住宅學報　第十六卷第一期　民國九十六年六月　學術論著　第1頁—20頁
JOURNAL OF HOUSING STUDIES, VOLUME 16 NO. 1, JUNE 2007

學術論著

以決策樹之迴歸樹建構住宅價格模型—台灣地區之實證分析

Regression Trees for Housing Price Models: An Empirical Study on Taiwan

陳樹衡*　郭子文**　棗厥庸***

Shu-Heng Chen, Tzu-Wen Kuo, Chueh-Yung Tsao

摘　要

以往房地產特徵方程式之估計多採用複迴歸模型，後期學者開始使用半參數或無母數方法估計特徵方程式，本研究以決策樹中的Cubist迴歸樹作爲房地產特徵方程式之估計模型，主要原因有三：其一，Cubist迴歸樹模型之設計符合房地產資料特性。其二，Cubist迴歸樹配適能力高且易於解釋。其三，當使用大量資料估計特徵方程式，Cubist迴歸樹相對於其他無母數方法運算上較有效率。本研究以台灣地區2002年至2004年間45,419筆房地產資料爲研究樣本，以複迴歸模型爲基準模型，研究發現迴歸樹之配適能力高於複迴歸模型，且並未有過度配適之問題。此外，特徵變數與房地產價格間存有非線性關係，個體變數較總體變數具有廣泛之解釋力。

關鍵詞：房價、特徵方程式、決策樹、Cubist迴歸樹

Abstract

The purpose of this paper is to use Cubist regression trees to estimate the hedonic equation, as the Cubist is expected to be more efficient than other nonparametric methods. In addition, the architecture of the Cubist is intuitive when applied to the housing price model. In this study, the regression method, which is frequently used in the estimation of the hedonic equation, is used as the benchmark model to be compared with. Based on 45,419 observations from the Taiwan area, it is found that the Cubist outperforms the benchmark model. Moreover, it is found that there are nonlinear relationships between the house prices and the characteristic variables. Finally, the micro characteristics exhibit higher explanatory power than the macro ones.

Key words: house prices, hedonic equation, decision trees, Cubist regression trees

(本文於2007年4月2日收稿，2007年5月2日審查通過，實際出版日期2007年10月)

*　政治大學經濟學系教授。Professor, Department of Economics, National Chengchi University.
**　眞理大學財務金融學系助理教授。Assistant Professor, Department of Finance and Banking, Aletheia University.
***　長庚大學工商管理學系助理教授。Assistant Professor, Department of Business Administration, Chang Gung University.
通訊作者，E-mail: cytsao@mail.cgu.edu.tw

一、前言

特徵價格理論(hedonic price theory)認爲房地產價格可由該房地產之特徵所決定，換言之，住宅價格可視爲各種特徵之隱含價格的函數。實證研究上，特徵價格理論之驗證往往透過實證資料的蒐集並進行統計分析，一般來說，經常使用的特徵變數包含住宅本身特性(如：使用坪數、房屋建材、所屬樓層等)、住宅環境(如：臨路路寬、交通便利性、鄰近公共設施等)與總體經濟條件(如：利率水準、國民所得、空屋率等)等。特徵價格理論與估計方法之建立，首歸功於Court (1939), Lancaster (1966),Griliches (1971a, 1971b)與Rosen (1974)等。Rosen (1974)提出特徵方程式(hedonic equation)估計法，自此之後，學術界普遍以迴歸模型作爲特徵方程式之估計模型，亦即使用房地產價格與特徵變數資料，配適出一線性迴歸或Box-Cox型態之模型，此類方法已廣泛使用於學術界與產業界。目前以台灣資料進行的實證研究甚多，例如劉振誠(1986)、辜炳珍、劉瑞文(1989)、張金鶚(1995)，以及林國民(1996)等。

傳統的特徵方程式估計法最大的優點在於方法簡單且易於解釋，但潛在的問題卻是模型假設可能不滿足與模型過於簡化等。實證上普遍發現房地產價格資料違反常態分配的假設，雖然經由Box-Cox轉換可稍稍減緩非常態的現象，但轉換後的資料往往仍存在右斜與高狹或低擴峰的特性。另一方面，以線性方式刻畫房地產特徵與價格間的關係似乎過於簡化，變數之間可能存在非線性關係且特徵隱含價格可能因所屬狀態之不同而有不同。Anglin & Gençay (1996), Gençay & Yang (1996), Pace (1998), Clapp (2004), Boa & Wan (2004), Bin (2004), 與Martins-Filho & Bin(2005)等，考慮以半參數(semi-parametric)與無母數(non-parametric)模型估計特徵方程式，Daniels & Kamp (1999), Kershaw & Rossini (1999), Wong et al. (2002)與Lomsombunchai et al.(2004)等，嘗試以類神經網路(artificial neural networks)刻畫地產特徵與價格間的關係。以上研究普遍發現這些方法可有效增加模型解釋能力。

本文主要目的在於應用決策樹方法於特徵方程式之估計。決策樹(decision trees)可視爲是一組決策法則(rule)，其中每一個法則包含條件式與決策式，條件式描述此法則啓用之時機，決策式刻畫啓用後對應之行動。

決策樹之概念是依事物的特徵，將事物區分爲不同的種類，每一種類再對應不同的決策模式。例如在房地產模型中，即根據房屋的特徵來區分不同類別，各類別中再建立本身價格模型。因此，若利用決策樹方法於特徵方程式之估計，最後特徵方程式將包含一組子特徵方程式，每個子特徵方程式規範不同特徵下房地產特徵與價格間的關係。在現有的決策樹分類器中，以ID3 (Quinlan, 1986)、C4.5 (Quinlan, 1993)、CART (Breiman et al., 1984)最廣爲採用。上述三種方法適用的資料型態是文字符號或是離散狀態，處理的問題主要爲分類(classification)問題。Quinlan (1996)改良傳統C4.5方法，使其對於連續屬性資料之處理較有效率。本研究採用的決策樹方法爲Quilan建立的Cubist迴歸樹(regression tree)，此方法爲C4.5的改良，可處理連續的資料型態(註1)。該方法在概念上與符號型態的決策樹相同，都是依據資料屬性進行分類，兩者差異處在於，Cubist迴歸樹的終端節點是一條迴歸方程式，而不是如C4.5決策樹的終端節點是該筆資料所屬類別(註2)。Kim et al. (2006) 對於決策樹的配適能力作了一個大規模的研究，他們考慮27種實務上最常用到的決策樹模型，並將這27種模型應用在52組文獻上曾考慮過的資料上，這52組中包含兩組Boston的房地產資料。研究結果發現，表現前5名的決策樹

中，有3名是Cubist迴歸樹及其變形，分別是Cubist committee model (第1名)、Cubist & nearest-neighbor (第4名)以及Cubist rule-based model (第5名)。Fan et al. (2006)使用決策樹模型於新加坡的房地產資料，研究發現購屋者對2至4房的公寓關心的是樓地板面積、房屋型態與屋齡等，而對於5房的公寓除了關心上述房屋特徵，也關心房屋樓層。

本文使用的特徵方程式估計法爲決策樹中的Cubist迴歸樹，主要原因有三。首先，Cubist迴歸樹模型之設計符合房地產資料特性。以下例子說明此特性：若考慮房價與臨路路寬間的關係，在大部分的情況下，這兩者之間的關係可能是正向的，也就是臨路路寬越寬，房價也就越高。以台北市爲例，仁愛路、敦化南北路兩旁的房屋，其價格必定比位在巷弄間的房屋高，即便兩者其他房屋特徵均相同。然而，臨路路寬影響房價的程度，應該會與房屋所在區域有關，直覺上來說，大都會區的影響程度會比鄉鎮地區的影響程度高。若實際情況眞爲如此，建立模型時應該將資料區分爲大都會區與鄉鎮市區，並各自配適迴歸模型，同時將觀察到這兩模型之差異。另一方面，房價與臨路路寬之間不必然一直都是正向關係，舉例來說，緊鄰高速公路、快速道路、或者聯絡都會區之間省道旁的房子，其房價可能與臨路路寬沒有關係，甚至可能是有負向關係。因此在某個路寬範圍內，房價與臨路路寬之間存在正向關係，而當路寬超過這個範圍，此正向關係可能就此消失，或者反而變成負向關係。若實際情況眞爲如此，在建立模型時應該將資料依路寬大小區分爲幾個區塊，再各自配適一迴歸模型，同時將觀察到這些模型間之差異。

上述例子說明了，在不同的條件下，房價與特徵變數之間的關係可能不同，這樣的想法相當符合直覺且經常發生。當面對的資料具有此種特性，則我們使用的計量模型應該具有能夠顯現個別差異的能力，而Cubist迴歸樹正是具有此種特性的計量模型。當上述房價與臨路路寬之間複雜的關係存在時，Cubist迴歸樹首先把資料分爲大都會區、鄉鎮地區、路寬在某個範圍內、路寬在某個範圍外等四個區塊，再個別配適迴歸模型，而不是將所有資料配適成單一迴歸模型。於是，Cubist迴歸樹企圖利用多條迴歸方程式，來刻畫自變數與應變數之間的關係，從這個層面來看，Cubist迴歸樹可視爲片段式迴歸模型(piecewise linear regression model)。和一般非線性模型，如類神經網路相比，決策樹的思想等於是用一個由許多區域性的線型模型，來逼近一個全域型單一的非線性模型。這樣一種思想，普遍存在當今建模的文獻中(Chen, 2002; Chen & Wang, 2003)。以上的情況可進一步延伸，也就是攸關房價因素的特徵，可能也會因爲地區的不同、房屋類型的不同而有差異。例如位於郊區的房子，其房價可能與是否含車位沒有直接關係，然而在都會區有含車位的房子，往往具有較高的單價。因此對於位於郊區的房子，我們無須將是否含車位的變數放入迴歸模型中，但是在都會區的鑑價模型中，是否含車位的變數將扮演不可或缺的角色。Cubist迴歸樹的另一個優點是，當資料已切割爲若干區塊，Cubist將各自在區塊中選取重要的自變數，建立個別的迴歸模型。總而言之，在力求得到最佳配適的前提下，Cubist迴歸樹將依據資料屬性，自動決定將資料分割成幾個區塊、如何分、該選取哪些變數，以及估計出個別迴歸模型。

利用Cubist迴歸樹估計特徵方程式的第二個原因是，Cubist迴歸樹配適能力高且易於解釋。採用非線性模型如半參數模型、無母數模型與類神經網路等方法，雖可得到較佳的模型配適度，但最終模型往往過於複雜，因而不易分析特徵變數與房價之間的關係，於是此類模型最大的缺點是，在學術上無法利用模型結果洞察經濟現象，在實務上也不易與不動產估價

師的直覺作連結，因此常招致黑箱作業(black-box)之批評。Cubist迴歸樹於特徵方程式估計上之應用，正好介於極度簡化模型(如迴歸模型)與極度複雜模型(如類神經網路)之間，它同時承襲了兩極端模型之優點，且避免了兩極端模型的缺點。一方面它可以得到優於迴歸模型之模型配適力(註3)，另一方面雖然Cubist迴歸樹屬非線性模型，但由於其片段式迴歸模型之設計，使得最終估計之模型易於分析且可得到豐富之經濟詮釋。

使用Cubist迴歸樹估計特徵方程式的第三個原因是，當使用大量資料估計特徵方程式，Cubist迴歸樹相對於其他無母數方法運算上較有效率。首先，因爲Cubist迴歸樹是由多個迴歸模型組成，因此估計模型所需耗費時間，包含決定如何切割資料所需時間，加上估計各別迴歸式所需之時間。相較於其他無母數方法如類神經網路需利用反覆疊代方式估計模型，Cubist迴歸樹顯得較有效率。其次，當使用最小平方法(ordinary least square)作爲迴歸模型之估計方法，大量的樣本與變數將使得估計過程中面臨的矩陣運算相當耗時，這使得利用所有資料估計單一迴歸模型的方式顯得十分笨重。

本文第二節詳述決策樹方法與Cubist迴歸樹，第三節說明資料與實驗設計，第四節爲實證結果分析，第五節爲結論。

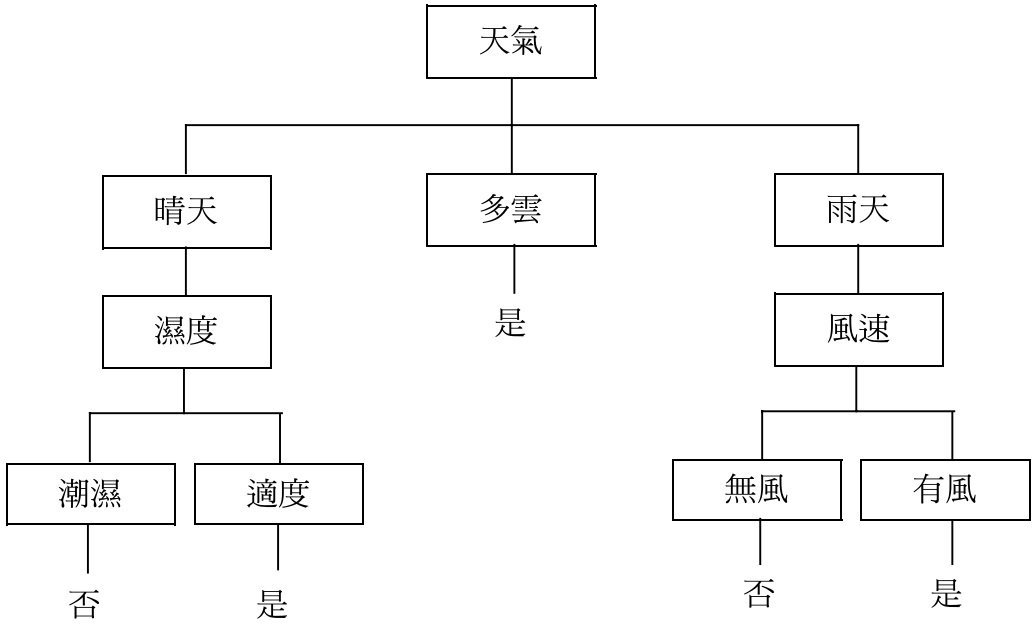
二、研究方法

本文使用Quilan建立的Cubist迴歸樹來估計特徵方程式，並使用K近鄰法修正決策樹，爲比較Cubist迴歸樹之配適與預測能力，研究中選取複迴歸模型作爲比較基準。以下分別介紹決策樹、Cubist迴歸樹、K近鄰法、以及本研究採用之模型績效衡量指標。

(一) 決策樹

決策樹可能是當今人工智慧，特別是機器學習(machine learning)這個領域中，最爲普遍使用的工具。它的普遍性可以從它做爲「標竿」(benchmark)的地位中看出：任何在機器學習領域中，正在發展的新工具，都以決策樹做爲其要超越的對象。決策樹顧名思義是一樹狀結構，只是如圖一顯示，它是一棵倒長的樹，樹根(root)在上，樹葉(leaf)在下，而中間是聯結樹根到樹葉的許多節點(node)或分叉點(branching point)，最底層的葉子點就代表一個最終的決策。以圖一的二元決策(binary decision)爲例，它代表了「是」(去打網球)或「否」(不去打網球)的兩種結果。而樹根及每一個節點則代表了一個狀態的界定(classification)，如圖一的樹根，便是對天氣做一界定，是晴天、多雲，還是雨天。而左邊的節點則是對濕度做一界定，是潮濕、還是適度，而右邊的節點則是對風速做一界定，是無風、還是有風。每一個界定都產生一個新的分叉，也稱子樹(subtree)，如此下去，一直到長到葉子爲止。而從樹根出發，經過節點，再到任一葉子之間的路徑(path)明顯地是唯一的。而每一條路徑就說明了一個決策的形成。再以圖一最左一條徑爲例，如今天之所以不打球的原因，是因今天「晴天」，而且又「潮濕」。依此類推，我們也知道其它路徑的意義，而所有的路徑之合，就整體的告訴我們在什麼情況下，會去打網球，而在什麼情況下，不打網球。因此，這就是決策樹命名的由來。

決策樹的概念，很容易地就可以應用於資料分析的問題。最早期的應用，是先從分類(classification)的問題開始。著名的軟體程式ID3就是處理分類的問題。



圖一 代表是否去打網球的決策樹

典型的分類問題可以說明如下，令

$Z_i=(X_i, Y_i)$

其中Zi爲資料集合的第i筆資料，它由兩部份所構成，其中X_i是一個特徵變數向量，由p個特徵變數所組成，

$\boldsymbol{X}_i=(X_{1i}, X_{2i}, X_{3i}, \cdots, X_{pi})$

其中每個特徵變數的值域是離散(discrete)及有限(finite)的，另外，Y_i是一個類別或決策變數(decision variable)。Y_i的值域基本上也是離散及有限的，在很多場合，也是最簡單的情況下，Y_i還是二元的。

而分類問題就是要從整個資料集合中找出X_i與Y_i的對應(mapping)關係，決策樹的構想，是將所有觀察值X_i ($i=1,\cdots, n$)依決策樹中所揭示的路徑分解開來，也就是每一個觀察值X_i都只能屬於整個決策樹中的一條路徑，不能有跨徑的行爲。而在每條路徑中的觀察值在最理想的情況下，應該是屬於同樣的類別。若不盡理想，至少也是大多數爲同樣類別，而異類只能是少數或例外。

在上述的特性下，決策樹的建構大致上是遵循下列很簡單也很直觀的程序。

第一步，我們在p個特徵變數中，先選取其一個，在不失一般性原則下，假設是X_1，再根據X_1的實現值，將所有資料X_i分類，在該特徵上是一樣的X_i，歸爲一類，按照X_1值域的規模(cardinality)，$|X_1|$，我們總共可以分成$|X_1|$類。

第二步，我們再對$|X_1|$類中的每一類，分別檢查同類中的X_i是否有同樣的決策變數(Y_i)值，若是，我們就以該值做爲葉子，將此一分枝結束；若否，我們則還要對其進行再分類。這個分類的程序會和「第一步」一樣，只是在同一條路徑上，每一個特徵變數頂多只需要出

現一次，所以，在以後的建構中，我們可以從還沒有出現過的特徵變數中來選。

第三步，依此重覆進行下去，直到葉子全部長出或特徵變數全部用盡(註4)。

以上三步雖然可以完成一棵決策樹的建構，但有兩個問題沒有處理。第一，按照以上的步驟所建構的決策樹不是唯一的。第二，以上的步驟並沒有說明從根到每一個分枝上，特徵變數的選擇是如何決定的？回答以上兩個問題，就牽涉到決策樹的數學基礎，在這裡我們不準備占用太多大量的篇幅來說明，除了指出對上述兩點，一般是利用最大亂度原則(maximum entropy principle)，在根及每一個節點上，選擇資訊利得(information gain)，也就是分類之前的資料亂度(entropy)與分類之後的亂度差最大的特徵變數，來做爲根或結點。

(二) Cubist迴歸樹

Cubist迴歸樹將估計模型以決策樹的方式呈現，依據資料屬性，將資料分割成若干區塊，再各自於區塊中選取重要的自變數，建立個別的迴歸模型。此方法在概念上與符號型態的決策樹相同，都是依據資料屬性進行分類，但Cubist迴歸樹的終端節點是一條迴歸方程式，而不是該筆資料的所屬類別，因此可以處理應變數爲連續的資料型態。估計模型可用一般化的方式表達如下(假設包含r個法則)：

法則 1：if條件式C_{11} and條件式$C_{12}\cdots$ and條件式C_{1m_1}

then $Y_i=a_{10}+a_{11}X_{1i}+a_{12}X_{2i}+\cdots+a_{1p}X_{pi}+\varepsilon_{1i}$, $i=1, 2, \cdots, n_1$

法則 2：if條件式C_{21} and條件式$C_{22}\cdots$ and條件式C_{2m_2}

then $Y_i=a_{20}+a_{21}X_{1i}+a_{22}X_{2i}+\cdots+a_{2p}X_{pi}+\varepsilon_{2i}$, $i=1, 2, \cdots, n_2$

法則r：if條件式C_{r1} and條件式$C_{r2}\cdots$ and條件式C_{rm_r}

then $Y_i=a_{r0}+a_{r1}X_{1i}+a_{r2}X_{2i}+\cdots+a_{rp}X_{pi}+\varepsilon_{ri}$, $i=1, 2, \cdots, n_r$

其中X_j是第j個解釋變數($j=1, 2, \cdots, p$)、m_j是法則j的條件式個數、n_j是滿足法則j的資料個數、a_{j0}是第j個法則迴歸式的截距項($j=1, 2, \cdots, r$)、a_{jk}是第j個法則中第k個解釋變數的係數($j=1, 2, \cdots, r$, $k=1, 2, \cdots, p$)。在本研究中，Y_i爲房地產價格(或取其對數)、$X=\{X_j, j=1, 2, \cdots, p\}$爲房地產特徵變數(或取其Box-Cox轉換)。

此模型將所有樣本資料區分爲r類，第j個法則有m_j個條件式，當資料滿足該類的所有條件時，以該類的線性迴歸式進行估計，也就是一般所謂的分段迴歸，不同的法則可以各自選擇所需要的解釋變數，並給予權重係數，當係數爲0，表示沒有使用該變數。當某一資料同時滿足多個法則時，其預測值爲多個線性模型預測值的平均。模型中的條件式一般以集合方式表現，可以包含定性或定量變數，例如$X_k \in$ {高，低} (定性變數)或$X_j>5$ (定量變數)，迴歸式則僅能包含定量變數。因此Cubist迴歸樹對於定性資料無須另設虛擬變數(dummy variable)來處理。

由以上一般化的模型可以看出，資料分類的多寡和每類法則包含的資料數目有關，所以要求終端節點包含樣本的比例越少，代表允許資料劃分較多的區塊，可得到較爲複雜的模型；若希望模型不要過於複雜，可以藉由提高終端節點包含樣本比例來簡化模型。

(三) K近鄰

本文的目的是建立房地產價格的區域化預測模型(local prediction models)，除了具有這種區域化概念的Cubist迴歸樹外，K近鄰(K nearest neighbors, KNN)另一種更簡單、更具直觀性的

方法。本文將使用結合Cubist迴歸樹與*K*近鄰的複合式方法，得出最佳的房地產模型。

如何判斷一間房屋的價值呢？一個相當直接但卻具有高度經濟意涵的方法，是直接在資料庫中找尋特徵類似的房屋，例如相同區域、相同房屋型態、相同建材、相同大小等，於是把該房屋的價格，當作是此待估價房屋價格的估計，這樣的方法正是*K*近鄰法。當給定一組房屋特徵資料，*K*近鄰法首先選取與該筆資料最相近的*K*筆資料，接下來將*K*筆資料對應的房價資料，做簡單平均或加權平均，或者建構一迴歸模型(Chen & Wang, 2003)，此平均數或迴歸模型中的應變數值，即當作未知房價的估計。使用*K*近鄰法時需選取近鄰點的個數，近鄰點的個數越多，代表資料具有平滑性，亦即具有相同屬性的資料可提供有用的訊息。

在本文使用的方法中，*K*近鄰法不是直接用來估計房地產價格，而是用來修正Cubist決策樹所估計之價格，以期得到更爲準確的鑑價模型。以使用1筆近鄰點爲例，其作法如下：假設給定一組房屋特徵(x)下，我們首先在歷史資料中搜尋與該房屋特徵最接近的特徵$x^{(1)}$，因爲是由歷史資料中搜尋，因此我們可得知此特徵最接近房屋的價格($y^{(1)}$)。接下來利用Cubist迴歸樹，分別預測此兩組特徵下的價格($\hat{y}$ 與 $\hat{y}^{(1)}$)，最後以 $\hat{y}+(y^{(1)}-\hat{y}^{(1)})$做爲模型最終的預測值。此方法背後的邏輯是，當每個模型都免不了有誤差的情況下，我們假設模型具有系統性的誤差，亦即利用房屋特徵(x)預測價格所產生的誤差，與其特徵最接近房屋的誤差近似，因此

$$y-\hat{y}\approx y^{(1)}-\hat{y}^{(1)}$$

故

$$y\approx \hat{y}+(y^{(1)}-\hat{y}^{(1)})$$

值得注意的是，由於*K*近鄰法的概念是特徵相似的資料應具有相似的型態(pattern)，因此若資料的特徵少且其中包含若干不重要的特徵，此方法的效果將受到影響。

(四) 衡量指標

本研究考慮四種衡量指標，分別是判定係數(R^2)、調整後判定係數(R_a^2)、絕對平均百分誤差(*MAPE*)與命中率(*hit_rate*)等。其定義分別是

$$R^2=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2},$$

其中y_i爲實際觀察值、$\hat{y}_i$ 爲模型配適或預測值、$\bar{y}$ 爲觀察值樣本平均數，

$$R_a^2=1-(1-R^2)\frac{n-1}{n-q},$$

其中n爲樣本數、q爲模型中估計係數之個數。在複迴歸模型中q即爲自變數個數加1 (若含常數項)，在Cubist迴歸樹中q設定爲所有法則估計係數之個數總和(包含條件式與迴歸式)。

$$MAPE=\frac{\sum|y_i-\hat{y}_i|/y_i}{n},$$

$$hit\_rate(\alpha)=\frac{\sum \mathrm{I}\{(1-\alpha)y_i\le\hat{y}_i\le(1+\alpha)y_i\}}{n},$$

其中$0\le\alpha\le1$且$\mathrm{I}\{L\}$等於1若L爲眞，其餘爲0。因此命中率顯示模型配適或預測值落在眞實

值上下100α%的比例。本研究選取三種α值，分別是0.05、0.1與0.2。以上四種指標除了*MAPE*外，皆是數值越大代表模型配適或預測能力越佳，*MAPE*則是數值越小越好。

三、資料與實驗設計

(一) 資料來源與變數說明

本文所使用的台灣房地產價格和相關建物特徵資料來自國內某銀行，期間自2002年2月至2004年5月，原始資料經處理後共有45,419筆，其中40,875筆爲訓練樣本，其餘爲測試樣本(註5) (註6)。本文使用25個特徵變數來估計房屋價格，特徵變數包含建物本身屬性(個體變數)和總體指標(總體變數)兩大類，其中個體變數包括：持份面積、地坪、公告現值、公共設施比率、路寬、屋齡、樓層、含車位與否、所屬樓層、房屋種類、主要建材、建物主要用途、周圍情形、縣市等變數；總體變數包括：戶口數、流入人口數、流出人口數、建照數、使用執照數、房價指數變動率、領先指標變動率、房屋建築業生產指數變動率、購屋貸款利率、台灣加權股價指數月底值變動率、所得成長率等。

這些特徵變數若依性質可分成定量變數和定性變數，其中持份面積、地坪、公告現值、公共設施比率、路寬、屋齡、樓層以及全部總體指標的變數屬於定量變數；而房屋種類、主要建材、建物主要用途、周圍情形、縣市、所屬樓層、含車位與否等變數則是屬於定性變數。定量變數中的持份面積、地坪、公告現值、路寬、屋齡、戶口數、流入人口數、流出人口數、建照數、使用執照數等變數經對數化處理；公共設施比率、樓層、房價指數變動率、領先指標變動率、房屋建築業生產指數變動率、購屋貸款利率、台灣加權股價指數月底值變動率、所得成長率等變數則未對數化。

表一是房地產價格對數化前後的基本統計量，由此表可看出對數化前後都顯著拒絕房地產價格是常態分配的假設，原始價格資料呈右偏及高狹峰之型態，對數化價格資料呈些微右偏及低闊峰之態勢。表二則列出三個重要定量變數對數化後的基本統計量，分別是ln持份面積、ln地坪與ln公告現值等。由此表可知，三個對數化後的資料都顯著拒絕常態分配的假設，持份面積和公告地價屬於左偏、高狹峰的分配，而地坪則屬於右偏、低闊峰的分配。

表三是三個重要定性變數的次數分配狀況，由此可看出，房屋種類中以套房、公寓、大廈佔樣本的多數，大約佔了7成，商業種類的建物相對少數，透天厝亦佔有不少的樣本。所屬樓層是1樓(不含全棟)的樣本約佔9%，頂樓約佔11%，全棟建物約佔23%，地下室佔樣本極少數。樣本區位分佈的狀況相當不平均，但大致與全台灣各區域房地產交易活絡程度接近。資料中以台北縣的樣本數最多，佔樣本數的27%，樣本數最少的縣市爲南投縣，6成以上的樣本來自桃竹苗以北的區域。

(二) 實驗設計

本研究結合Cubist迴歸樹與*K*近鄰兩種方法作爲房地產特徵方程式之估計模型，並與傳統複迴歸模型的預測結果做比較。研究中使用的解釋變數除了特徵變數本身外，爲刻畫特徵變數與房地產價格間可能存在之非線性關係，本文考慮個體變數(除含車位與否)的平方項與交叉相乘項，因樣本資料涵蓋的時間只有兩年多，所以不考慮總體變數的平方項與交叉相乘項，以及其與個體變數的交乘項。使用Cubist迴歸樹估計模型時一共使用53個變數，包含25個特徵

表一 房地產價格基本統計量

	房地產價格	ln房地產價格
平均數	4,076,014	15.11
標準差	1,974,413	0.47
最小值	1,347,000	14.11
最大值	10,750,029	16.19
偏態(p-value)	1.00 (0.00)	0.08 (0.00)
峰度(p-value)	0.56 (0.00)	-0.71 (0.00)
K-S檢定 (p-value)	21.45 (0.00)	6.75 (0.00)

表二 重要定量變數基本統計量

	ln持份面積	ln地坪	ln公告現值
平均數	3.60	2.12	11.61
標準差	0.39	0.82	0.88
最小值	1.95	-3.00	5.80
最大值	5.17	6.88	14.87
偏態(p-value)	-0.13 (0.00)	0.11 (0.00)	-0.24 (0.00)
峰度(p-value)	0.33 (0.00)	-0.01 (0.65)	0.49 (0.00)
K-S檢定(p-value)	2.88 (0.00)	13.80 (0.00)	6.09 (0.00)

變數、7個平方項、21個交乘項等。至於研究中用來當作基準的複迴歸模型，則需要使用虛擬變數來處理定性變數，舉例來說，若想要測試房屋所屬縣市對建物價格的影響，則需將所屬縣市以虛擬變數的方式定義，16個縣市分類共需設立15個虛擬變數，因此複迴規模型一共需要使用80個變數。

在Cubist迴歸樹與*K*近鄰所使用的參數設定方面，本研究採用交叉驗證(cross validation)法決定，進行步驟說明如下：首先將全部訓練樣本隨機區分爲10個段落，輪流保留其中1個段落的資料，利用其他9個段落的資料進行模型配適，並衡量所配適的模型在保留資料上的預測能力。最終模型參數的決定，將是使估計模型最具有預測能力的那一組。本文考慮四種不同的近鄰個數，分別是0、3、6與9，另外考慮兩種不同的終端節點樣本比例，分別是0.5%與1%，因此共有八組候選模型。表四爲使用交叉驗證法所得到的樣本外預測結果，我們分別以調整後判定係數(R_a^2)、絕對平均百分誤差(*MAPE*)、命中率(*hit_rate*)作爲判斷指標，並採取投票策略(voting policy)作爲選取依據，也就是先針對每一個判斷指標，決定表現最好的參數組合並給予註記，最終參數的選取，將是擁有最多註記的模型。依據以上的選取方式，我們選擇使用9個近鄰的模型，且終端節點樣本比例爲0.5%。

表三 重要定性變數次數分配

房屋種類			所屬樓層			縣市		
分類	次數	%	分類	次數	%	分類	次數	%
套房、公寓、大廈	32107	70.69	1樓	4081	8.99	台北市	7129	15.70
工業住宅	392	0.86	頂樓	5014	11.04	基隆市	726	1.60
辦公室、住辦、店面	3440	7.57	地下樓層	54	0.12	台北縣	12271	27.02
透天厝、別墅	8293	18.26	全棟	10448	23.00	新竹市	498	1.10
其他	1187	2.61	其他	25822	56.85	新竹縣	468	1.03
						桃園縣	6264	13.79
						苗栗縣	148	0.33
						台中市	4124	9.08
						台中縣	1990	4.38
						南投縣	121	0.27
						高雄市	5429	11.95
						高雄縣	2476	5.45
						屏東縣	283	0.62
						台南縣市	913	2.01
						宜蘭花蓮	648	1.43
						雲林彰化	1931	4.25

表四 交叉驗證結果

Minimum Coverage = 0.5%					
近鄰個數	R_a^2	*MAPE*	*hit_rate* (5%)	*hit_rate* (10%)	*hit_rate* (20%)
0	87.39%	12.50%	30.19%	53.52%	80.77%
3	87.39%	12.50%	30.19%	53.52%	80.77%
6	88.41%	12.02%	30.78%	54.54%	82.36%
9	88.66%	11.90%	30.92%	54.64%	82.67%
Minimum Coverage = 1%					
近鄰個數	R_a^2	*MAPE*	*hit_rate* (5%)	*hit_rate* (10%)	*hit_rate* (20%)
0	87.59%	12.67%	27.56%	51.11%	80.54%
3	87.48%	12.51%	30.41%	53.20%	80.65%
6	88.54%	12.02%	30.68%	54.45%	82.30%
9	88.77%	11.91%	30.81%	54.55%	82.67%

四、實證分析

(一) 模型估計

研究結果發現Cubist迴歸樹模型包含36個法則，也就是本研究估計出的特徵方程式，包含36個子迴歸模型。此結果顯示房地產價格與特徵變數之間確實存在複雜的非線性關係。經由統計顯示，模型平均使用4.28個變數(包含定性與定量變數)來區分這36個區域，標準差是1.23。在這36條迴歸式中，平均使用21.08個變數(定量變數)建構每一條迴歸模型，標準差是5.28。Cubist迴歸樹的解讀方式可以下例說明之，我們以模型中第21個法則爲例(註7)，其結果爲：

if

縣市∈{台北市，新竹縣}

$(\ln 地坪)^2 \le 8.106732$

ln公告現值*ln持份面積≤42.64024

樓層*ln屋齡>5.598218

公設比*ln屋齡≤0.3192939

then

$$\begin{aligned}\ln 房地產價格 = {} & 11.5527 + 0.092\ \ln 公告現值 * \ln 持份面積 \\ & - 0.0242\ \ln 公告現值 * \ln 屋齡 - 0.114\ (\ln 持份面積)^2 \\ & + 0.063\ \ln 持份面積 * \ln 屋齡 - 0.038\ \ln 路寬 * \ln 公告現值 \\ & + 0.0111\ (\ln 公告現值)^2 + 0.067\ \ln 路寬 * \ln 屋齡 \\ & + 0.57 公設比 * \ln 路寬 + 0.05\ \ln 地坪 * \ln 持份面積 \\ & + 0.071\ \ln 持份面積 * \ln 路寬 - 0.039\ \ln 地坪 * \ln 屋齡 \\ & + 0.035\ (\ln 地坪)^2 - 0.04 購屋貸款利率 - 0.03\ \ln 公告現值\end{aligned}$$

在“if”的部分界定此法則適用範圍，“then”的部分刻畫房地產價格迴歸模型。在本例中，限定的範圍是在台北市或新竹縣，同時地坪對數化平方小於等於8.11、公告現值對數化乘上持份面積對數化小於等於42.64、樓層乘上屋齡對數化大於5.60、再加上公設比乘上屋齡對數化小於等於0.32。在此限制下，房地產價格將與公告現值對數化乘上持份面積對數化、持份面積對數化乘上屋齡對數化、公告現值對數化平方、路寬對數化乘上屋齡對數化、公設比乘上路寬對數化、地坪對數化乘上持份面積對數化、持份面積對數化乘上路寬對數化、地坪對數化平方有正向關係，與公告現值對數化乘上屋齡對數化、持份面積對數化平方、路寬對數化乘上公告現值對數化、地坪對數化乘上屋齡對數化、購屋貸款利率、公告現值對數化有負向關係。

在基準模型方面，複迴歸模型的估計結果顯示大部分變數皆呈現統計顯著，在5%顯著水準下47個定量變數中只有6個變數未達統計顯著，分別是戶口數對數化、流入人口數對數化、使用執照數對數化、房價指數、房屋建築業生產指數成長率、樓層乘上地坪對數化等，33個虛擬變數(dummy variable)中則有7個變數未達統計顯著，分別是所屬樓層中的地下樓層、房屋種類中的住宅與工宅、主要建材中的鋼骨結構、主要用途中的住家與工業、周圍情形中的住商混和等(註8)。

(二) 模型表現

爲比較Cubist迴歸樹與傳統複迴歸模型對房地產資料之配適與預測能力，本研究考慮四種衡量指標，分別是判定係數(R^2)、調整後判定係數(R_a^2)、絕對平均百分誤差(*MAPE*)與命中率(*hit_rate*)等。表五爲模型配適與預測結果，由此表可看出，在所有衡量指標中Cubist迴歸樹的配適與預測能力均優於傳統複迴歸模型。舉例來說，Cubist迴歸樹樣本內的R^2高達89.54%，傳統複迴歸模型則爲86.02%，經由係數個數調整後的R_a^2仍顯示Cubist迴歸樹優於複迴歸模型，Cubist迴歸樹的絕對平均百分誤差較複迴歸低2%，再者，Cubist迴歸樹在不同等級的區間內皆顯示其較佳的命中率。在樣本外的預測能力方面，表五顯示在所有指標中Cubist迴歸樹依然優於複迴歸模型。本研究進一步將兩模型於訓練樣本與測試樣本內的命中率進行統計檢定，考慮以下假設

$$\begin{cases} H_0 : hit\_rate(\alpha)_1 \le hit\_rate(\alpha)_2 \\ H_a : hit\_rate(\alpha)_1 > hit\_rate(\alpha)_2 \end{cases}$$

其中$hit_rate(\alpha)_1$爲Cubist迴歸樹之母體命中率，$hit_rate(\alpha)_2$爲複迴歸模型之母體命中率。檢定結果顯示無論在訓練樣本或測試樣本，以及3種不同的比例等級中，Cubist迴歸樹之命中率均顯著大於複迴歸模型之命中率。

使用統計模型進行資料分析時，除了希望模型具有高度之解釋力，也希望模型穩健，以此模型衍生出之經濟意涵才具有較高之信賴度。本研究發現Cubist迴歸樹雖然較迴歸模型複雜且富有彈性，但在房地產資料之應用上顯示其高度穩定性。由表五中發現Cubist迴歸樹在訓練樣本與測試樣本間的表現差異不大，平均來說在測試樣本中的表現較訓練樣本降了0.8%，而複迴歸模型則平均降了2.61%。綜合上述分析，Cubist迴歸樹在特徵方程式之估計，不但顯現其高度之配適與預測能力，且未發生模型過度配適之問題。

(三) 變數分析

我們進一步分析哪些變數用來區分36個迴歸模型，表六顯示條件式中使用變數之頻率，頻率計算方式爲該變數出現在條件式的次數，除以法則總數。舉例來說，表六顯示「所屬樓層」在條件式中出現的比例爲33.33%，表示三分之一的法則使用了所屬樓層，作爲區分樣本

表五　模型表現

		R^2	R_a^2	*MAPE*	*hit_rate*(5%)	*hit_rate*(10%)	*hit_rate*(20%)
訓練樣本	Cubist迴歸樹	89.54%	89.45%	11.53%	31.73%	55.85%	83.76%
	複迴歸	86.02%	85.99%	13.75%	24.78%	46.92%	77.28%
	t-檢定量(p-value)				44.15 (0.00)	51.10 (0.00)	46.77 (0.00)
測試樣本	Cubist迴歸樹	88.92%	88.83%	11.90%	29.90%	54.73%	83.51%
	複迴歸	83.00%	82.69%	15.10%	23.28%	43.88%	73.83%
	t-檢定量(p-value)				14.28 (0.00)	20.70 (0.00)	22.53 (0.00)

表六　Cubist迴歸樹變數使用頻率

變數	ln持份面積	(ln持份面積)^2	公設比	(公設比)^2	ln路寬	(ln路寬)^2
條件式	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
迴歸式	0.00%	94.44%	0.00%	44.44%	13.89%	0.00%
變數	ln屋齡	(ln屋齡)2	ln戶口數	ln流入人口數	ln流出人口數	ln建照數
條件式	5.56%	22.22%	27.78%	0.00%	0.00%	0.00%
迴歸式	88.89%	33.33%	50.00%	38.89%	44.44%	11.11%
變數	ln使用執照數	房價指數	領先指標成長率	房屋建築業生產指數成長率	購屋貸款利率	台灣加權股價指數月底值成長率
條件式	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
迴歸式	22.22%	16.67%	13.89%	0.00%	52.78%	5.56%
變數	所得成長率	樓層	(樓層)2	ln地坪	(ln地坪)2	ln公告現值
條件式	0.00%	0.00%	0.00%	11.11%	27.78%	0.00%
迴歸式	5.56%	25.00%	0.00%	88.89%	50.00%	30.56%
變數	(ln公告現值)^2	車位	ln持份面積* 樓層	ln持份面積*ln屋齡	ln持份面積*ln路寬	公設比* ln持份面積
條件式	19.44%	0.00%	0.00%	0.00%	0.00%	5.56%
迴歸式	72.22%	8.33%	77.78%	72.22%	69.44%	50.00%
變數	ln持份面積*ln公告現值	ln持份面積*ln地坪	樓層* ln屋齡	樓層* ln路寬	樓層* ln公告現值	樓層* ln地坪
條件式	77.78%	0.00%	13.89%	0.00%	0.00%	0.00%
迴歸式	88.89%	88.89%	36.11%	27.78%	80.56%	55.56%
變數	ln屋齡* ln路寬	ln屋齡* ln公告現值	ln屋齡* ln地坪	ln路寬* ln公告現值	ln路寬* ln地坪	ln公告現值*ln地坪
條件式	0.00%	19.44%	0.00%	0.00%	0.00%	0.00%
迴歸式	52.78%	91.67%	61.11%	75.00%	61.11%	86.11%
變數	公設比*樓層	公設比* ln屋齡	公設比* ln路寬	公設比* ln公告現值	公設比* ln地坪	
條件式	0.00%	55.56%	0.00%	0.00%	0.00%	
迴歸式	36.11%	33.33%	27.78%	58.33%	50.00%	
變數	所屬樓層	房屋種類	主要建材	周圍情形	建物用途	縣市
條件式	33.33%	13.89%	0.00%	0.00%	5.56%	88.89%
迴歸式	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

說明1：條件式變數使用頻率計算方式爲該變數出現在條件式的次數，除以法則總數。

說明2：迴歸式變數使用頻率計算方式爲該變數出現在迴歸式的次數，除以法則總數。

資料之變數之一。在特徵方程式之應用上，若條件式中使用了某一特徵變數，代表決定房地產價格之因素與程度，會因爲該變數的不同而有所不同。變數使用的頻率越高，代表其他特徵變數與房地產價格之間的關係，越容易受到該變數所影響。由表六中我們發現，6個定性變數中只有4個變數(所屬樓層、房屋種類、建物用途與縣市)被用來區分這36個區域，主要建材與周圍情形則完全沒有使用。分析原因可能是因爲台灣地區並不像歐美國家，有明顯的商業、工業與住宅區等明確之劃分，因此周圍情形較不影響房地產價格，縣市差異才是眞正影響房價的主要因素。其次，在地窄人稠的台灣，擁擠的都會區中所屬樓層也扮演重要關鍵。再者，房屋種類往往反應了房屋主要建材，因此在給定房屋種類之資訊下，主要建材之邊際解釋力較弱，因此也被忽略。最後，我們發現建物用途也會影響特徵變數的隱含價格，但所佔比例較少。在條件式中使用定量變數方面，47個定量變數中有11變數使用於條件式中。進一步分析這11個定量變數有兩點有趣發現，首先，11個變數中有8個變數是特徵變數的交乘項或平方項，其次，僅有一個總體變數(戶口數)使用於條件式中，其餘特徵變數皆屬房屋個體變數。綜合上述結果，本文發現區分子特徵方程式的定量變數主要以房屋個體變數爲主，且是以非線性方式區分。

表六也顯示迴歸式中使用變數之頻率，頻率計算方式仍是以該變數出現在迴歸式的次數，除以法則總數。迴歸式中使用某變數之頻率越高，表示有越高比例的樣本資料，使用該變數作爲房地產價格之解釋變數。舉例來說，表六顯示「樓層」在條件式中出現的比例爲25.00%，表示歸類在這四分之一法則的房地產案件，樓層變數會影響該房地產價格。因模型設計在迴歸式中只會使用定量變數，因此6個定性變數之使用頻率皆爲零。研究發現47個定量變數中只有5個變數從未使用，包含ln持份面積、公設比、$(公設比)^2$、$(ln路寬)^2$、房屋建築業生產指數成長率與$(樓層)^2$等。值得注意的是，一般認爲影響房地產價格的重要因素是持分面積，但本研究發現持分面積本身無解釋力，其平方項才是決定房地產價格之重要變數(註9)。所有變數中持分面積平方是使用率最高的變數(94.44%)，次高使用率的變數爲屋齡與公告現值之交乘項(91.67%)，第三高之變數爲屋齡、地坪、持分面積與公告現值之交乘項，以及持分面積與地坪之交乘項(88.89%)。此外，這5個未曾使用的變數若與複迴歸模型中未達統計顯著的變數相比，僅有房屋建築業生產指數成長率這項變數重複。

上述分析發現非線性變數在特徵方程式中似乎有較高的使用率，另一方面，本研究也發現較常使用的變數大部分與個體變數有關。爲進一步分析變數屬性與變數形態之差異，我們將47個定量變數依變數屬性區分爲個體變數(36)與總體變數(11)，同時依據變數形態也可區分爲線性變數(19)與非線性變數(28)。表七爲依據變數屬性與變數形態區分之平均使用次數，由表七可發現模型中使用個體變數與非線性變數之平均次數，明顯高於總體變數與線性變數之平均次數，經由統計檢定發現兩者之平均差異均達統計顯著。綜合上述結果，本研究發現大部分的變數皆可部分地解釋房地產價格資料，其中個體變數較總體變數具有更廣泛之解釋力，再者，特徵變數與房地產價格間普遍存在非線性關係。

最後，本研究探討哪些個體特徵變數在模型的迴歸式扮演重要角色，我們鎖定的特徵變數包含持份面積、公設比、路寬、屋齡、樓層、地坪，以及公告現值等7個變數。由於本研究考慮的變數包含個體變數本身、個體變數平方項，以及個體變數之兩兩交乘項，因此針對每一個個體特徵變數，一共有8個變數與之相關。爲了清楚呈現各特徵變數的相對重要性，表八

表七　變數於Cubist迴歸式平均使用次數(依變數屬性與變數形態區分)

總體變數 平均次數	個體變數 平均次數	平均數差	t-檢定量	p-value
8.55	17.81	-9.27	-2.63	0.01
線性變數 平均次數	非線性變數 平均次數	平均數差	t-檢定量	p-value
9.30	20.25	-10.95	-3.93	0.00

表八　主要個體變數平均使用次數

特徵變數	持份面積	公設比	路寬	屋齡
平均使用次數	5.42	3.00	3.28	4.69
頻率	67.71%	37.50%	40.97%	58.68%
特徵變數	樓層	地坪	公告現值	
次數	3.39	5.42	5.83	
頻率	42.36%	67.71%	72.92%	

爲這7個個體特徵變數平均使用次數，其中平均使用次數代表該變數在36個迴歸式中，每條迴歸式平均使用幾個與該特徵變數有關之變數，平均使用次數越多，象徵特徵方程式中充斥越多與該特徵變數有關之結構，因此該特徵變數也越重要。舉例來說，「公告現值」平均使用次數爲5.83，顯示在8個與公告現值有關的變數中，每條迴歸式平均使用了5.83個變數，相較於平均只出現3.00次的「公設比」，公告現值明顯較公設比在特徵方程式的結構中，佔有更重要的地位。表八顯示公告現值、持份面積與地坪是較重要的特徵變數，公設比、路寬與樓層的重要性相對較低。綜合上述結果，本研究發現每個個體特徵變數皆影響房地產價格，影響較大的特徵除了建物權狀上登載的面積外，土地權狀上登錄的持份面積也極具影響力，此外，政府部門的公告現值是重要參考資訊。

五、結論

計量經濟源自於對經濟理論之驗證，透過資料之蒐集與計量方法之適當使用，研究者可驗證經濟理論在實證上是否成立，或進一步提出修正經濟理論之方向。隨著統計與數量方法之快速進步與多樣化發展，計量經濟模型不再只是被動擔任輔佐角色，它主動出擊迎戰實證資料，希望藉由適當的模型設計以便從資料中擷取有用資訊，此資訊將可提供未來經濟理論發展之重要依據。本研究企圖利用迴歸樹模型於大量房地產資料上，企圖發掘決定房地產價格之重要因素與其影響方式，研究結果顯示迴歸樹不僅在設計上符合房地產資料之特性，實證研究發現其配適與預測能力皆顯著優於傳統複迴歸模型。因此，本文之研究成果在學術上可提供後續房地產相關理論發展之參考，在實務上可作爲房地產從業人員估算房價之簡便工具。

利用全台灣45,419筆房地產資料爲研究樣本，本研究發現房地產特徵方程式可以36個片段

式迴歸模型表示，每筆房地產依其特徵有不同之價格決定方式，在不同的子特徵方程式中，特徵變數影響房地產價格的方式與程度皆可能不同。綜合分析以迴歸樹模型估計出之特徵方程式，發現用來區分子特徵方程式的定性變數主要以縣市、所屬樓層以及房屋種類爲主，定量變數則以房屋個體變數爲主，且是以非線性方式區分。同時，本研究發現大部分的變數皆可部分地解釋房地產價格資料，其中個體變數較總體變數具有更廣泛之解釋力，再者，特徵變數與房地產價格間普遍存在非線性關係。我們進一步追蹤哪些原始特徵變數具有較高影響力，發現影響較大的特徵除了建物權狀上登載的面積外，土地權狀上登錄的持份面積也極具影響力，此外，政府部門的公告現值是重要參考資訊。

在理論與實務上房地產價格的波動往往與某些總體經濟指標有高度的連動性，但本研究發現總體變數對房地產價格的影響力較薄弱，分析原因可能是因爲樣本僅涵蓋2年期的資料，因此總體效果較不易於模型中呈現出來。後續研究將是探討總體變數之效果以及其與個體變數之交互影響，同時爲因應此研究議題，需改良原始的迴歸樹模型使其適用於縱斷面與橫斷面資料，我們將panel迴歸樹的發展列入未來研究項目。

註 釋

註1：www.rulequest.com/cubist-info.html.

註2：本文第二節將詳述決策樹方法與Cubist迴歸樹。

註3：類神經網路在理論上可得到最佳之模型配適，但需仰賴正確的模型設定，如：隱藏層數、節點數、學習方法、學習次數等。目前學術上並無一有效的方法決定此些設定，因此使用上可能因不正確的設定而得到欠佳的結果。

註4：若到了這一步，分在同一路徑的資料仍有不同的決策變數值，則在實務上，我們可用多數決來產生葉子值，並爲該路徑劃下句點。

註5：資料處理分爲兩階段，第一階段將輸入錯誤或缺失的資料剔除，第二階段將剩餘資料配適一迴歸模型，利用Belsley et al. (1980)的判斷準則進行異常點的刪除，資料刪除比率爲4.46%。

註6：因決策樹可將資燎進行分割，並顯示不同區塊資料之特性，因此適用於大量樣本數的資料庫。

註7：因文章篇幅限制，僅呈現36個法則中的1個法則，如需完整模型請洽作者。

註8：因文章篇幅限制未列出迴歸模型中81個係數(含常數項)之估計結果，如需完整模型請洽作者。

註9：爲簡潔呈現研究結果，以下撰文均不提及經對數轉換之變數，實際使用之變數型態請參照第三節之說明。

參考文獻

林國民

1996 《高雄自有住宅特徵價格之研究》碩士論文，國立成功大學。

張金鶚主持

1995 《台灣地區住宅價格指數之研究》，行政院經建會委託研究。

辜炳珍、劉瑞文

1989 《房地產價格指數編查之研究》，行政院主計處。

劉振誠

1986 《住宅價格影響因素之研究以台北市松山、中山、大安、古亭區爲例》碩士論文，國立中興大學。

Anglin, P. & R. Gençay

1996 "Semiparametric Estimation of a Hedonic Price Function," *Journal of Applied Econometrics*. 11: 633-648.

Belsley, D. A., E. Huh & R. E. Welsch

1980 *Regression Diagnostics: Identifying Influential Data and Source of Collinearity*. New York: John Wiley and Sons.

Bin, O.

2004 "A Prediction Comparison of Housing Sales Prices by Parametric versus Semiparametric Regression," *Journal of Housing Economics*. 13: 68-84.

Boa, H. & A. Wan

2004 "On the Use of Spline Smoothing in Estimating Hedonic Housing Price Models: Empirical Evidence Using Hong Kong Data," *Real Estate Economics*. 32(3): 487-507.

Breiman, L., J. H. Friedman, R. A. Olsen & C. J. Stone

1984 *Classification and Regression Trees*. CA: Wadsworth.

Chen, S.-H., ed.

2002 *Genetic Algorithms and Genetic Programming in Computational Finance*. Oxford: Kluwer.

Chen, S.-H. & P. Wang

2003 *Computational Intelligence in Economics and Finance*. New York: Springer.

Clapp, J.

2004 "A Semiparametric Method for Estimating Local House Price Indices," *Real Estate Economics*. 15(1): 127-160.

Court, A.

1939 "Hedonic Price Indexes with Automotive Examples," in *The Dynamics of Automobile Demand*. 99-117, New York: General Motors Corporation.

Daniels, H. & B. Kamp

1999 "Application of MLP Networks to Bond Rating and House Pricing," *Neural Computing*

*& Application*s. 8: 226-234.

Fan, G.-Z., S. E. Ong & H. C. Koh

2006 "Determinants of House Price: A Decision Tree Approach," *Urban Studies*. 43(12): 2301-2316.

Gençay, R. & X. Yang

1996 "A Forecast Comparison of Residential Housing Prices by Parametric versus Semiparametric Conditional Mean Estimators," *Economics Letters*. 52: 129-135.

Griliches, Z.

1971a "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change," in *Price Indexes and Quality Change: Studies in New Methods of Measurement*. 55-87. ed. Z. Griliches, Cambridge: Harvard University Press.

1971b "Hedonic Price Indexes Revisited," in *Price Indexes and Quality Change: Studies in New Methods of Measurement*. 3-15. ed. Z. Griliches, Cambridge: Harvard University Press.

Kim, H., W.-Y. Loh, Y.-S. Shih & P. Chaudhuri

2007 "Visualizable and Interpretable Regression Models with Good Prediction Power," *IIE Transactions Special Issue on Data Mining and Web Mining* (forthcoming).

Kershaw, P. & P. Rossini

1999 "Using Neural Networks to Estimate Constant Quality House Price Indices," in P*roceedings of the Fifth Annual Pacific-Rim Real Estate Society Conference*. Kuala Lumpur, Malaysia.

Lancaster, K.

1966 "A New Approach to Consumer Theory," *Journal of Political Economy*. 74: 132-157.

Lomsombunchai, V., C. Gan & M. Lee

2004 "House Price Prediction: Hedonic Price Models vs. Artificial Neural Nets," *American Journal of Applied Statistics*. 1(3): 193-201.

Martins-Filho, C. & O. Bin

2005 "Estimation of Hedonic Price Functions via Additive Nonparametric Regression," *Empirical Economics*. 30: 93-114.

Pace, R.

1998 "Appraisal Using Generalized Additive Models," *Journal of Real Estate Research*. 15: 77-99.

Quinlan, J. R.

1986 "Introduction of Decision Tree," *Machine Learning*. 1: 81-106.

1993 *C4.5: Programs for Machine Learning*. CA: Morgan Kaufmann.

1996 "Improved Use of Continuous Attributes in C4.5," *Journal of Artificial Intelligence Research*. 4: 77-90.

Rosen, S.

1974 “Hedonic Prices and Implicit Markets: Product Differentiation in Perfect Competition,” *Journal of Political Economy*. 82(1): 34-55.

Wong, K. C., A. T. P. So & Y. C. Hung

2002 “Neural Network vs. Hedonic Price Model: Appraisal of High-Density Condominiums,” in *Real Estate Valuation Theory*. 181-198. ed. K. Wang and M. L. Wolverton, Boston: Kluwer Academic.