Journal of the Chinese
Statistical Association
Vol. 60, (2022) 95–124

基於廣義加成模型之房屋價格預測方法

黃士峰 [1,2†] **廖育苹** [2]

[1] 國立高雄大學應用數學系

[2] 國立高雄大學統計學研究所

摘　要

本研究針對不動產實價登錄資料，提出一結合廣義加成模型（generalized additive model，GAM）的系統性作法預測 2014 到 2019 年間高雄市的房屋價格，其中資料欄位包含數值型與類別型的解釋變數，例如：房屋位置、面積、型態…等。我們提出以 GAM 捕捉解釋變數對房價的非線性影響趨勢，並透過使用 AIC（或 BIC）決定哪些解釋變數需要透過非線性轉換方可提升模型的配適及預測表現。在實證研究方面則透過移動視窗法，以每 3、6、12、或 24 個月的資料進行模型配適，再將其應用於預測下 1 個月的房屋價格，以評估所提出模型的表現。特別地，我們採用機器學習領域中的 K 最近鄰（K nearest neighbor，KNN）方法所估計的房屋價格作為比較基準。數值結果顯示所提出的模型對高雄市房屋價格的配適與預測表現皆優於 KNN 法。

關鍵詞：廣義加成模型、實價登錄、薄板樣條。

JEL classification: C53, R32.

[†]通訊作者: 黃士峰
E-mail: huangsf@nuk.edu.tw

1. 緒論

在薪資成長趕不上物價上漲的年代，買不起房子成為近年來台灣年輕人最關心的話題之一，以高雄市為例，雖然房價沒有大台北地區這麼高不可攀，但在城市不斷進步、交通網逐漸完善、與大型購物商場及高科技產業進駐的誘因下，房價也從原本市中心的蛋黃區開始起漲，漲價區域亦逐漸向蛋黃區外圍擴散。房屋除了提供擁有人居住以外，同時也是一項在金融市場上重要的投資標的與理財工具。因此，關心房地產景氣且對於房價評估有興趣的市場參與人，除了包含具自購需求的消費者以外，房仲業、銀行、投資客、政策制定者、研究人員…等，也都希望能夠精進對房價評估的準確度（ Schulz and Werwatz, 2004 ）。影響房屋價格起伏的因素眾多，舉凡大環境的總體經濟情況與政府政策、房屋的地點、周遭環境、附近居民工作屬性…等都是可能的因素（ Kim and Park, 2005 ），但有部分因素往往面臨不易收集或量化（如附近居民工作屬性）的困難，因此一般房地產市場參與人或社會大眾也不易將其應用於評估過程。此一現象在台灣實施不動產實價登錄後有了改變，這一項政策使得每一位房地產市場參與人與社會大眾皆能透過實價登錄平台，自行上網查詢已完成交易之建物的相關資訊，如建物區段門牌、交易年月日、移轉層次、總樓層數、建物型態、建築完成年月、建物移轉總面積、建物現況格局（房、廳、衛）、有無管理組織、總價…等連續、計數、類別型態資料，而非過去僅能依賴房仲業者所提供的賣方資訊，同時也解決資料不易收集與量化的困難。基於這一些優點，本研究便嘗試透過收集不動產實價登錄網站上的公開資訊，並透過提出一適當的統計模型，探討實價登錄網站上的公開資訊對於評估高雄市房屋價格的適切性。

文獻上關於評估房價的方法非常多樣，Pagourtzi *et al.*（2003）將其分為傳統型（traditional type）與進階型（advanced type）兩大類，其中進階型模型表現都較傳統型模型優異，文獻上又以特徵價格法（hedonic pricing method）為最被廣為使用的進階型模型。特徵價格法首先將房價細分為數個組成部分（或稱為特徵），但實務上由於部分影響房價的重要特徵並非市場上的交易標的，故無法直接觀測到其價格，如：建物的周遭環境、便利性…等，故必須先將所有的重要特徵進行量化（如：透過計算該建物至最近的超級市場、捷運站、高速公路交流道的距離，作為度量該建物便利性的量化指標），再透過複迴歸（multiple regression）方法對房價進行建模，並將任一特徵所對應之迴歸係數解釋為其對房價的邊際貢獻（ Goodman, 1998；Malpezzi, 2003；Sirmans *et al.*, 2005 ）。換句話說，特徵價格法著眼於量化或評價無法直接觀測

到的建物特徵後，再透過建立迴歸模型將其應用於房價評估上（ Janssen *et al.*, 2001 ）。

然而，當特徵與房價間存在非線性關聯、特徵的量化指標存在巨大跳點或變異、或建物與建物間存在空間關聯性…等現象發生時，特徵價格法往往無法適當地捕捉這一些現象（ Selim, 2009 ）。因此，更多不同型式的房價評估方法便應運而生，如：人工神經網絡方法（artificial neural network，ANN）、模糊邏輯法（fuzzy logic method）、廣義線性模型（generalized additive model，GAM）、空間分析方法（spatial analysis method）…等。其中 Selim（2009）提出一應用 ANN 的方法預測土耳其房屋價格，其研究結果顯示人工神經網路能有效的捕捉房價與特徵間的非線性關聯，同時增加預測的準確度。Kuşan *et al.*（2010）引入模糊邏輯系統建模並預測房價，透過選擇與建物相關的環境因子、交通因子…等，應用於土耳其埃斯基謝希爾市（Eskisehir city）地區，其研究結果顯示所提出的方法對該城市不同區域的房價預測具有良好的表現。Dbrowski and Adamczyk（2010）以 GAM 對波蘭華沙的房價進行建模與預測，透過因子分析從 205 個經濟指標中選出合適的解釋變數，並藉由平滑函數對解釋變數進行非線性轉換後，再對房價進行建模與預測，其研究結果發現此一方法可有效降低預測誤差。Zhang *et al.*（2015）應用地理資訊系統（geographic information system）的資料，提出一空間分析法對中國武漢市房價、建物附近的道路密度、建物距離最近的湖泊多遠、與湖泊位於城市哪個區域等資料之間的關聯性進行建模，以將地理資訊納入房價評估的考量。綜合以上各方法的特點與基於台灣不動產實價登錄網站所提供的資料型態，本研究著眼於提出一在 GAM 的架構下，對高雄市房價進行建模與預測。

GAM 是由 Hastie and Tibshirani（1986）所提出，主要目的為有效刻劃解釋變數影響反應變數的非線性趨勢（nonlinear trend）。與傳統 GLM（generalized linear model）不同之處在於，GAM 將 GLM 中的解釋變數改為透過一合適的非線性函數轉換以捕捉解釋變數的非線性效應。在文獻上，GAM 已經被應用於解決各類問題，尤其在環境流行病學（environmental epidemiology）領域，研究證實 GAM 能協助分析者得到滿意的解釋和預測效果（ Dominici *et al.*, 2002；Wood and Augustin, 2002；Yang *et al.*, 2012；Wood *et al.*, 2015；De Souza *et al.*, 2018；Simpson, 2018；Wu and Zhang, 2019 ）。基於 GAM 具備處理不同解釋變數型態且刻劃各解釋變數所造成之非線性效應的特點，十分契合台灣不動產實價登錄網站所提供之公開資料型態與部分變數對房價呈現非線性影響的現象，故本研究提出一在 GAM 的架構下，考慮地理位置及相關

的房屋資訊影響房價的非線性趨勢，並透過選模準則判斷應該將哪一些解釋變數進行適當的非線性轉換以反映其影響房價的非線性趨勢，對高雄市房價進行建模與預測。

此外，為了評估所提出方法的配適與預測表現，本研究亦引用 K 最近鄰（K nearest neighbor，KNN）法（ Altman, 1992；Lowe, 1995；Cunningham and Delany, 2021 ）對房價進行預測，KNN 法主要是透過度量預測標的建物的特徵與訓練集內所有建物特徵的距離後，從訓練集內選取若干特徵與標的建物最為相近者，將其所對應之房價的（加權）平均值視為標的建物的房價預測值。由於透過 KNN 法預測房價十分合乎直覺且亦為大部分購屋者評估房屋價值的原則，因此本研究將 KNN 法所預測的房價視為評估基準並與所提出方法的預測結果進行比較。在實證研究中，我們採用內政部高雄市 2014 年 1 月到 2019 年 12 月之間的不動產實價登錄資料，同時為了反映市場最新房價趨勢以及部分解釋變數對房價影響的非線性趨勢亦隨時間改變而動態調整之現象，我們提出以移動視窗（rolling window）的方式進行分析。在本研究中，移動視窗的設計為以每 3、6、12 或 24 個月資料進行建模後再用以預測下 1 個月的房價，接著將視窗向後移動 1 個月再重複以上步驟，亦即視窗大小為 3、6、12、或 24 個月，模型更新頻率為 1 個月。實證結果顯示在上述四種不同視窗大小的移動視窗架構下，所提出方法無論在房價配適或是預測效果的表現皆優於 KNN 法。

本文在本小節後的編排如下：第二節為資料介紹，說明不動產實價登錄資料之來源以及資料的預處理過程，並舉例說明部分解釋變數影響房價的非線性趨勢；第三節介紹所提出的建模步驟與估計方法；第四節呈現模擬與實證分析結果，探討所提出估計方法的合理性與預測的準確性；第五節對本研究的結果進行討論與總結；理論推導過程則置於附錄中。

2. 資料介紹

本節首先介紹實價登錄網上下載的資料內容與本研究所提出的資料預處理過程，接著再透過圖形呈現實價登錄網上可供下載與建物相關資料的特徵及其影響房價的非線性效應。

2.1 實價登錄不動產資料說明

本研究資料於行政院內政部不動產交易實價查詢服務網之 Open data 下載（https://plvr.land.moi.gov.tw/DownloadOpenData）。表 1 列出幾筆資料的範例，其中資料欄位包含鄉鎮市區、交易標的、土地區段位置/建物區段門牌、土地移轉總面

表 1: 實價登錄不動產原始資料範例。

鄉鎮市區	大寮區	鼓山區	小港區	三民區
交易標的	房地 (土地 + 建物)	房地 (土地 + 建物)+ 車位	土地	車位
土地區段位置/建物區段門牌	高雄市大寮區仁德路 121~150 號	高雄市鼓山區美術北五街 31~60 號	孔宅段 231~1260 地號	高雄市三民區褒忠街 151~180 號
土地移轉總面積平方公尺	80.62	16.06	95	2.69
都市土地使用分區	住	住	住	住
非都市土地使用分區				
非都市土地使用編定				
交易年月日	1030120	1030120	1030123	1030120
交易筆棟數	土地 1 建物 1 車位 0	土地 1 建物 1 車位 1	土地 1 建物 0 車位 0	土地 0 建物 0 車位 3
移轉層次	全	十層		地下二層
總樓層數	四層	二十七層		二十二層
建物型態	透天厝	住宅大樓 (11 層含以上有電梯)	其他	其他
主要用途	見其他登記事項	住家用		見其他登記事項
主要建材	鋼筋混凝土造	鋼筋混凝土造		鋼筋混凝土造
建築完成年月	961011	1020301		981130
建物移轉總面積平方公尺	175.54	204.79	0	42.71
建物現況格局-房	4	3	0	0
建物現況格局-廳	1	2	0	0
建物現況格局-衛	3	2	0	0
建物現況格局-隔間	有	有	有	有
有無管理組織	無	有	無	有
總價元	5800000	14500000	6550000	1500000
單價元/平方公尺	33041	70804	68947	
車位類別		坡道平面		坡道平面
車位移轉總面積平方公尺	0	16.66	0	46.95
車位總價元	0	0	0	1500000
備註				單獨車位交易

積（平方公尺）、都市土地使用分區、非都市土地使用分區、非都市土地使用編定、交易年月日、交易筆棟數、移轉層次、總樓層數、建物型態、主要用途、主要建材、建築完成年月、建物移轉總面積（平方公尺）、建物現況格局-房、建物現況格局-廳、建物現況格局-衛、建物現況格局-隔間、有無管理組織、總價（元）、單價（元/平方公尺）、車位類別、車位移轉總面積（平方公尺）、車位總價（元）、備註等欄位。

本研究自表 1 中選取以下幾個欄位的資料作為影響房價的解釋變數並對建物單價（萬元/坪）進行建模：土地區段位置/建物區段門牌、土地移轉總面積、移轉層次、建物型態、建築完成年月（屋齡）、建物移轉總面積、建物現況格局-房、建物現況格局-廳、建物現況格局-衛、建物現況格局-隔間、有無管理組織等。

2.2 資料預處理

本研究選取高雄市 2014 年 1 月 1 日到 2019 年 12 月 31 日間交易資料進行分析，其中在都市土地使用分區這個欄位選擇「住」；在主要用途這個欄位選擇「住家用」或「國民住宅」；在移轉層次這個欄位選擇「單獨一層」或「全」，不考慮「地下層」；在建物型態上選取「透天厝」、「公寓」、「住宅大樓」、或「華廈」等 4 種類型，並去除建築完成年月日有缺失值的資料；將單價、總價、土地移轉總面積、建物移轉總面積單位平方公尺皆改為坪，以反映一般台灣民眾在討論房價時常以坪數作為單位的習慣；在移轉層次、建物型態、建物現況格局-隔間與有無管理組織等欄位的文字進行數值轉換，其中在移轉層次（或交易樓層）的部分，將國字轉換為數值，如：三層轉換為 3，「全」為透天厝，本研究將其轉換為 0，意即在模型中暫不考慮移轉層次對透天厝房價的影響；在建物型態的部分，1 表示透天厝、2 表示公寓、3 表示華廈以及 4 表示住宅大樓；在建物現況格局-隔間與有無管理組織的部分，1 代表有、0 代表無；並利用建築完成年月日這一欄的資料計算屋齡；利用土地區段位置/建物區段門牌這個欄位的資料，經轉換後得到建物的經度及緯度。此外，本研究刪除土地移轉總面積及建物移轉總面積大於 200 坪或小於 1 坪、屋齡大於 60 年以及房、廳、衛大於 10 的資料，以降低離群值對模型估計所造成的偏誤。

表 2 列出本文所採用的變數代號、名稱、與其量化後的數值型態，特別地，這一些變數中包含連續型態（如：經緯度、土地移轉面積…等）、計數型態（如：交易樓層、屋齡…等）、與類別型態（如：建物型態、有無隔間…等）的變數，這一項資料固有的特性也加深對房價建模的困難度。

表 2: 變數代號、名稱、與其數值型態。

變數代號	變數名稱	數值型態
$y^{(0)}$	單價	正實數（萬元/坪）
a_1	經度	正實數
a_2	緯度	正實數
x_1	log(土地移轉總面積)	實數
x_2	log(建物移轉總面積)	實數
x_3	交易樓層	非負整數
x_4	建物型態	透天厝（1）、公寓（2）、華廈（3）、住宅大樓（4）
x_5	屋齡	非負整數
x_6	建物現況格局 (房)	非負整數
x_7	建物現況格局 (廳)	非負整數
x_8	建物現況格局 (衛)	非負整數
x_9	有無隔間	無（0）、有（1）
x_{10}	有無管理組織	無（0）、有（1）

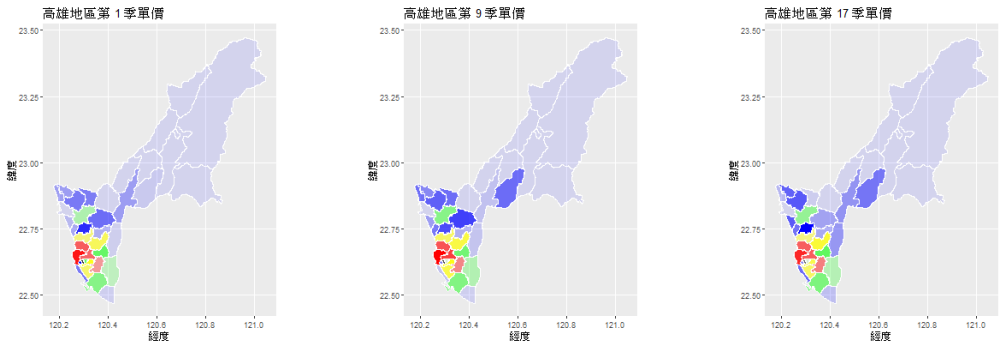


圖 1: 由左至右分別為 2014 年 1 月至 3 月、2016 年 1 月至 3 月、2018 年 1 月至 3 月高雄市各行政區內建物成交之平均單價（萬元/坪）的熱圖，其中價格高低以紅、黃、綠、藍色系依序排列，同一色系顏色越深表示價格越高。

2.3 資料特徵

首先，圖 1 呈現 2014 年 1 月至 3 月、2016 年 1 月至 3 月、2018 年 1 月至 3 月高雄市各行政區內建物成交之平均單價（萬元/坪）的熱圖，由圖中可見房屋的平均成交單價除了會隨時間改變以外，也的確會受到地理位置的影響，離原本高雄市中心較近

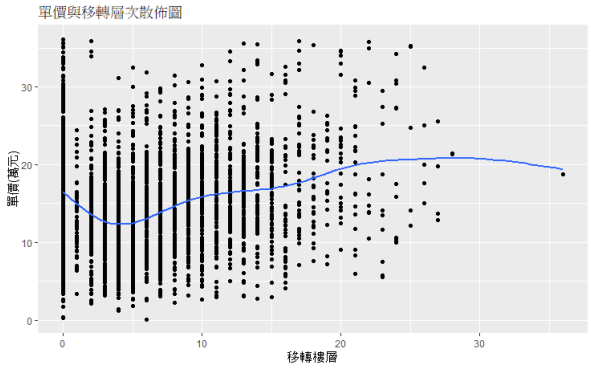


圖 2: 2014 年 4 月至 6 月，高雄市房價與移轉層次散佈圖，橫軸為建物交易移轉樓層，0 代表此建物為透天厝，其餘代表其交易樓層，縱軸為單價（萬元/坪），圖中藍線為經由薄板樣條函數轉換後的曲線，可視為移轉樓層對房價影響的趨勢線。

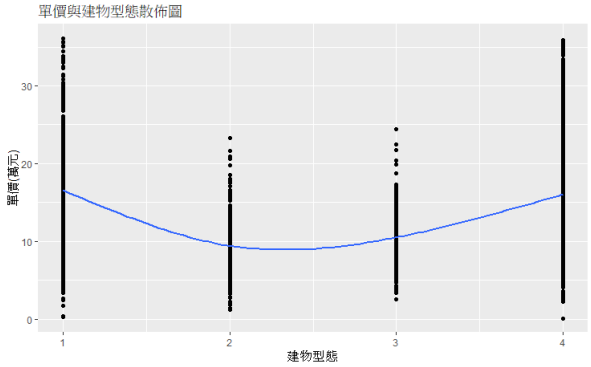


圖 3: 2014 年 4 月至 6 月，高雄市房價與建物型態散佈圖，橫軸為交易的建物型態，型態 1~4 分別表示透天厝、公寓（5 樓含以下無電梯）、華廈（10 層含以下有電梯）、住宅大樓（11 層含以上有電梯），縱軸為單價（萬元/坪），圖中藍線為經由薄板樣條函數轉換後的曲線，可視為建物型態對房價影響的趨勢線。

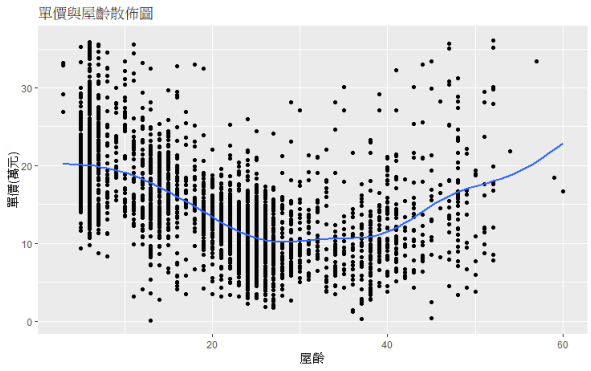


圖 4: 2014 年 4 月至 6 月，高雄市房價與建物屋齡（年）散佈圖，圖中藍線為經由薄板樣條函數轉換後的曲線，可視為屋齡對房價影響的趨勢線。

的行政區房價較高。因此，本研究除了透過表 2 中的經緯度資料 (a_1, a_2) 描述建物所在位置外，亦同時將其視為影響房價的因子。

接下來我們以 2014 年 4 月到 6 月間之資料為例，圖 2~4 分別呈現房價（$y^{(0)}$）和表 2 中之交易樓層（x_3）、建物型態（x_4）、屋齡（x_5）等三個變數的散佈圖，由圖中可以發現，房價與上述三個變數間存在非線性關係，換言之，直接透過線性模型且經由上述變數對房價進行建模（如：複迴歸模型）並不合適。特別地，在圖 4 的右上角區域呈現出屋齡越高但單價反而越高的現象，透過建物所在位置的經緯度（a_1、a_2）分析後發現，造成此現象的原因為此部分的建物大多位於高雄市市中心，建物型態（x_4）多為透天厝，近年來因市中心地價逐年攀升，位於市中心的透天厝儘管屋齡漸增，但房價卻反而水漲船高。同時經由上述討論亦可發現，房價與單一變數之間的關係也會同時受到其他變數的影響，因此，對房價建立模型應同時考慮多重變數的作用。

此外，雖然房價與上述三個變數間的關係無法以簡單的線性函數描述，但經由配適 Wood and Augustin（2002）與 Wood（2003）所提出的薄板樣條函數（thin plate spline）轉換後，可以發現所得到的曲線（即圖 2~4 中的藍線）可以有效地描繪出上述三個變數對房價影響的趨勢，故可視為個別變數對房價之非線性效應的趨勢線。在表 2 中所列舉的變數，經由本研究所考慮的資料中發現房價與解釋變數 x_1、x_2、x_9、x_{10} 間適合以線性關係描述，但房價與解釋變數 x_3、x_4、x_5、x_6、x_7、x_8 間則會在部份連續 3 個月的資料中呈現明顯的非線性關係，為節省空間，在此不一一贅述。根據此一發現，本研究在後續經由表 2 中所列舉的解釋變數對房價建立模型時，提出應將影響房價之非線性趨勢納入建模考量，以增加模型的配適及預測能力。

3. 模型與方法

誠如上一節所言，對房價建立模型時，由於房價與解釋變數間存在非線性關係，因此，直接透過複迴歸模型對原始資料進行建模並不合適。為了突破此一困難，本文提出先透過非線性擬合技術找到解釋變數影響房價的非線性趨勢後，再結合複迴歸模型對房價進行建模較為合適。因此，本研究首先提出以下的模型對實價登錄網上的房價資料進行建模：

$$y_i = \alpha_0 + f_1(a_{1i}, a_{2i}; \boldsymbol{\theta}_1) + \sum_{j \in \mathrm{A}} f_j(x_{ji}; \boldsymbol{\theta}_j) + \sum_{k \in \{1,2,9,10\} \cup \mathrm{B}} \alpha_k x_{ki} + \varepsilon_i \qquad (1)$$

在 1 式中 y_i 為第 i 筆觀測資料之建物單價 $y_i^{(0)}$ 經下式轉換後的反應變數：

$$y_i = \frac{y_i^{(0)} - m_1}{m_2} \tag{2}$$

其中 m_1、m_2 分別為 2014 年 1 月到 3 月房價資料之平均數及標準差。此外，1 式中的 a_{1i}、a_{2i} 分別為第 i 筆觀測資料之經度、緯度，x_{ji}、x_{ki} 為第 i 筆觀測資料之交易樓層、建物型態…等列舉於表 2 中的解釋變數值，A 與 B 為兩個解釋變數的指標集合，且滿足 $\mathrm{A} \cup \mathrm{B} = \{3,4,5,6,7,8\}$ 以及 $\mathrm{A} \cap \mathrm{B} = \emptyset$，$\alpha_0$ 與 α_k 為參數，f_1 為經度和緯度之 2 維非線性轉換函數且 $\boldsymbol{\theta}_1$ 為其相對應的參數向量，f_j 為第 j 個解釋變數之非線性轉換函數且 $\boldsymbol{\theta}_j$ 為其參數向量，ε_i 為獨立且分佈為 $N\left(0,\sigma^2\right)$ 之誤差項。為了便於闡述，本研究皆採用薄板樣條函數對 1 式中的變數進行非線性轉換，並未深入探討採用不同的非線性轉換函數或是在非線性轉換函數中採用不同結點（knots）設定之個數與位置對建模與預測的影響。

此外，為了刻劃 2.3 節中所提及的房價與解釋變數 x_1、x_2、x_9、x_{10} 間呈現線性關係，但與解釋變數 x_3、x_4、x_5、x_6、x_7、x_8 間則會在部份連續 3 個月的資料中呈現明顯的非線性關係之資料特性，我們將 1 式中的模型設計為同時包含 LM（linear model）以及 GAM 兩部分，在 1 式右側，$\alpha_0 + \sum_{k\in\{1,2,9,10\}\cup \mathrm{B}} \alpha_k x_{ki}$ 為 LM 部份，$f_1(a_{1i}, a_{2i}; \boldsymbol{\theta}_1) + \sum_{j\in \mathrm{A}} f_j(x_{ji}; \boldsymbol{\theta}_j)$ 則為 GAM 部份，但在 x_3~x_8 等 6 個解釋變數中，哪一些解釋變數應安排於 LM 或 GAM 部份，則透過資料與稍後介紹的選模機制決定。特別地，1 式中的模型亦將經緯度資料直接安排於 GAM 部份（即 $f_1(a_{1i}, a_{2i}; \boldsymbol{\theta}_1)$ 項），並為其建立一 2 維薄板樣條轉換函數，其中 f_1 與 f_j 的薄板樣條函數表達式如下：

$$f_1(a_{1i}, a_{2i}; \boldsymbol{\theta}_1) = \sum_{h=1}^{t_1} \theta_{1h} \boldsymbol{b}_{1h}(a_{1i}, a_{2i}), \boldsymbol{\theta}_1 = (\theta_{11}, \theta_{12}, \ldots, \theta_{1,t_1})^T,$$

$$f_j(x_{ji}; \boldsymbol{\theta}_j) = \sum_{h=1}^{t_j} \theta_{jh} \boldsymbol{b}_{jh}(x_{ji}), \boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{j,t_j})^T,$$

其中 $\boldsymbol{b}_{1h}$ 與 $\boldsymbol{b}_{jh}$ 為基底函數，θ_{1h} 與 θ_{jh} 為相對應的係數，t_1 與 t_j 分別為 f_1 與 f_j 所採用的基底函數個數，並採用廣義交叉驗證（generalized cross validation, GCV）決定基底函數個數 t_1 與 t_j 的設定。

特別地，在本研究所採用的移動視窗架構下，我們發現解釋變數 x_3~x_8 是否對房價有非線性影響及其影響的趨勢皆會隨時間變化而有所不同，故此動態變化的特性顯示解釋變數 x_3~x_8 對房價的非線性效應在時間上具有局部特性，因此，對於不同時間的資料，皆需要重新審視此 6 個解釋變數的非線性效應。由於在這 6 個解釋變數中，哪一些解釋變數應該安排於 1 式的 LM 或 GAM，總共有 2^6 種可能的組合，故為了進一步刻劃最佳組合會隨著時間動態改變的特性，本研究考慮所有這 2^6 種可能組合所對應的候選模型，並透過模型選取準則 AIC 或 BIC，從中選擇出同時兼顧配適與預測能力的模型，以下列出本研究計算 AIC 與 BIC 的方式：

$$\text{AIC} = -2 \times \log L(\alpha_0, \alpha_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_j, \sigma^2) + 2 \times df,$$

$$\text{BIC} = -2 \times \log L(\alpha_0, \alpha_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_j, \sigma^2) + \log(n) \times df,$$

其中 n 為樣本數，

$$L(\alpha_0, \alpha_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_j, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left\{ \frac{\sum_{i=1}^{n} (y_i - \alpha_0 - f_1 - \sum_{j \in A} f_j - \sum_{K \in B^*} \alpha_k x_{ki})^2}{2\sigma^2} \right\}$$

為概似函數，$B^* = \{1, 2, 9, 10\} \cup \text{B}$ 且令 $K = \#\{B^*\} = 10 - J$ 為屬於 LM 部分的變數個數，其中 $J = \#\{A\}$ 為 x_3~x_8 中需要進行薄板樣條轉換的解釋變數個數，$df = \sum_{j \in \text{A}} \text{tr}(\mathbf{A}_j(\lambda)) + (K + 1) + 1$ 為模型的有效最大自由度，當中包含了對第 j 個薄板樣條轉換函數所對應的自由度 $\text{tr}(\mathbf{A}_j(\lambda))$、LM 部份（含截距項）的自由度 $K + 1$、以及估計 σ^2 的 1 個自由度。最後，上述 GCV 的定義與 $\mathbf{A}_j(\lambda)$ 的計算細節請參見附錄。

值得注意的是 1 式中的模型僅針對經緯度資料建立聯合非線性轉換函數，尚未廣泛考慮變數間的交互作用，然而，根據第 2.3 節的觀察發現，屋齡（x_5）與地點的交互作用可能也為影響高雄市房價的因素，故我們進一步考慮以下同時將 a_1、a_2、x_5 進行聯合非線性函數轉換以反應屋齡與地點交互作用的模型設定：

$$y_i = \alpha_0 + f_1(a_{1i}, a_{2i}; \boldsymbol{\theta}_1) + f_*(a_{1i}, a_{2i}, x_{5i}; \boldsymbol{\theta}_*) + \sum_{j \in \mathrm{A}} f_j(x_{ji}; \boldsymbol{\theta}_j) + \sum_{k \in \{1,2,9,10\} \cup \mathrm{B}} \alpha_k x_{ki} + \varepsilon_i \tag{3}$$

其參數估計可藉由與處理 1 式模型類似的程序得到，於此不再贅述。我們將在實證研究時討論 3 式中的交互作用效應是否可提升對房價的預測準確度。

4. 數值研究

為了評估 GAM 在房價配適與預測上的表現，本文同時引入 KNN 法對房價進行預測，並將其視為 GAM 預測結果的比較對象。KNN 為一種無母數分析方法，主要透過度量不同樣本特徵間的距離，以判斷未知樣本屬於哪一個已知類別或是與哪一些樣本較為接近。應用於本文所探討的問題時，我們將每一筆交易資料在表 2 中所列的解釋變數視為該筆交易房價的特徵向量，對任一欲預測房價的資料，則利用搜尋在訓練集資料中，與其特徵向量距離最接近的 k 筆資料所對應房價之加權平均數，作為該筆交易資料的房價預測值，相關計算細節如下：

首先將訓練集的資料記為 $(\mathbf{y}^{(0)}, \mathbf{X}^{(0)})$，其中 $\mathbf{y}^{(0)} = (y_1^{(0)}, y_2^{(0)}, \ldots, y_n^{(0)})^{\mathrm{T}}$ 為 n 筆房價資料所形成的向量，

$$\mathbf{X}^{(0)} = \begin{pmatrix} a_{1,1}^{(0)} & a_{1,2}^{(0)} & x_{1,1}^{(0)} & \cdots & x_{1,10}^{(0)} \\ a_{2,1}^{(0)} & a_{2,2}^{(0)} & x_{2,1}^{(0)} & \cdots & x_{2,10}^{(0)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,1}^{(0)} & a_{n,2}^{(0)} & x_{n,1}^{(0)} & \cdots & x_{n,10}^{(0)} \end{pmatrix}$$

為列舉於表 2 中的解釋變數相對於房價資料之觀測值所形成的矩陣，$\mathbf{X}^{(0)}$ 中的第 i 列表示由第 i 筆房價資料所對應的解釋變數觀測值所形成的向量。由於 $\mathbf{X}^{(0)}$ 中包含數值型與類別型的解釋變數，對於數值型的解釋變數，即表 2 中的 $a_1, a_2, x_1, x_2, x_3, x_5, x_6, x_7, x_8$ 等 9 個變數，我們對每一個解釋變數皆各自進行標準化以去除單位效應，如：對屋齡 x_5，將 $\mathbf{X}^{(0)}$ 中的觀測值向量 $\mathbf{x}_5^{(0)} = (x_{1,5}^{(0)}, \ldots, x_{n,5}^{(0)})^{\mathrm{T}}$ 轉換成 $\mathbf{x}_5 = (x_{1,5}, \ldots, x_{n,5})^{\mathrm{T}}$，其中 $x_{i,5} = (x_{i,5}^{(0)} - \hat{\mu}_5)/\hat{\sigma}_5$，$i = 1, \ldots, n$，$\hat{\mu}_5$ 為 $\mathbf{x}_5^{(0)}$ 的樣本平均數，$\hat{\sigma}_5$ 為 $\mathbf{x}_5^{(0)}$ 的樣本標準差。另一方面，對於表 2 中的 x_4、x_9、x_{10} 等 3 個類別型的解釋變數，我們則將其以虛擬變數（dummy variable）的型式表示，如：建物型

態 x_4 原本分為透天厝、公寓、華廈、住宅大樓等四個類別，我們將該四個類別分別以 $x_4^{dum} = (x_{4-1}, x_{4-2}, x_{4-3}, x_{4-4}) = (1,0,0,0)$、$(0,1,0,0)$、$(0,0,1,0)$、$(0,0,0,1)$ 表示。綜合以上對數值型與類別型變數的轉換，我們將轉換後的資料記為 $\mathbf{X} = (\boldsymbol{z_1}, \ldots, \boldsymbol{z_n})^{\mathrm{T}}$，其中第 i 筆資料為

$$\boldsymbol{z}_i = (a_{i,1}, a_{i,2}, x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}^{dum}, x_{i,5}, x_{i,6}, x_{i,7}, x_{i,8}, x_{i,9}^{dum}, x_{i,10}^{dum})^{\mathrm{T}}，i = 1, ..., n。$$

假設觀察到一筆經由上述處理後的特徵向量，記為 $\boldsymbol{z}^*$，首先計算在上述的 $\mathbf{X}$ 中，與 $\boldsymbol{z}^*$ 距離最接近的 k 個 $\boldsymbol{z}_i$ 向量，記為 $\boldsymbol{z}_{(1)}, ..., \boldsymbol{z}_{(k)}$，其中距離的計算方式如下：

$$d_i = c_1 d_{i,num} + c_2 d_{i,dum}，i = 1, ..., n \tag{4}$$

在 4 式中，$d_{i,num}$ 為 $\boldsymbol{z}_i$ 與 $\boldsymbol{z}^*$ 中 9 個經標準化後之數值型變數觀測值間的歐式距離，$d_{i,dum}$ 為 $\boldsymbol{z}_i$ 與 $\boldsymbol{z}^*$ 中 3 個經虛擬變數化後之類別型變數觀測值間的漢明距離（Hamming distance），c_1 與 c_2 分別為 $d_{i,num}$ 與 $d_{i,dum}$ 的權重。由於本研究對數值型變數採用標準化後，每一個變數的全距約為 6，因此，9 個數值型變數的最大可能的歐氏距離約為 18；另一方面，$\boldsymbol{z}_i$ 與 $\boldsymbol{z}^*$ 在 $x_{i,4}^{dum}, x_{i,9}^{dum}, x_{i,10}^{dum}$ 等 3 個變數間最大可能的漢明距離約為 6；在設定 c_1 與 c_2 時，我們為了同時考量數值型與類別型變數個數的貢獻（分別為 9 與 3），但同時亦要均衡數值型與類別型變數最大距離（分別為 18 與 6），故在定義 d_i 時，提出將 $d_{i,num}$ 與 $d_{i,dum}$ 分別除以 18 與 6 以去除最大距離的效應後，再分別乘以 9 與 3 以和變數個數成正比，因此可得 $c_1 : c_2 = (3/18) : (1/6) = 1 : 1$，故設定 $c_1 = c_2 = 0.5$。最後，以 $\boldsymbol{z}_{(1)}, ..., \boldsymbol{z}_{(k)}$ 所對應之房價（記為 $y_{(1)}, ..., y_{(k)}$）的加權平均值作為該筆資料的房價估計值，其中權重設定為正比於 $\boldsymbol{z}_{(i)}$ 與 $\boldsymbol{z}^*$ 距離 $d_{(i)}$ 的倒數，即 KNN 對 $\boldsymbol{z}^*$ 所對應之房價的預測值為 $\hat{y}_{KNN}^{(0)}(\boldsymbol{z}^*) = m_1 + m_2 \hat{y}_{KNN}(\boldsymbol{z}^*)$，其中 $\hat{y}_{KNN}(\boldsymbol{z}^*) = \sum_{j=1}^{k} y_{(i)}/d_{(i)}$，$m_1$、$m_2$ 則定義於 2 式。特別地，在本文所採用的移動視窗架構下，k 值的選取會隨著資料不同而改變，每一次的移動視窗 k 值均由 $\{1, \ldots, 20\}$ 中挑選，並以訓練集中的資料，選出使得 5 折交互驗證（5-fold cross-validation）之均方誤差（mean-square error，MSE）最小的 k 值，並依此 k 值應用下一個月的建物特徵向量以預測其所對應的房價。

4.1 模擬研究

本節首先透過模擬研究探討透過 GAM 是否能夠描繪出與真實資料相似的特性及行為。以 2014 年 4 月到 2014 年 6 月間之資料為例，採用表 2 所列的變數作為解釋變數觀測值，在移除房價大於（或小於）平均值 3 倍標準差的筆數後，共計有 3375 筆資料，模擬流程如下：

1. 在參數設定方面，本例設定解釋變數 x_3 至 x_8 皆具有非線性效應，並採用上述資料透過配適 1 式中的模型後，將所得到的薄板樣條函數 f_1、f_j，與係數估計值 $\alpha_0^{(0)}$、$\alpha_k^{(0)}$、$\boldsymbol{\theta}_1^{(0)}$、$\boldsymbol{\theta}_j^{(0)}$ 設為本模擬研究的參數設定值，同時得到殘差的標準差設定值 $\sigma_{(0)}$。

2. 應用上述步驟所得到的參數設定值、薄板樣條函數與原始資料的解釋變數，透過下式生成新的反應變數 y_i^*，$i = 1, \ldots, 3375$：

$$y_i^* = \alpha_0^{(0)} + f_1\left(a_{1i}, a_{2i}; \boldsymbol{\theta}_1^{(0)}\right) + \sum_{j\in\{3,4,5,6,7,8\}} f_j\left(x_{ji}; \boldsymbol{\theta}_j^{(0)}\right) + \sum_{k\in\{1,2,9,10\}} \alpha_k^{(0)} x_{ki} + e_i$$

其中誤差項 $\{e_i，i = 1, 2, ..., 3375\}$ 為一組自 $N\left(0, \sigma_{(0)}^2\right)$ 分佈獨立生成的隨機變數。圖 5~8 呈現模擬資料 y_i^* 與真實資料 y_i 分別對解釋變數 x_1~x_{10} 的散佈圖，由於所模擬出來的 y_i^* 與解釋變數間的散佈情況與實際資料的散佈情況相似，若以真實資料 y_i 對模擬資料 y_i^* 進行線性迴歸，其 R^2 約為 0.69，因此，我們得知透過 1 式的模型，經由代入上述的解釋變數後，可生成出與真實房價接近的模擬價格。

3. 由於步驟 2 所生成的 y_i^* 與解釋變數間的真實關係即為 1 式的模型，本步驟在 1 式的假設下，同樣將解釋變數 x_3 至 x_8 放進薄板樣條函數中，應用 R 'mgcv' package 於 y^* 與解釋變數的資料上，進行參數估計。最後經由圖 9 的殘差分位圖及殘差與估計值散佈圖進行殘差診斷，圖 9(a) 的殘差分位圖顯示無顯著證據拒絕殘差為常態分配的假設，圖 9(b) 的殘差與估計值散佈圖則顯示殘差不隨著估計值大小變化，因此無顯著證據拒絕解釋變數與誤差獨立的假設，故殘差診斷結果為模型配適良好。

在上述的模擬案例中，我們發現 GAM 能有效地模擬出實際資料中解釋變數影響房價的非線性趨勢。此外，應用 R 'mgcv' package 於 GAM 的參數估計時，亦在上述

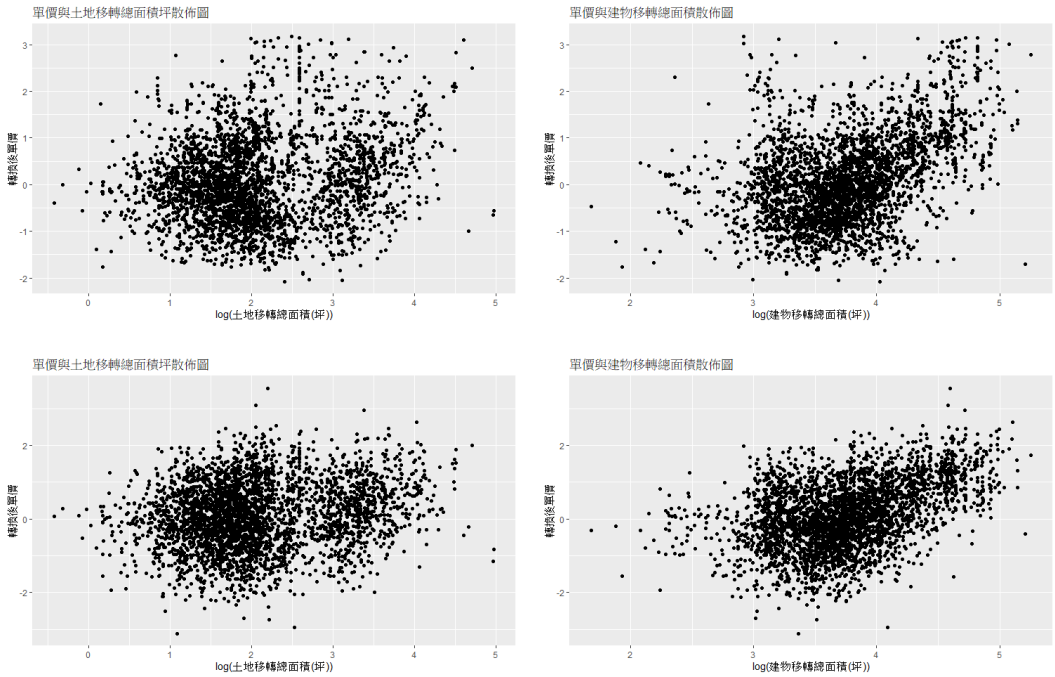


圖 5: 左側為轉換後房價與 log(土地移轉總面積)（即 x_1）之散佈圖，右側為轉換後房價與 log(建物移轉總面積)（即 x_2）之散佈圖，其中上排的 y 座標為真實資料，下排的 y 座標為模擬資料散佈圖。

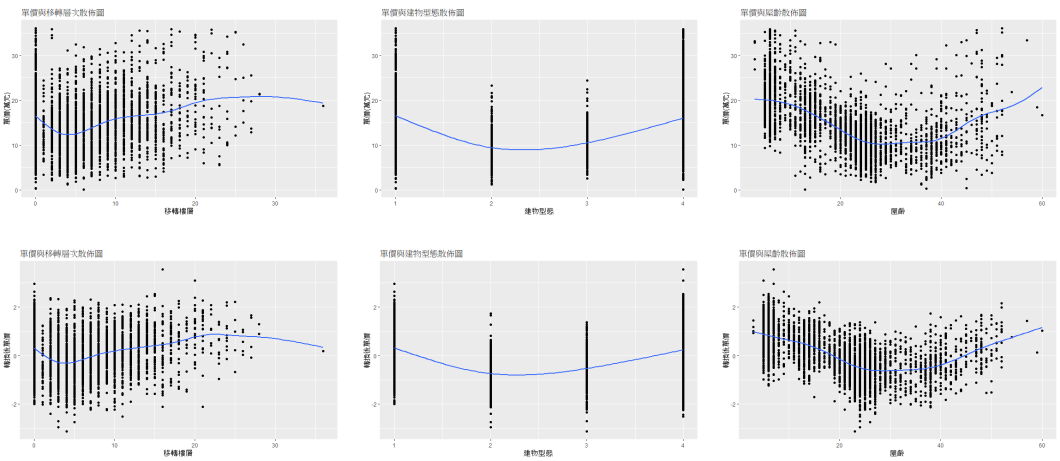


圖 6: 左側為轉換後房價與移轉樓層（x_3）之散佈圖，中間圖為轉換後房價與建物型態（x_4）之散佈圖，右側為轉換後房價與屋齡（x_5）之散佈圖，其中上排的 y 座標為真實資料，下排的 y 座標為模擬資料散佈圖。藍線代表各解釋變數對轉換後房價影響的趨勢線。

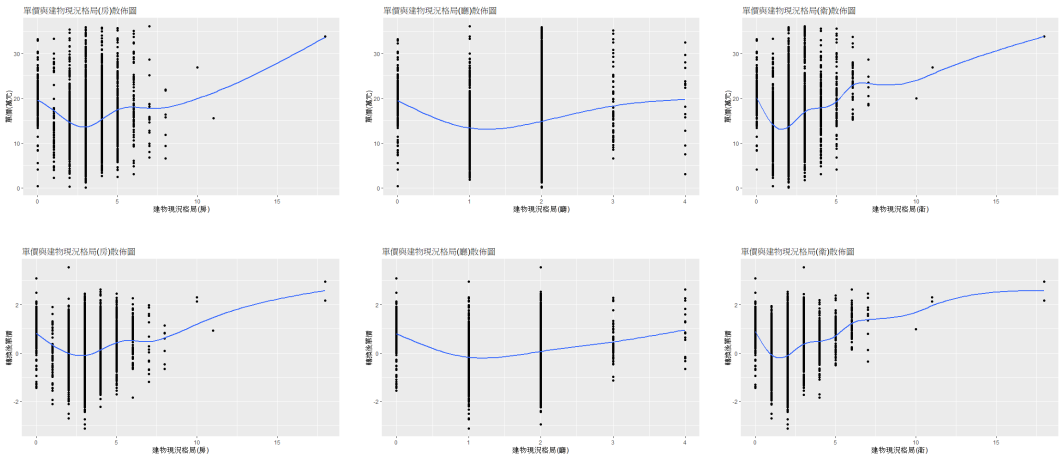


圖 7: 左側為轉換後房價與建物現況格局（房）（即 x_6）之散佈圖，中間圖為轉換後房價與建物現況格局（廳）（即 x_7）之散佈圖，右側為轉換後房價與建物現況格局（衛）（即 x_8）之散佈圖，其中上排的 y 座標為真實資料，下排的 y 座標為模擬資料散佈圖。藍線代表各解釋變數對轉換後房價影響的趨勢線。

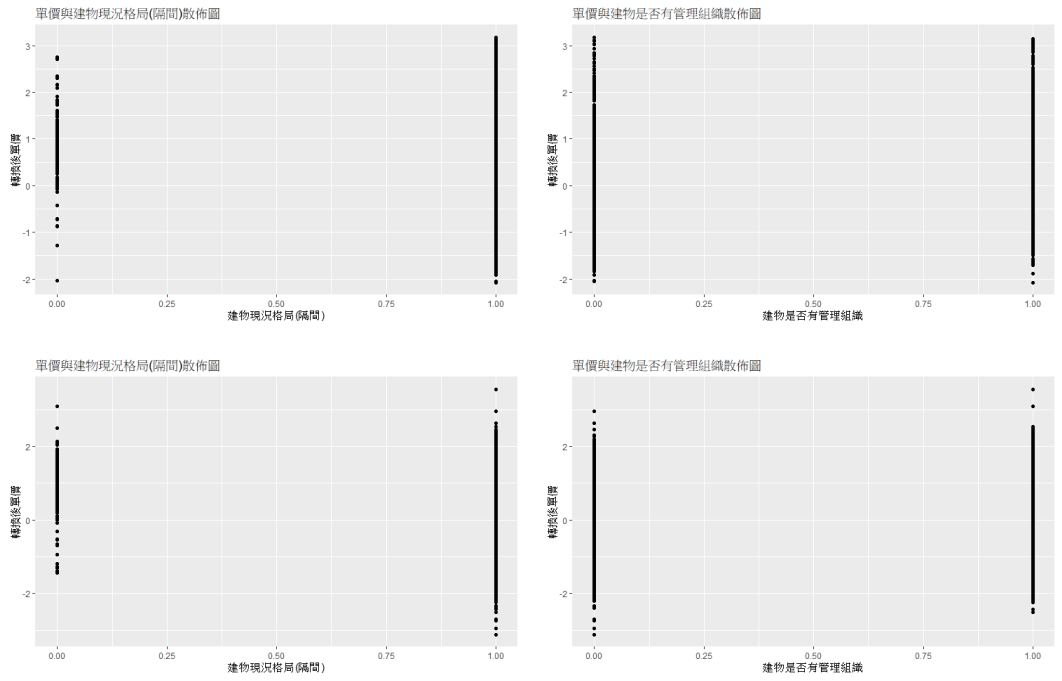
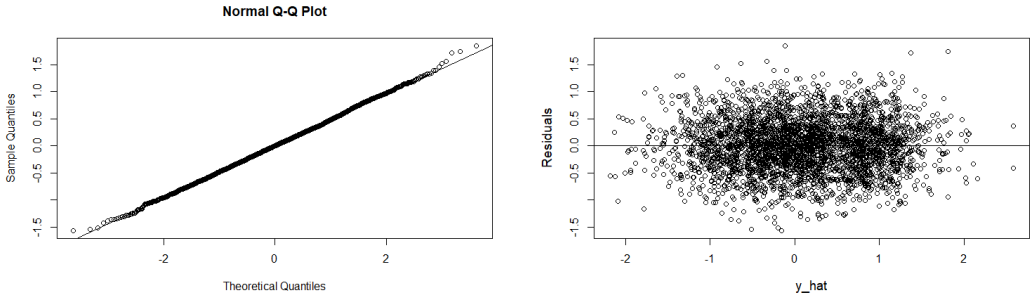


圖 8: 左側為轉換後房價與有無隔間（x_9），右側為轉換後房價與有無管理組織（x_{10}），其中上排的 y 座標為真實資料，下排的 y 座標為模擬資料散佈圖。



(a) (b)

圖 9: （a）殘差分位圖；（b）殘差與估計值散佈圖。

案例獲得良好的殘差診斷結果。本研究亦曾以不同時間區間的資料進行與上述案例相同的模擬實驗，所呈現的結果皆與上述案例非常類似，於此不再贅述。以下我們進一步探討應用所提出的方法於實務資料分析時的配適與預測表現。

4.2 實證分析

本研究在實證分析上採用高雄市 2014 年 1 月到 2019 年 12 月間的實價登錄資料，以每 w 個月的資料（w=3, 6, 12, 或 24），對房價以及在表 2 中所列舉的解釋變數，透過所提出的方法建立一 GAM，再依所建立的模型對下個月的房價進行預測。接著將時間平移 1 個月後再接續上述過程，直至得到 2019 年 12 月的房價預測結果為止，以探討所提出模型的配適與預測表現。與模擬研究中相同地，我們亦移除在每 w 個月中房價大於（或小於）平均值 3 倍標準差的相關資料，特別地，由於透天厝與移轉樓層為 1（可能為大樓店面）的建物價格型態較為不同，故以下的實證分析亦移除移轉層次小於 2 的資料。

在第 3 節我們提出透過 AIC 或 BIC 等選模準則決定 1 式中解釋變數 x_3~x_8 哪些應透過薄板樣條轉換函數以捕捉影響房價的非線性趨勢，進而提升模型解釋力與預測力。由於對此 6 個解釋變數而言，共有 2^6 種不同的組合，為了比較不同組合所對應之 GAM 的配適效果，本研究採用計算以下的均方根誤差（root-mean-square error，RMSE）做為評估配適效果的指標：

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(0)} - \hat{y}_i^{(0)}\right)^2}$$

表 3: 在 1 式的模型下，對 2014 年 4 月到 6 月的資料，前 10 個 AIC 值最小或 BIC 值最小之 GAM 的 RMSE。

集合 A	RMSE	AIC	集合 A	RMSE	AIC
3,4,5,6,7,8	2.692	2553.358	4,5,8	2.718	2886.100
4,5,6,7,8	2.696	2553.902	5,8	2.722	2888.142
3,5,6,7,8	2.696	2560.179	4,5,7,8	2.707	2888.541
5,6,7,8	2.700	2560.441	5,7,8	2.711	2890.640
3,4,5,6,8	2.699	2561.557	4,5,6,8	2.704	2898.339
4,5,6,8	2.704	2563.152	5,6,8	2.708	2900.797
3,4,5,7,8	2.702	2564.236	4,5,6,7,8	2.696	2903.271
4,5,7,8	2.707	2564.884	5,6,7,8	2.700	2906.305
3,5,6,8	2.703	2568.386	3,4,5,8	2.712	2906.821
5,6,8	2.708	2569.665	3,5,8	2.717	2908.736

其中 $y_i^{(0)}$ 代表第 i 個房價的真實值，$\hat{y}_i^{(0)} = m_1 + m_2\hat{y}_i$ 則為第 i 個房價的估計值，$\hat{y}_i$ 為透過 1 式得到的估計值，m_1、m_2 則定義於 2 式。若某一種組合的 RMSE 越小，即代表該組合所對應之 GAM 對房價的配適值與真實值越接近。

以 2014 年 4 月到 6 月的資料為例，表 3 列出在 1 式的模型下，分別根據 AIC 或 BIC 所挑選出來的前 10 種組合之 RMSE。經比較 2 個表格的數值結果後可發現，AIC 值最小的 GAM 建議將 x_3~x_8 等解釋變數皆進行薄板樣條函數轉換；但 BIC 值最小的 GAM 則建議將 x_4、x_5、x_8 等解釋變數進行薄板樣條函數轉換。由此可見，不同的選模準則可能會建議不同的 GAM。經由比較上述經由 AIC 或 BIC 建議的 GAM 之 RMSE 後，對這三個月的資料而言，表 3 中所有模型的對房價配適的 RMSE 皆約為 2.7 萬元左右，但 AIC 所建議的 GAM 較 BIC 所建議的 GAM 具有略低的 RMSE，故其配適效果稍佳。

此外，圖 10 呈現 KNN 模型和透過 AIC 或 BIC 所建議之 1 式與 3 式模型下的 GAM 在移動視窗設定為 3 個月下的 RMSE 熱圖，顏色越淡代表 RMSE 越小，其中 GAM.A1 代表在 1 式模型下透過 AIC 所建議的結果，GAM.B2 代表在 3 式模型下透過 BIC 所建議的結果，依此類推。由圖 10 可發現，四個 GAM 模型的配適效果皆明顯地優於 KNN 模型，特別地，GAM.A2 與 GAM.B2 的配適效果又略優於 GAM.A1 與 GAM.B1。圖 11 呈現本文所提出之 KNN 建模時，在移動視窗 $w = 3$ 個月與 5 折交互驗證的架構下，69 期資料 k 值的時間序列圖與直方圖，由圖中可見，k 值落於 5

至 12 間，平均約為 7.7，且 k 值的選取的確會依資料不同而動態調整。

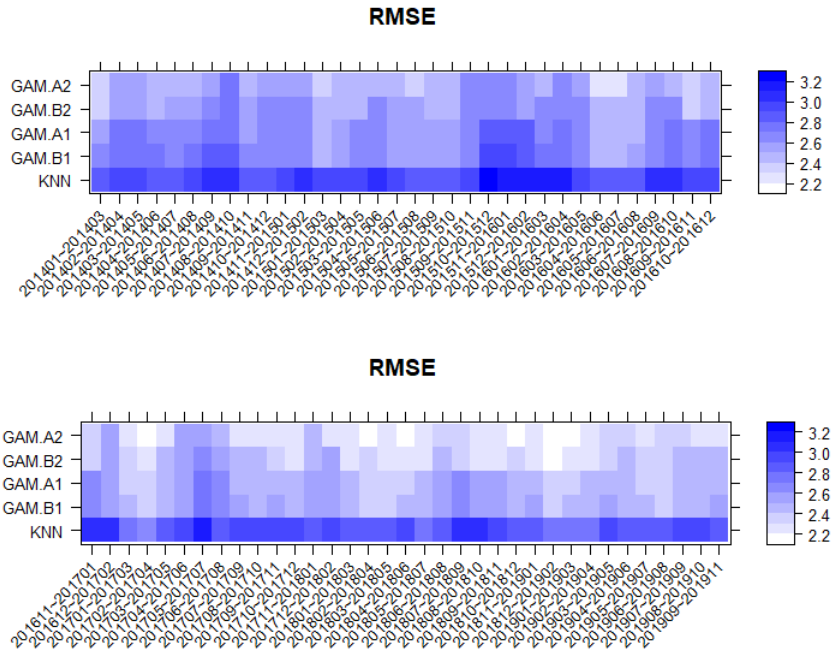
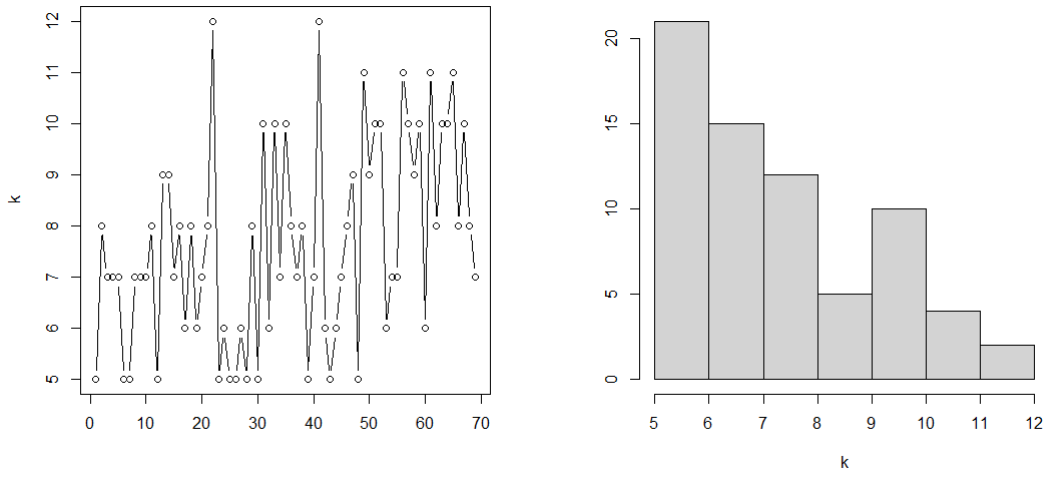


圖 10: KNN、GAM.A1、GAM.B1、GAM.A2、與 GAM.B2 在移動視窗為 3 個月時的 RMSE 熱圖。



(a) (b)

圖 11: KNN 模型在移動視窗為 3 個月時，k 值的（a）時間序列圖與（b）直方圖。

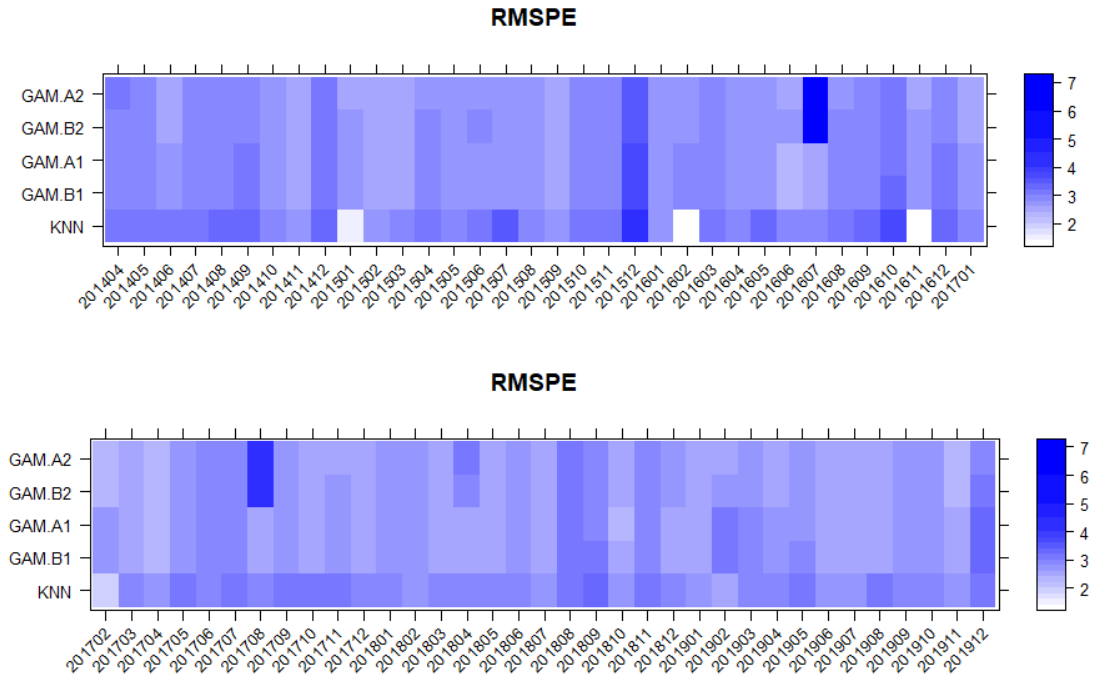


圖 12: KNN、GAM.A1、GAM.B1、GAM.A2、與 GAM.B2 在移動視窗為 3 個月時的 RMSPE 熱圖。

接下來，為了評估不同準則所建議之模型的預測效果，本研究採用計算以下的均方根預測誤差（root-mean-square prediction error，RMSPE）做為評估預測效果的指標：

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(0)} - \tilde{y}_i^{(0)}\right)^2}$$

其中 $y_i^{(0)}$ 代表第 i 個房價的真實值，$\tilde{y}_i^{(0)} = m_1 + m_2\tilde{y}_i$ 則為第 i 個房價的預測值，$\tilde{y}_i$ 為利用前 w 個月的資料（w=3,6,12, 或 24）所建立之模型的預測值，m_1、m_2 則定義於 2 式。若某一模型的 RMSPE 越小，即代表該模型對房價的預測效果越好。圖 12 呈現 KNN、GAM.A1、GAM.B1、GAM.A2、GAM.B2 在移動視窗為 3 個月下的 RMSPE 熱圖，顏色越淡代表 RMSPE 越小。由圖 12 可發現，在 69 期的預測月份中，本文所提出的 KNN 法在其中 3 個月預測表現非常優異，但在大多數的情況下，四種 GAM 模型的 RMSPE 皆優於 KNN 模型。

接下來，為了進一步比較 KNN 法和 GAM.A1、GAM.B1、GAM.A2、GAM.B2

表 4: 移動視窗為 3、6、或 12 個月時，不同模型與 KNN3 法在四種評估指標上，其「RMSE、RMSPE 的比例值小於 1」以及「N3、N5 的比例值大於 1」的百分比，其中比例值大於 0.90 者以粗體表示。

		Model				
	Window size	KNN	GAM.A1	GAM.B1	GAM.A2	GAM.B2
$\frac{\text{RMSE(Model)}}{\text{RMSE(KNN3)}} < 1$	3	0.00	**1.00**	**1.00**	**1.00**	**1.00**
	6	**0.98**	**1.00**	**1.00**	**1.00**	**1.00**
	12	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	24	**1.00**	**0.98**	**0.96**	**1.00**	**1.00**
$\frac{\text{RMSPE(Model)}}{\text{RMSPE(KNN3)}} < 1$	3	0.00	0.88	0.87	0.87	0.86
	6	0.89	0.89	0.89	0.88	0.89
	12	**0.90**	0.88	0.88	**0.92**	**0.90**
	24	**0.90**	0.88	0.88	**0.92**	**0.92**
$\frac{\text{N3(Model)}}{\text{N3(KNN3)}} > 1$	3	0.00	0.88	**0.90**	**0.93**	**0.93**
	6	0.89	0.89	0.89	**0.92**	**0.91**
	12	**0.92**	0.87	0.87	**0.92**	**0.92**
	24	**0.92**	**0.90**	0.88	**0.90**	**0.90**
$\frac{\text{N5(Model)}}{\text{N5(KNN3)}} > 1$	3	0.00	0.88	**0.90**	**0.93**	**0.93**
	6	**1.00**	**0.97**	**0.97**	**0.97**	**0.97**
	12	**0.92**	0.87	0.87	**0.92**	**0.92**
	24	**0.92**	**0.90**	0.88	**0.90**	**0.90**

等模型在不同移動視窗設定下的表現，除了上述提及的 RMSE 與 RMSPE 以外，我們亦考慮計算每坪房價預測誤差小於 3 萬元與小於 5 萬元之比例，分別以 N3 與 N5 表示。表 4 呈現移動視窗設定為 3、6、12、或 24 個月時，不同模型與移動視窗為 3 個月時的 KNN 法（記為 KNN3）在上述四個評估指標上，其「RMSE、RMSPE 的比例值小於 1」以及「N3 與 N5 的比例值大於 1」的百分比，上述四個百分比越高，即代表該模型優於 KNN3 法的程度越高。首先，由表中 KNN 的欄位可發現，上述四項百分比的數值大多高於 9 成，因此，對 KNN 方法而言，移動視窗設定為 6、12、或 24 個月時的表現優於移動視窗設定為 3 個月時的表現。另一方面，在 KNN 與 GAM 模型的比較上，我們也發現所考慮的四個 GAM 模型在 RMSE、RMSPE、N3、與 N5 等四個評估指標與四種不同移動視窗大小的考量下，皆有高於 8 成以上的比例優於 KNN3 法，較為特別的是 GAM.A2 與 GAM.B2 的表現似乎更優於 GAM.A1 與

表 5: 移動視窗為 3、6、12、或 24 個月時，不同模型的 RMSE、RMSPE、N3、與 N5 之成對樣本 t 檢定的 p 值，其中 μ_{A1} 表示 GAM.A1 模型之 RMSE、RMSPE、N3、或 N5 的母體期望值，依此類推，且 p 值小於 0.05 者以粗體表示。

		Window size	3	6	12	24
(I)	$H_0 : \mu_{A1} = \mu_{B1}$	RMSE	**2.20E-06**	**2.20E-06**	**1.10E-15**	**6.00E-16**
	$H_1 : \mu_{A1} < \mu_{B1}$	RMSPE	**0.0027**	**0.0006**	**0.0004**	**0.015**
	$H_0 : \mu_{A1} = \mu_{B1}$	N3	0.5694	**0.0006**	**0.0122**	0.5221
	$H_1 : \mu_{A1} > \mu_{B1}$	N5	0.5694	**0.0082**	**0.0122**	0.5221
(II)	$H_0 : \mu_{A2} = \mu_{B2}$	RMSE	**1.60E-15**	**1.20E-11**	**1.70E-13**	**1.30E-09**
	$H_1 : \mu_{A2} < \mu_{B2}$	RMSPE	0.5201	0.1157	**0.0049**	**0.0208**
	$H_0 : \mu_{A2} = \mu_{B2}$	N3	**0.0121**	**0.0014**	0.3778	0.1515
	$H_1 : \mu_{A2} > \mu_{B2}$	N5	**0.0121**	0.0671	0.3778	0.1515
(III)	$H_0 : \mu_{A1} = \mu_{A2}$	RMSE	**2.20E-16**	**2.20E-16**	**2.20E-16**	**2.20E-16**
	$H_1 : \mu_{A2} < \mu_{A1}$	RMSPE	0.7732	**2.50E-05**	**8.80E-13**	**1.70E-11**
	$H_0 : \mu_{A1} = \mu_{A2}$	N3	**2.00E-08**	**1.10E-07**	**9.10E-10**	**6.30E-10**
	$H_1 : \mu_{A2} > \mu_{A1}$	N5	**2.00E-08**	**4.40E-05**	**9.10E-10**	**6.30E-10**
(IV)	$H_0 : \mu_{B1} = \mu_{B2}$	RMSE	**3.00E-16**	**2.20E-16**	**2.20E-16**	**2.20E-16**
	$H_1 : \mu_{B2} < \mu_{B1}$	RMSPE	0.7973	**1.60E-06**	**1.20E-12**	**4.60E-11**
	$H_0 : \mu_{B1} = \mu_{B2}$	N3	**6.20E-06**	**5.00E-08**	**7.90E-11**	**1.90E-09**
	$H_1 : \mu_{B2} > \mu_{B1}$	N5	**6.20E-06**	**0.0001**	**7.90E-11**	**1.90E-09**
(V)	$H_0 : \mu_{A2} = \mu_{KNN}$	RMSE	**2.20E-16**	**2.20E-16**	**2.20E-16**	**3.60E-15**
	$H_1 : \mu_{A2} < \mu_{KNN}$	RMSPE	0.2140	**8.80E-07**	**5.30E-05**	**0.0164**
	$H_0 : \mu_{A2} = \mu_{KNN}$	N3	**1.50E-10**	**8.60E-11**	**1.00E-05**	**0.0375**
	$H_1 : \mu_{A2} > \mu_{KNN}$	N5	**1.50E-10**	**7.00E-06**	**1.00E-05**	**0.0375**

GAM.B1，且 GAM.A2 的表現又略優於 GAM.B2。此外，對 GAM.A2 與 GAM.B2 而言，當移動視窗設定為 24 個月時，「N3、N5 的比例值大於 1」的比重卻低於移動視窗設定為 6 個月或 12 個月時的表現，透露出該兩個方法或許移動視窗較適合設定為 6 個月或 12 個月。

為了更深入討論由表 4 所觀察到的現象，我們針對（I）GAM.A1 vs. GAM.B1；（II）GAM.A2 vs. GAM.B2；（III）GAM.A1 vs. GAM.A2；（IV）GAM.B1 vs. GAM.B2；（V）GAM.A2 vs. KNN 等 5 個情況，在 RMSE、RMSPE、N3、與 N5 等四個評估指標與四種不同的移動視窗設定下，分別採用成對樣本 t 檢定（paired sample t-test）評估任意兩個模型間的差異是否顯著，檢定結果的 p 值呈現於表 5。由表 5

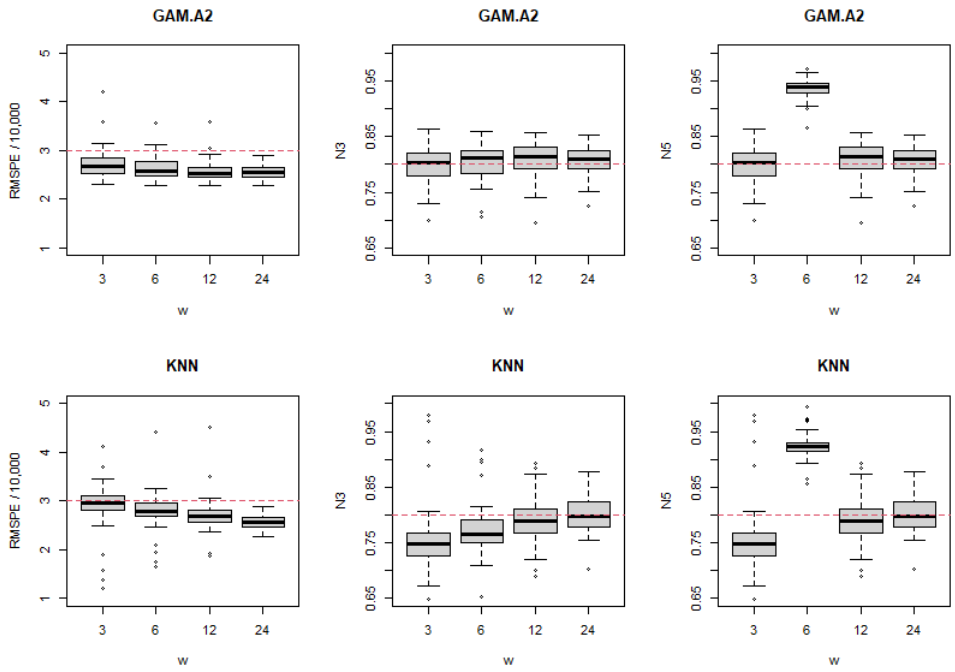


圖 13: 在不同移動視窗下，KNN 法與 GAM.A2 的 RMSPE、N3 與 N5 的盒形圖。

（I）~（IV）的結果可推論出 GAM.A2 與 GAM.B2 為四種 GAM 模型中表現較好的模型，其中 GAM.A2 又略優於 GAM.B2。此外，表 5 中情況（V）的檢定結果也呈現 GAM.A2 的表現優於 KNN 法，圖 13 進一步呈現 KNN 法與 GAM.A2 在不同移動視窗下，RMSPE、N3 與 N5 等三種評估預測表現之指標的盒形圖，由圖中可以發現，GAM.A2 在四種不同視窗的設定下，其 RMSPE 大多低於 3 萬元，且其 N3 與 N5 亦大多具有接近或高於 8 成的穩定表現，三項評估預測表現指標的結果皆優於 KNN 法，尤其當移動視窗設定為 6 或 12 個月時，GAM.A2 的表現更為優異，較為特殊的是評估指標 N5 在移動視窗設定為 6 個月時，表現特別突出。

根據表 4 與表 5 所呈現的結果，我們進一步討論 GAM.A2 在不同移動視窗設定下的表現，並於表 6 中呈現（I）$w = 3$ vs. $w = 6$；（II）$w = 6$ vs. $w = 12$；（III）$w = 6$ vs. $w = 12$；（IV）$w = 6$ vs. $w = 24$；（V）$w = 12$ vs. $w = 24$ 等情況下，成對樣本 t 檢定在四個評估指標下的 p 值。由表 6 的 RMSE 指標可發現移動視窗設定較小時，配適效果較好；然而在預測效果的評估上（即 RMSPE、N3、與 N5 等指標），情境（I）與（II）的檢定結果指出移動視窗設為 6 或 12 個月時，GAM.A2 的預測表現優於移動視窗設為 3 個月的表現；情境（III）的檢定結果指出移動視窗設為 12 個月時，RMSPE 顯著優於移動視窗設為 6 個月時的表現，但 N5 則反過來變成移動視窗

表 6: GAM.A2 模型在不同移動視窗下，RMSE、RMSPE、N3、與 N5 的之成對樣本 t 檢定的 p 值，其中 μ_w 表示移動視窗為 w 個月時，GAM.A2 模型之 RMSE、RMSPE、N3、或 N5 的母體期望值，p 值小於 0.05 者以粗體表示。

GAM.A2	(I)	(II)	(III)	(IV)	(V)
	$H_0:\mu_6=\mu_3$ $H_1:\mu_6<\mu_3$	$H_0:\mu_{12}=\mu_3$ $H_1:\mu_{12}<\mu_3$	$H_0:\mu_{12}=\mu_6$ $H_1:\mu_{12}<\mu_6$	$H_0:\mu_{24}=\mu_6$ $H_1:\mu_{24}<\mu_6$	$H_0:\mu_{12}=\mu_{24}$ $H_1:\mu_{12}<\mu_{24}$
RMSE	0.9929	1.0000	1.0000	1.0000	**2.20E-16**
RMSPE	0.0680	**0.0376**	**0.0031**	0.0555	**0.0145**
	$H_0:\mu_6=\mu_3$ $H_1:\mu_6>\mu_3$	$H_0:\mu_{12}=\mu_3$ $H_1:\mu_{12}>\mu_3$	$H_0:\mu_{12}=\mu_6$ $H_1:\mu_{12}>\mu_6$	$H_0:\mu_{24}=\mu_6$ $H_1:\mu_{24}>\mu_6$	$H_0:\mu_{12}=\mu_{24}$ $H_1:\mu_{12}>\mu_{24}$
N3	**0.0004**	**0.0004**	0.1352	0.8949	**0.0001**
N5	**2.20E-16**	**0.0004**	1.0000	1.0000	**0.0001**

設為 6 個月時的表現顯著優於移動視窗設定為 12 個月時結果（此一現象可與圖 13 中位於右上角的子圖之發現相呼應），另外，指標 N3 的檢定結果雖不顯著，但移動視窗設為 12 個月時的表現略優，故整體的預測效果以移動視窗設定為 12 個月時的表現略優於移動視窗設定為 6 個月時的結果；情境（IV）的判讀方式與情境三類似，可得出移動視窗設為 6 個月時的表現略優於移動視窗設為 24 個月的結果；情境（V）的檢定結果則顯示移動視窗設為 12 個月時無論在配適與預測的表現，皆顯著優於移動視窗設為 24 個月的結果。此外，本文所提出方法的計算時間與移動視窗大小成正比，亦即移動視窗設為 24 個月時所需的計算時間為移動視窗設為 3 個月時的 8 倍。綜合以上發現，我們建議在實務上使用本研究所提出的方法預測房價時，若允許較多的計算時間，可將移動視窗大小設定為 12 個月，若需要在短一點的時間內得到結果，則建議將移動視窗大小設定為 6 個月。

5. 結論與討論

本研究針對政府公開之高雄市不動產實價登錄資料進行分析及預測，透過模擬實驗發現所提出的模型能生成出與真實房價接近的模擬價格，實證分析結果則顯示透過所提出方法所建議的 GAM 無論在配適或是預測方面的表現，都優於 KNN 法，尤其是在 3 式的模型中引入建物地點與屋齡的交互作用後，提升效果更為明顯。

雖然本研究嘗試結合 GAM、選模準則、與實價登錄網上的資料，對高雄市房價進行建模與預測時可得到不錯的效果，但誠如第 3 節所提及，本研究皆採用薄板樣條函

數擷取解釋變數影響房價的非線性趨勢，未來可進一步探討採用不同非線性轉換函數或是採用不同之結點設定對 GAM 配適與預測表現的影響。此外，本研究在 1 式的模型中僅考慮對經緯度使用 2 維薄板樣條函數進行非線性轉換，在 3 式的模型中再透過配適一 3 維薄板樣條函數以描述建物地點之經緯度與屋齡的交互作用，其餘解釋變數的交互作用項對房價是否也存在非線性效應？若存在應如何設計於模型中？若考慮交互作用項勢必使得候選模型個數大量增加，此時應如何有效挑選出適合的模型？皆為未來值得深入研究的課題。

另一方面，實務上亦可直接套用其他機器學習方法（如：Random Forest、SVR、XGB、LGBM⋯等）或是深度學習模型（如：RNN、LSTM⋯等）對房價進行預測，然本文主要呈現透過 GAM 可以建立一個具有競爭力且易於解釋的房價預測模型，例如：圖 2~4 即呈現房價與其他解釋變數間的非線性相關，能讓我們更清楚了解房價與解釋變數間的關聯。此外，本文採用 KNN 法作為比較對象，主要是因為 KNN 法實為一般在交易房產時，人們較為常用的比價方式，也就是找鄰近性質相近的近期成交案進行推估。本文的實證研究結果呈現 GAM 除了具上述所提之解釋能力外，亦具有較 KNN 法更好的預測能力。未來我們將嘗試應用機器學習或深度學習模型優異的分類能力，並根據其分類結果，為每一類別再分別建立 GAM，以便兼顧預測力與解釋力。然而，此一研究方向是否可顯著地提升房價預測能力，仍待未來更進一步的探討。

附錄

以下簡述第 3 節所提到之 $\mathbf{A}(\lambda)$ 的計算細節與 GCV 的定義。令 d 代表與反應變數進行薄板樣條轉換的解釋變數個數，當 $d = 1$ 時，Wood and Augustin（2002）提出以下的薄板樣條轉換方式：f 由 $m + 2$ 個基底所形成，則 GAM 為

$$y = f(x) + \varepsilon, \tag{5}$$

其中 $f(x) = \sum_{j=1}^{m+2} \alpha_j b_j(x)$，$b_j(x)$ 為基底函數且滿足 $b_j(x) = |x - x_j^*|^3$，$j = 1, ..., m$、$b_{m+1}(x) = 1$、$b_{m+2}(x) = x$，α_j 為相對應的係數，x_j^*，$j = 1, ..., m$ 為事先設定的內點。

假設 (x_i, y_i)，$i = 1, \ldots, n$ 為 n 筆觀測到的資料，定義 $\mathbf{y} = (y_1, y_2, ..., y_n)^{\mathrm{T}}$,

$$\mathbf{X} = \begin{bmatrix} b_1(x_1) & \ldots & b_{m+2}(x_1) \\ b_1(x_2) & \ldots & b_{m+2}(x_2) \\ b_1(x_3) & \ldots & b_{m+2}(x_3) \\ & \vdots & \\ b_1(x_n) & \ldots & b_{m+2}(x_n) \end{bmatrix} = \begin{bmatrix} \mathbf{b}(x_1)^{\mathrm{T}} \\ \mathbf{b}(x_2)^{\mathrm{T}} \\ \mathbf{b}(x_3)^{\mathrm{T}} \\ \vdots \\ \mathbf{b}(x_n)^{\mathrm{T}} \end{bmatrix}, \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_{m+2} \end{bmatrix} \text{和 } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

則 5 式的矩陣表達式為 $\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$。Wood and Augustin（2002）提出透過以下的準則對 $\boldsymbol{\alpha}$ 進行估計：

$$\hat{\boldsymbol{\alpha}}(\lambda) = \arg_{\alpha}\{\min||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2 + \lambda J(f)\} \tag{6}$$

其中 λ 為懲罰係數，$J(f) = \int (f''(x_i))^2 \, dx_i$ 為懲罰項。由於 $f''(x_i) = \sum_{j=1}^{m} \alpha_j b_j''(x_i) = \mathbf{b}''(x_i)^{\mathrm{T}}\boldsymbol{\alpha}$，故可得 $J(f) = \int \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{b}''(x_i)\mathbf{b}''(x_i)^{\mathrm{T}}\boldsymbol{\alpha} dx_i = \boldsymbol{\alpha}^{\mathrm{T}} \int \mathbf{S}(x_i) dx_i \boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha}$，因此可將 6 式右側改寫成

$$\min||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2 + \lambda\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{H}\boldsymbol{\alpha} = \min \boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{H})\boldsymbol{\alpha} - \mathbf{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{X}^{\mathrm{T}} \tag{7}$$

透過 7 式可解出 $\hat{\boldsymbol{\alpha}}(\lambda) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{H})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$。令 $\mathbf{A}(\lambda) = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{H})^{-1}\mathbf{X}^{\mathrm{T}}$，則可得 $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\alpha}}(\lambda) = \mathbf{A}(\lambda)\mathbf{y}$。其餘細節（包含如何確保 $\hat{\boldsymbol{\alpha}}(\lambda)$ 的可識別性以及推廣到 $d \geq 2$ 時的相關推導過程）請參閱 Wood and Augustin（2002），於此不再贅述。最後，懲罰係數 λ 的選取則是透過最小化以下的 GCV（記為 $V(\lambda)$）決定：

$$V(\lambda) = \frac{n||\mathbf{y} - \mathbf{A}(\lambda)\mathbf{y}||^2}{[n - \mathrm{tr}(\mathbf{A}(\lambda))]^2}$$

其中 $\mathrm{tr}(\mathbf{A}(\lambda))$ 稱為模型有效的估計自由度。

參考文獻

[1] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician*, 46(3), pages 175-185.

[2] Cunningham, P., and Delany, S. J. (2021). k-Nearest neighbour classifiers-A Tutorial. *ACM Computing Surveys (CSUR)*, 54(6), pages 1-25.

[3] Dbrowski, J., and Adamczyk, T. (2010). Application of GAM additive non-linear models to estimate real estate market value. *Geomatics and Environmental Engineering*, 4(2), pages 55-62.

[4] De Souza, J. B., Reisen, V. A., Franco, G. C., Ispány, M., Bondon, P., and Santos, J. M. (2018). Generalized additive models with principal component analysis: an application to time series of respiratory disease and air pollution data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(2), pages 453-480.

[5] Dominici, F., McDermott, A., Zeger, S. L. and Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, 156(3), pages 193-203.

[6] Goodman, A. C. (1998). Andrew Court and the invention of hedonic price analysis. *Journal of Urban Economics*, 44(2), pages 291-298.

[7] Hastie, T. J., and Tibshirani, R. J. (1986). Generalized additive models. *Statistical Science*, 1(3), pages 297-310.

[8] Janssen, C., Söderberg, B., and Zhou, J. (2001). Robust estimation of hedonic models of price and income for investment property. *Journal of Property Investment & Finance*, 19(4), pages 342-360.

[9] Kim, K., and Park, J. (2005). Segmentation of the housing market and its determinants: Seoul and its neighbouring new towns in Korea. *Australian Geographer*, 36(2), pages 221-232.

[10] Kuşan, H., Aytekin, O. and Özdemir, İ. (2010). The use of fuzzy logic in predicting house selling price. *Expert systems with Applications*, 37(3), pages 1808-1813.

[11] Lowe, D. G. (1995). Similarity metric learning for a variable-kernel classifier. *Neural Computation*, 7(1), pages 72-85.

[12] Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In: O'Sullivan T., and Gibb K. Housing Economics and Public Policy. John Wiley & Sons. pages 67-89.

[13] Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., and French, N. (2003). Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance*, 21(4), pages 383-401.

[14] Schulz, R., and Werwatz, A. (2004). A state space model for Berlin house prices: Estimation and economic interpretation. *The Journal of Real Estate Finance and Economics*, 28(1), pages 37-57.

[15] Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, 36(2), pages 2843-2852.

[16] Simpson, G. L. (2018). Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution*, 6.

[17] Sirmans, S., Macpherson, D., and Zietz, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, 13(1), pages 3-43.

[18] Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), pages 95-114.

[19] Wood, S. N., and Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157(2-3), pages 157-177.

[20] Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1), pages 139-155.

[21] Wu, Z., and Zhang, S. (2019). Study on the spatial–temporal change characteristics and influence factors of fog and haze pollution based on GAM. *Neural Computing and Applications*, 31(2), pages 1619-1631.

[22] Yang, L., Qin, G., Zhao, N., Wang, C., and Song, G. (2012). Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. *BMC Medical Research Methodology*, 12, pages 1-13.

[23] Zhang, Z., Tan, S., and Tang, W. (2015). A GIS-based spatial analysis of housing price and road density in proximity to urban lakes in Wuhan City, China. *Chinese Geographical Science*, 25(6), pages 775-790.

[Received January 2022; accepted April 2022.]

Housing Price Forecasting by Generalized Additive Models

Shih-Feng Huang[1,2†] **and Yu-Ping Liao**[2]

[1]Department of Applied Mathematics, National University of Kaohsiung

[2]Institute of Statistics, National University of Kaohsiung

ABSTRACT

This study proposes a systematic approach based on the generalized additive model (GAM) and the real price registration website data to predict the housing prices in Kaohsiung from 2014 to 2019. The data include numeric and categorical covariates such as house location, area, type, etc. We employ the GAM to capture the nonlinear effects of covariates on housing prices, where the AIC (or BIC) is used to determine which covariates need to be transformed nonlinearly to improve the fitting and prediction performances. In our empirical study, a rolling window approach with a window size of 3, 6, 12, or 24 months and a one-month adaption frequency is employed to investigate the performance of the proposed method. In particular, we adopt the housing prices estimated by a machine learning method, the K nearest neighbor (KNN), as a comparison benchmark. The numerical results reveal that the proposed GAM has better fitting and predicting performances for Kaohsiung housing prices than the KNN method.

Key words and phrases: generalized additive models, real price registration, thin plate spline.

JEL classification: C53, R32.

[†]Corresponding to: Shih-Feng Huang
E-mail: huangsf@nuk.edu.tw