計值、參數統計量（如 t 值）、 R^2 等等，如何解讀如此龐大的資訊並非易事 (Wheeler, 2010)。因此產生精確且有效的視覺化圖像對呈現 GWR 結果尤為重要。

透過模擬研究，本文發現 GWR 在目標變數的估計結果尚稱準確，但係數曲面的估計結果相當不理想，尤其是解釋變數的變動範圍較小時，係數曲面的估計值非常不穩定，未必能反映理論值的空間特性。如果需要描繪 GWR 的係數曲面，我們覺得 CGWR 更為合適，無論是係數曲面的形狀、或是係數 MSE 誤差都比較穩定。建議在考量空間資料視覺化時，可同時參考 GWR 與 CGWR 的結果，如果兩者的曲面一致，表示其結果更為可靠；反之，則建議參考 CGWR 的曲面估計結果。目標變數的估計值也是 CGWR 優於 GWR，但其誤差差異明顯小於係數曲面的估計值。另外，我們也發現在固定效果下，解釋變數愈大有助於期望值估計，但在隨機效應的假設下，解釋變數愈大反而導致目標變數估計值變差。

本文實證分析得出 CGWR 優於 GWR，此與先前 Leong 等人 (2017) 類似，不但有較佳的 R^2 配適度，殘差分析也顯示 CGWR 的離群值較少，其分佈也較集中在水平線上，配合平行坐標圖的分析係數分佈，也認為 CGWR 更適合用於實質詮釋。另外，本文也發現 CGWR 的估計曲面都比較平滑（圖 3-8），其環寬也較大，尤其在解釋變數的變異較小時（ d 值較小）、環寬值也較大。這些結果意謂 CGWR 可能有過度平滑化的傾向，適當的平滑固然可以降低估計誤差，但也可能扭曲了資料蘊含的訊息，未來將繼續探索如何避免 CGWR 過度修勻。

另一方面，本文考量的修正模型 CGWR 並非唯一的可能，MGWR 也常用於相關研究。由於兩種 GWR 修正模型的原理相似，使用的疊代估計方法也雷同，差異在於 CGWR 在疊代上採用 Jacobi 式，MGWR 則使用 Back-fitting 方法。Back-fitting 於更新參數的速度雖然比較快，但兩者收斂方向是否一致則缺乏相關研究，後續將比較

CGWR 和 MGWR 的差異。另外，本文將疊代的初始值設為 0，此與 MGWR 相關研究建議使用 GWR 為初始值不同，兩者之優劣需再詳加比較。

本研究限於篇幅，只使用了兩類參數曲面假設（線性曲面、二次曲面），未來將加入其他實證分析出現的曲面，像是常見的多峰曲面。其次，本文的電腦模擬僅考慮一個解釋變數，多個解釋變數下的結果則需另外驗證，根據我們先前的研究經驗，如果解釋變數的參數曲面間為負相關，則需調整 CGWR 的估計步驟，此與迴歸分析中解釋變數存有共線性 (Multi-collinearity) 類似，這個方向的研究在 GWR 相關文獻較為缺乏。最後，本文的解釋變數分配為連續均勻分配，但沒有測試對稱、右偏等分配，讀者引用本文結果需謹慎留意，避免因設定情境差異過大而過度推論。

參考文獻

- [1] Brunson, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical analysis*, 28(4), pages 281-298.
- [2] Brunson, C., Fotheringham, A. S., and Charlton, M. E. (1998). Geographically Weighted Regression-Modelling Spatial Non-stationarity. *Statistician*, 47(3), pages 431-443.
- [3] Chan, T., Chiang, P., Su, M., Wang, H., and Liu, M. S. (2014). Geographic Disparity in Chronic Obstructive Pulmonary Disease (COPD) Mortality Rates among the Taiwan Population. *PLOS ONE*, 9(5), pages e98170.
- [4] Edayu Z. N., and Syerrina, Z. (2018). A statistical analysis for geographical weighted regression. The 9th IGRSM International Conference and Exhibition on Geospatial & Remote Sensing, Malaysia.
- [5] Fotheringham, A. S., Brunson, C., and Charlton, M. E. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons.
- [6] Fotheringham, A.S., Yang, W., and Kang, W. (2017). Multiscale Geographically Weighted Regression (MGWR). *Annals of the American Association of Geographers*, 107(6), pages 1247-1265.

-
- [7] Leong, Y., and Yue, C. J. (2017). A Modification to Geographically Weighted Regression. *International Journal of Health Geographics*, 16(1).
 - [8] Matthews, S. A., and Yang, T-C. (2012). Mapping the Results of Local Statistics: Using geographically Weighted Regression. *Demographic Research*, 26, pages 151–166.
 - [9] Silverman, B. W. (1985). Spline Aspects of Spline Smoothing Approaches to Non-parametric Regression Curve Fitting. *Journal of the Royal Statistical Society, Series B*, 47(1), pages 1–52.
 - [10] Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, pages 234–240.
 - [11] Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass: Addison-Wesley Pub. Co.
 - [12] Wheeler, D. C. (2010). Visualizing and Diagnosing Coefficients from Geographically Weighted Regression Models. In: Jiang B., and Yao X. (Eds.) *Geospatial Analysis and Modelling of Urban Structure and Dynamics*, Springer Dordrecht, pages 415–436.
 - [13] Yu, H., Fotheringham, A. S., Li, Z., Oshan, T., Kang, W., and Wolf, L. J. (2020). Inference in Multiscale Geographically Weighted Regression. *Geographical Analysis*, 52(1), pages 87–106.
 - [14] 李宗儒、陳昭榮、李妙純 (2021)。臺灣大腸癌死亡率之空間分析。《臺灣衛誌》，40 卷 2 期，頁 225-240。
 - [15] 涂明蕙 (2020)。臺灣居民健康與壽命之空間分析。國立政治大學商學院統計學系碩士論文。

Using Geographically Weighted Regression for Spatial Data Visualization

Jack C. Yue^{1†}, Yin-Yee Leong² and Rou-Yun Kuo¹

¹ Department of Statistics, National Chengchi University

² Department of Risk Management and Insurance, Feng Chia University

ABSTRACT

Data appear in many different forms in the age of big data, and applications of data analysis have become more extensive recently. Data Visualization is the core of Exploratory Data Analysis, which is particularly important for the analysis of spatial data. Visualization tools, such as graphs and tables, can effectively present the main characteristics of the data, including spatial homogeneity and spatial autocorrelation. Geographically Weighted Regression (GWR) describes the local relationship between target variables and explanatory variables, and is used to show the change of variable relationship with geographic locations. This paper discusses the applicable timing of GWR, illustrates the limitations and possible problems of GWR through computer simulation, and tests whether the modification of GWR is effective.

Key words and phrases: Geographically Weighted Regression, Spatial Statistics, Data Visualization, Exploratory Data Analysis, Simulation.

JEL classification: C21, C61.

[†]Corresponding to: Jack C. Yue
E-mail: csyue@nccu.edu.tw