國立交通大學

資訊科學與工程研究所

碩士論文

基於實價登錄的房價模型研究

Real Estate Price Models of Based on Real Price Registration

研 究 生：邱司杰

指導教授：易志偉　教授

中華民國　103　年　8　月

基於實價登錄的房價模型研究

Real Estate Price Models of Based on Real Price Registration

研 究 生：邱司杰　　Student：Szu-Chieh Chiu

指導教授：易志偉　　Advisor：Chih-Wei Yi

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

August 2014

Hsinchu, Taiwan, Republic of China

中華民國 103 年 8 月

基於實價登錄的房價模型研究

學生：邱司杰　　指導教授：易志偉 教授

國立交通大學

資訊學院資訊學程

摘要

近年來世界各國吹起開放政府資料的熱潮，我國內政部也將實價登錄制度推上檯面，以促進台灣房地產交易資訊的透明化。本篇主要工作在於建立包含實價登錄資料在內的多源的房地產交易資料收集系統，並利用資料視覺化技術，透過簡單易懂的圖表呈現，讓使用者能迅速了解複雜的房市交易資料並掌握房市走向。最後依據實價登錄資料建立房價預測模型估算房價。

我們收集了 101 年 8 月到 102 年 9 月的實價登錄資料以及有巢氏代售房屋資料。視覺化方面以架設網站的方式做資料的呈現，除了長條圖、折線圖，還用了堆疊圖、散布圖等多樣化圖表呈現台灣房市。房價模型建構則依據實價登錄系統中之交易資料為依據，利用 SPSS 統計分析軟體建立線性迴歸的房價模型以進行房價推算。我們以坪數、屋齡、格局、房屋種類等資訊為參數，透過半對數模型對台北市各區的房屋價格的估算，提供大眾買屋或賣屋的價格參考依據。

關鍵字：實價登錄、房價模型、資料視覺化。

Price Model of Real Estate Based on Real Price Registration

Student：Szu-Chieh Chiu Advisors：Prof. Chih-Wei Yi

Degree Program of Computer Science

National Chiao Tung University

Abstract

In recent years, open data is a trend in many country to access public-interesting data. In the meanwhile, the Ministry of the Interior, Taiwan has been promoting the Real Price Registration System to make the real estate transaction data public. In this work, we would like to build a system to consistently real estate price data from various sources for future research works, visualize the collected data for easy understanding, and develop skills to model the house price market.

In this work, we built a database to collect house price data, including the real estate transaction data obtained from the Real Price Registration System (from August 2012 to September 2013), and the real estate pricing data periodically crawled from the U-Trust website. A web site was created to demonstrate the collected data by various visualization techniques, including bar charts, line charts, stack charts, and other diversified graphs to reveal the trends of the housing market in Taiwan. House price models were constructed by SPSS linear regression based on the transactions records from the Real Price Registration System. The parameters used in the models include the floor number, building age, pattern, house types, etc. We developed the price models for the districts of Taipei City.

Keywords：Real Price Registration, House Price Prediction Model

誌　　謝

感謝這兩年來，易志偉老師細心的教導，讓我在 NOL 實驗室學習關於巨量資料的收集、處理跟呈現，並教導如何做研究以己解決問題的能力，以及傾聽對未來的想法以及生活規劃並給予協助，感覺像是很親切的家人一樣，讓我在這兩年過得很快樂。

感謝我周遭的人一路支持我，給我鼓勵。我的家人一路陪伴我、照顧我。我們家的狗，財財，載我回到家的時候都熱情的迎接我，讓我感受到家的溫暖。感謝室友們的照顧，讓我可以快快樂樂的度過碩士班時光。感謝我們家的兩隻貓貓，Eigen 和 Pseudo，在我壓力很大的時候以超及撫慰人心的可愛臉旁看著我，讓我壓力全消。每天早上準時喵喵叫我起床，讓我不會遲到。

感謝實驗室的同學，在研究的路上給我很多幫助。尤其是我們這組的學弟妹，MarsW 和傻笑，幫了我不少忙。

感謝系羽跟實驗室的夥伴們，在我做研究心煩中，總是可以當作我的宣洩管道，聽我發牢騷，陪我玩耍、跟我打球，讓我在研究過程中可以是當的調劑身心。

邱司杰 於

國立交通大學網路工程研究所碩士班

中華民國一O三年八月

目錄

圖目錄

表目錄

chapter 1 研究簡介

本研究主要在收集多源的房地產買賣資訊，除實價登錄資料外，還有其他房仲網的待售房屋資料。用視覺化的方式呈現在網頁上，並利用實價登錄資料來建立房價模型。本章節主要介紹研究動機、主題、方法及貢獻。1.1 研究動機是在新聞中看到實價登錄的相關報導，在質疑資料的正確性，如買低報高的情況。1.2 研究主題是以線性迴歸的方式建立房價模型，希望能藉此更了解台灣房地產買賣的模式，更能掌握房市走向。1.3 研究說明是在說明研究方法，從資料的收集、整理到呈現的工作流程。1.4 預期貢獻列了本論文的主要貢獻，包括房價推估及多源房地產買賣資訊收集的機制。最後，1.5 為章節介紹。

1.1 研究動機

1.1.1 實價登錄

台灣政府行政院於 2012 年 8 月 1 日公布實施「不動產成交案件實際資訊申報登錄制度」，簡稱實價登錄，做為建置房地產交易開放資料平台的法源依據，期望透過透明的交易資訊平抑飆漲的房價，並為未來可能施行的實價課稅預作準備。所謂實價登錄制度是指不動產交易買賣雙方必須依據政府所指定的表格填寫交易相關資料，包含房屋住址、房地交易總價、建物格局...等，當中的不動產交易係指買賣，而繼承、贈與雖有轉移但無金錢流動，故無須申報。其目的是為了讓房地產交易符合所謂的「公平、公正、公開」三公原則，提供房市交易價格資訊給一般民眾，有助於改善單一購屋人與廣大房市中「資訊不對等」的情況。

然而，政府的一番好意並未如其所望。實價登錄固然提供民眾一個公開且具有一定公信力的管道了解房市，但因政策不夠健全完整又缺乏具約束力罰則，造成有虛報的情況。尚因相關的例外情況，未能完整及正確的呈現房市資訊，甚至讓有心人士主導房市走向助漲房價。政府好意的實價登錄也因而有了嘲諷意味濃厚的新名稱─虛價登錄。

1.1.2 實價登錄中的角色

在實價登錄申報過程當中，主要參與人有買賣雙方、房屋仲介以及地政士。賣方將資產轉移給買方獲得金錢者，而買方則是付出金錢獲得資產的人。仲介的工作是夾在買賣雙方中間提供雙方相關的交易資訊、提供一些看房、介紹服務，成交後提供履約保證，依各家房仲行情抽取傭金。地政士也就是俗稱的土地代書，扮演著將買賣雙方提供的資料上傳登記的重要角色。民國 90 年通過的地政士法正式將土地代書正名為地政士，必須通過國家考試才能開業。

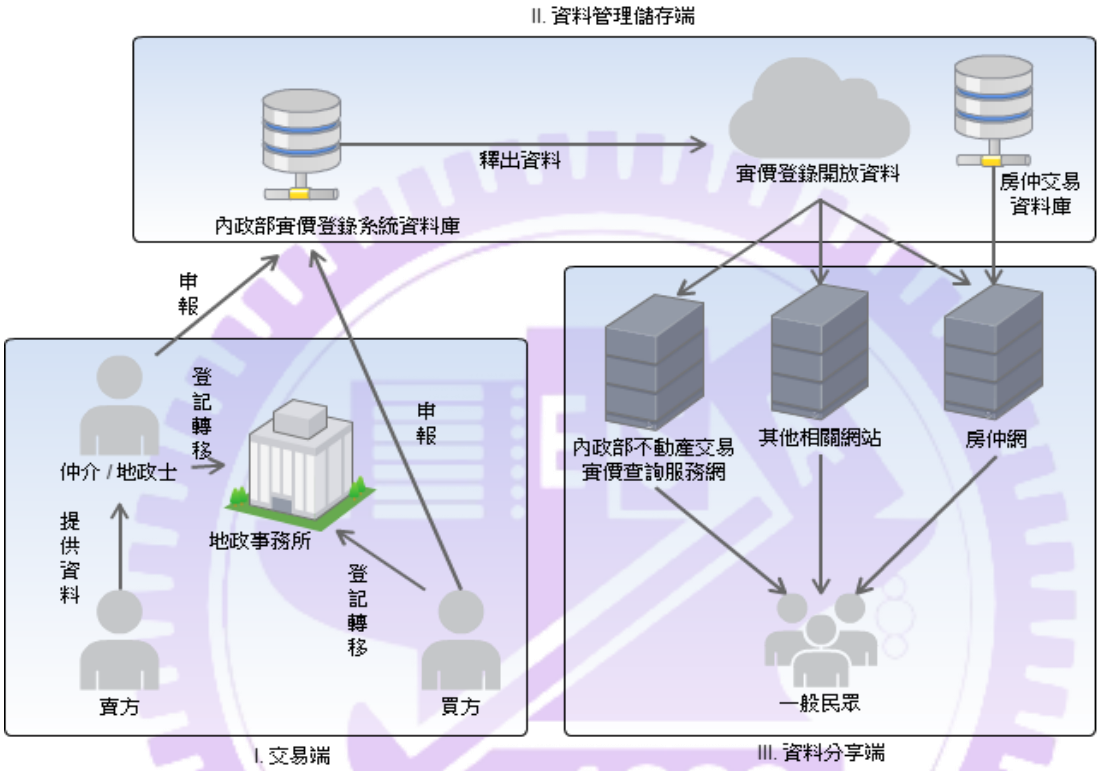
實價登錄相關的流程可參考 Figure 1 中所示之情境示意圖，主要分為交易端、資料管理儲存端及資料分享端。

交易端也就是在房地產交易過程的部分，若有經房屋仲介或有請地政士代簽契約的話，應由房屋仲介或地政士在交易完成後向地政事務所登記轉移，一併申報實價登錄。若未經仲介或地政士，則移轉確定後，應由承買人會同出賣人向不動產所在地地政事務辦理所有權移轉登記及申報實價登錄，但已建置完成的不動產成交案件，30 天內須實際至資訊申報登錄系統網站，一般民眾可透過自然人憑證登錄，公司行號（地政士、不動產經紀業者）則可透過工商憑證登錄。在交易資料登錄前，需確定各欄位資料的正確性，不單單只有交易價格，還有土地面積、建物面積、房屋格局等等各式各樣的資料。

資料管理儲存端負責資料的彙整、儲存及公開，主要分為內政部的時價登錄資料及一般民間房仲交易資料。實價登錄資料登錄至內政部實價登錄系統資料庫後，公部門會對資料進行處理，主要有兩個部分：第一，將交易價格最高與最低各 5%去除，去除極端值的目的是為了防止炒作。第二，將地址區段化，其目的為保護個人隱私。內政部將處理為的資料已 open data 的方式釋出，提供一般民眾下載。房仲交易資料則由各房仲業者管理及應用，為不公開的資料。

資料分享端是將各式資料以視覺化或是原始檔案的方式提供給一般民眾瀏覽、下載或使用。實價登錄方面除了內政部本身的不動產交易實價查詢服務網外

其他相關網站，如迪歐地圖[1]就是以實價登錄公開的資料為基礎建立的房價地圖。一般房仲網站也會將實價登錄的交易資料做簡單的呈現，提供想購屋的民眾一些參考的依據。



1.1.3 資料的失真

樂屋網的文章【樂屋講座】揭露放大鏡下的「實價登錄」[2]提到目前「實價登錄」政策的漏洞，如：建設公司自建自售不用登錄、不透過仲介之租賃案件不登錄、代銷公司登錄時間落差大，部分代銷公司為了逃避登錄會將契約終止期延長，會將代銷契約拉長以回避申報登錄等等。

再者，政府課稅方案非以買賣價格來課，而是以公告現值張課稅基礎。好房News「實價登錄真能反應房子價值？」[3]中提到實價登錄是還原真實房價的過渡手段，未來的實價課稅如何落實是一大關鍵。

以上兩則新聞提到了幾個造成實價登錄資料不完整的原因，而什麼叫做資料的不完整呢？資料的不完整意旨原始資料無法忠實呈現當時的交易狀況。以下列出可能造成資料不完整的幾個原因：

1. 內政部公部門在公開資料前將交易價格最高與最低各 5%去除。
2. 同筆交易中含兩個或兩個以上標的，在登錄時只登錄第一個標的地址，其餘不需登錄。
3. 作業疏失，如買方不熟悉房地契約內容而申報有誤或地政士或仲介未確實查證賣方所提供資料的正確性。
4. 買賣雙方刻意提高或壓低價格，使交易不具市場代表性。

前三點為法規或是少數疏失，較難改變及避免，而針對第四點則可深入探討造成買賣雙方刻意提高或壓低價格的原因。

1.1.3.1 低報交易價格

與不動產相關的稅收項目有房屋稅、地價稅、房屋財產所得稅、奢侈稅、空地稅、增值稅等等，雖然政府並沒有明確的針對實價登錄進行課稅，但實價登錄無疑是讓政府能更精準的課稅。因此，買方為逃避納稅可能串通賣方將交易價格低報。

1.1.3.2 高報交易價格

高報交易價格的理由除了外，還可以向銀行貸到比較高的貸款。在 Mobile01 論壇中也有人討論高雄的橋頭建案，討論串中提到同一建案在實價登錄中卻明顯呈現三種不同價位且價差高達 600 萬元。其中可能的原因是建商在建案初期為了續建的資金問題以自己已超出行情的價購買並登記在實價登錄中，一方面能夠提供銀行貸款的依據，另一方面也能在實價登錄中提高自己建案的行情，讓不知情的民眾買貴。

1.2 研究主題 –房價模型建構

有了大量的房價交易資料後，我們除了可以做視覺化呈現，讓一般民眾更輕易的了解房市資訊外，我們對資料還有一些不一樣的期待。既然我們有了這些歷史資料，鑑往知來，我們可以預測未來的房價。

為了達到預測房價的目的，我們以實價登錄資料為基礎以線性迴歸的方式建立了房價模型。用了交易資料中許多屬性，如交易坪數、房屋格局、建物樓層等做為預測依據。在建構模型的過程中會對使用的實價登錄資料稍做篩選，如對各欄位剔除前後共 5%的極端值，使模型能預測得更加準確。

建立房價模型不但能夠用來預測房價，還能用預測的方式，先預估相同條件的房子價格，找出大量資料中的價格異常的資料。房價模型也能提供政府另一個瞭解房市的方法，也能當作之後課稅的參考。因此，房價模型的建立的很重要的。

1.3 研究方法

拿到實價登錄資料第一件事是要對資料做分析、統計、視覺化的呈現。由於實價登錄是長期累積的資料，因此對於資料的維護和追蹤也是很重要的。對資料有初步的認識後，房價的推估是下一步，利用分群及線性規劃等技巧推算房價。

1.3.1 異源資料收集

除了內政部釋出的實價登錄資料以外，我們還蒐集其他房仲網站等公開的市場資訊，讓參考的資料更全面。其他會影響房價的因素也是重要的參考依據，主要可以分成三大類：交通方面有捷運站位置；嫌惡設施方面有公墓、靈骨塔、殯儀館、火化場、焚化廠及垃圾掩埋場；還有其他像是中小學學區、醫院、郵局 ATM 等。工作重點包含資料收集、前處理及資料庫建置。

資料收集：異源資料的收集與維護可分成前中後三期，分別為資料的獲得、資料庫的建置以及自動化的維護流程。前期資料的獲得除了一些開源資料外，其他需以網路爬蟲將網頁內容擷取下來，如房仲網的資料即為非開源但公開的資料。中期為建立資料庫，大部分網頁資訊為半結構化資料，對資料的遺失或空缺需加以處理。後期的目的為建立長期資料收集的作業流程，由於政府提供了我們長期

穩定的資料來源，而好的資料維護、追蹤及更新才能將資料呈現出長期的趨勢，因此需要一套自動化的流程來定期更新維護資料庫。

資料前處理：擁有可運用的資料後下一步是將資料呈現出來，但在呈現資料前還需要一些前置處理，如錯誤資料修正、資料型態整理、異常資料排除等。完成前置處理後呈現的資料較具參考價值。

資料庫建置：為了方便日後大量資料的存取及維護，資料庫的建立是必要的。由於資料來源不同，特性也有一定的差異，因此我們將經過前處理的資料依來源分類，建立其專屬的資料庫。

1.3.2 資料視覺化呈現網站

使用者介面：而呈現方式以網頁為主，主要使用 d3.js 做資料的視覺化，呈現以價錢、坪數、樓層、建物種類等等屬性為主題的統計資料。

統計資料呈現：以 google map 呈現房價地圖。提供以縣市和鄉鎮市區的查詢，選擇想了解的區域按下「搜尋」，地圖上即會顯示該區域的房價資訊。隨著地圖的縮放，價錢資訊會以平均價格或是單一價格呈現。點擊單一價格會出現該筆交易的詳細資訊。

互動房價模型分析介面：以台北市為預估目標，參考使用者輸入的相關條件，如坪數、屋齡、樓高等，由線性迴歸模型建構房價模型做預估，除了呈現預估房價外，針對不同區域建構的房價模型會呈現價格預測圖及其區間統計圖。

1.3.3 Linear Regression 房價模型建立

特徵分析: 除了資料的分析、整理、呈現，我們利用實價登錄的資訊，如建物面積、房屋類型、建材、樓層等，配合特徵模型及課題模型等以迴歸方法來建立房價模型。為了讓模型預估的更精準，我們利用分群演算法再建構房價模型前先將資料分群，在對每一群建構出專屬的房價模型，進一步改善。房價模型除了可用來估算房價外，對於未來長期的觀察房價趨勢也很有幫助。

房價模型：本篇以政府提供實價登錄資料的特定欄位如鄉鎮市區、建物型態、

屋齡等以行政區進行初步的分群。再配合 Linear Regression 建構出該區的房價預測模型。

1.3.4 模型誤差分析

房價模型的預測準確度我們用線性迴歸的誤差 SSE (Sum of Square Error) 來分析。誤差分析除了用來檢測模型準確度，還可以用在資料的篩選。再篩選方面我們做了行政分區及去頭尾極端值，不論是分區或去除極端值都能有效降低 SSE。

1.4 預期貢獻

本研究的貢獻可分成三大部分：第一，資料收集機制的建立；第二，視覺化資料的呈現；第三，房價模型的建構。

1.4.1 資料收集機制建立

要做資料呈現之前必須要有穩定資料來源。因此本篇制訂了一套資料收集的標準作業流程，讓系統能呈現最新的資料。資料收集機制主要包含資料庫建置、定期資料更新、資料整理呈現這三大步驟。

1.4.2 視覺化資料呈現

收集完的資料沒有透過視覺化的呈現終究還是資料，人們很難從中獲得資訊。我們將資料做簡單的統計，透過網頁的方式讓使用者看到經過視覺化呈現後的資料，輕易的從中獲得想了解的資訊、看到一些重大事件對房市交易的影響或是房市的整體走勢。

1.4.3 房價模型建構

在房地產交易的過程中，買方為了不想買貴會想了解自己中意的房屋價位在哪，而賣方若也能充分了解自己房屋的特性及價格不但可以讓交易更快速順利，也不怕自己開價太低而少賺一筆。除了買賣雙方，政府若能掌握房地產交易的金額，對未來實價課稅也有一定的幫助。

因此，房價推估是很重要的。我們將經過前置處理的實價登錄資料當作來源，對每個區域建構出該區的房價迴歸模型。使用者可以輸入自己對房屋的需求，如坪數、樓層、格局等，經過指定的模型推算出房價。

1.5 章節介紹

本文提供了一套收集台灣房地產資料的流程與方法，並能以現有的資料建構出房價模型，針對不同的條件推算出價格。

在第二章中，我們將探討一些有關實價登錄相關的政策、新聞，房價模型的相關研究，以及在做資料視覺化的重要性為何及資料呈現方面的限制。

第三章會針對內政部釋出的實價登錄資料的來源、格式集呈現前需要作的前處理做介紹，並介紹資料庫設計的想法及概念，也完整描述資料庫建置、資料匯入的流程。第四章擴大收集更多源的房地產資訊，以房仲網為目標收集其代售房屋資訊，能對台灣房價資料有更全面的了解。

第六章資提到資料視覺化，也就是網站的架設及各網站功能、圖表的說明。第七章解說了房價模型的建置、SPSS 的使用及利用 SPSS 分析實價登錄資料的結果。第八章介紹了資料收集的詳細流程。第九章為結論。

chapter 2 相關背景介紹

本章將介紹房價模型研究的相關的背景知識，以及實作及數據分析會使用到的工具，包括：2.1 實價登錄相關的新聞及政策；2.2 建構房價模型的相關研究；2.3 系統環境及限制，包含系統架構、google map 的限制、動態程式架構及 d3.js 視覺化呈現函式庫；2.4 統計分析軟體 SPSS 的介紹。

2.1 相關新聞、政策

隨著記載時間的增長，從實價登錄資料能觀察到一些有趣的現象。今日新聞在「房價年漲逾 1 成-UNIQLO 設點拉抬周邊房價」[4]中提到 UNIQLO 設點後，對周邊房價產生較大漲幅的，由 102 年 10 月新北市新莊區中正路 52 號優衣褲漲幅最為明顯，實價登錄從 101 年每坪 36 萬元，漲至 102 年的 39.5 元，漲幅 9.7%。

而代書在實價登錄中扮演的角色一直讓外界質疑，政府也想從將資料真實性的責任歸咎於代書，但代書們並不這麼認為。中時電子報在「代書：公親變事主，不公平；建商：假資料炒房，想太多」[5]中提到台北市地政士公會理事長張義權認為，交易實價唯有買賣雙方或仲介最了解，地政士（昔稱代書）扮演的角色則是「代理」申報與登錄。若要求地政士成為第一順位申報義務人，登錄有誤立刻受罰，簡直是公親變事主！華固建設總經理洪嘉昇表示，地政士每接一個房屋交易案，只賺幾千元，而房價動輒上千萬元，中間的酬勞、對價關係差太多了，故意造假、虛價登錄的可能性微乎其微。

實價登錄相關法規即實價登錄地政三法包含不動產經紀業管理條例、地政士法以及平均地權條例之修正條文。重點包括：

1. 申報登錄時機：權利人或地政士或不動產經紀業者應於買賣案件辦竣所有權移轉登記 30 日內，向主管機關申報登錄土地及建物成交案件實際資訊。
2. 相關處罰規定：違反申報登錄土地及建物成交案件實際資訊義務，將處

以 3 萬至 15 萬元罰鍰。

3. 施行日期：考量修法後施行日期之銜接，爰增訂修正條文之施行日期，由行政院定之規定。行政院已核定自 101 年 8 月 1 日起施行。
4. 提供資料區段化：登錄之資訊，除涉及個人資料外，得供政府機關利用並以區域化、去識別化方式提供查詢。

2.2 房價模型介紹

估算房價最常見的方式是用特徵價格法，在智庫百科的解釋如下[6]：特徵價格法又稱 Hedonic 模型法和效用估價法，認為房地產由眾多不同的特徵組成，而房地產價格是由所有特徵帶給人們的效用決定的。由於各特徵的數量及組合方式不同，使得房地產的價格產生差異。因此，如能將房地產的價格影響因素分解，求出各影響因素所隱含的價格，在控制地產的特徵 (或品質) 數量固定不變時，就能將房地產價格變動的品質因素拆離，以反映純粹價格的變化。

特徵價格法的基本思路是將房地產商品的價格分解，以顯現出其各項特徵的隱含價格，在保持房地產的特徵不變的情況下，將房地產價格變動中的特徵因素分解，從價格的總變動中逐項剔除特徵變動的影響，剩下的便是純粹由供求關係引起的價格變動。

在「住宅價格指數之研究---以台北市為例」[8]一文中使用特徵價格模型建構台北市的房價模型，資料是由太平洋房屋公司所提供其 77～82 年委託交易成交之住宅案例，台北市的部分共 4,328 筆。該篇將特徵分成戶的特徵、棟的特徵、鄰里環境特徵及其他個體特徵四類，戶的特徵、棟的特徵、鄰里環境特徵及其他個體特徵。戶的特徵包含登記總面積、所在樓層及衛浴設備套數；棟的特徵包含屋齡及地上總樓層數。

研究結果影響住宅價格最顯著的是登記總面積，這和預期相符，也反應我國住宅消費注重面積的情形；另外，所在樓層也如預期，呈現二次曲線的變化，也就是一樓價格最高，三至五樓價格最低，隨著所在樓層的增加，房價也逐漸升高。

衛浴設備套數幾乎每年都呈現相當顯著的結果，且與登記總面積的線性重合亦不顯著，顯見衛浴設備套數仍反應除而積以外的其他原因。就屋齡來看，平均每年折舊 3.3 萬～4.3 萬或房價的 0.47%～1.53%。地上總樓層數也呈現建物造價對住宅價格的影響是相當顯著的；而僅次於登記總面積影響住宅價格最顯著的是區位，平均來說，舊市區的住宅總價要較新市區的住宅總價貴 146～201 萬或房價的 13.53%～36.11%。

住宅學報中「不動產自動估價與估價師個別估價之比較—以比較法之案例選取、權重調整與估值三階段差異分析」[10]文中提到，複迴歸模型是最普遍被應用的大量估價模型，其以特徵價格理論為基礎。而其對於自動估價系統預測結果的衡量標準主要有二項，第一個是平均絕對百分比誤差(mean absolute percentage error, MAPE) ，觀察整體誤差絕對值的統計量，若平均絕對百分比誤差越小表示其估價表現越好；第二個衡量標準為命中率 (hit-rate)，計算各個測試樣本估值與原始成交價格的差距，並觀察誤差在誤差範圍內的命中次數比例是否達到標準，可估計命中比例。

而在「台灣地區特徵性房價函數估計係不一致問題之探討 台灣地區特徵性房價函數估計係不一致問題之探討」[12]一文中提到在台灣地區，傳統上多以特徵性函數方式估計房價函數，並固定住宅品質但卻忽略了不同地區與時間之特徵性一定相同的情況。也就是說，各地區房價可能會隨著域特性與時間發展而影響房價函數的估計參。故利用行政院主計處「住宅狀況調查」資料，找出在影響房價的變數中何者之估計係數會隨著時間或地區的不同而有所變化，繼修正房價函。

除了一般的迴歸模型，也有人使用半參數法來預估房價。「半參數法於國內不動產大量估價之可行性評估」[11]一文中提到半參數法可更正確衡量住宅屬性與房價之關係，提高估價命中率。然而，可能受國內不動產成交資料的變數不完整與資料品質影響，半參數法對我國不動產房價預測的改善幅度未如國外明顯，

建議未來可加強不動產交易個案的區位、景觀、以及交通可及性等變數的收集。此外，個別區域的預測結果明顯優於整體台北市，以 BRUTO 方法選模時各區域間變數的函數型式亦有差異，顯示未來在建構房價的大量估價模型時，區域範圍界定宜採用小範圍劃分，部分變數的函數型式可採用非線性設定為佳。

2.3 資料視覺化

網站 Desiring clicks 中的【資訊視覺化】資訊視覺化與工程的應用[7]一文中提到視覺化在工程上的應用。

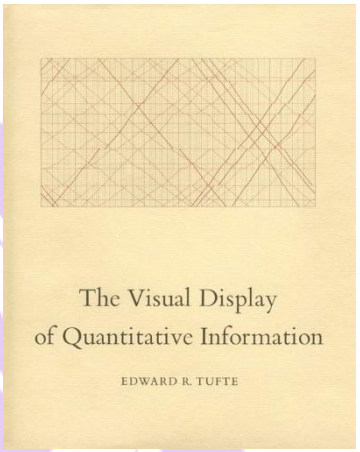


Figure 2 Book Cover: The Visual Display of Quantitative Information

甚麼是視覺化？幾十台火車的同時運行與交錯路線要怎麼同時表達在一張圖上？在 Edward R Tufte 所著的 The Visual Display of Quantitative Inforamtion 中，封面上的圖表(Figure 2 Book Cover: The Visual Display of Quantitative Information)就是個很有趣的資訊視覺化的例子。圖上橫軸代表時間，縱軸代表距離。斜度越大代表該班列車速度越快，由左上往右下的是南下列車，往右上往左下的則代表北上列車，而斜線斷掉的地方表示列車暫時停靠在該火車站。乘客可以清楚的從橫軸上的座標找出自己所在的城市，然後快速的估算出到目的地要花多少時間，需要怎麼轉車。只要畫面上的斜線和橫線相交，基本上代表乘客可以輕鬆的規劃出轉車路線。

在歷史發展的過程中，自然與工程是一向非常重要的發展指標。無論是颱風的資料分析、降雨資料分析、到建築物的結構力學分析與呈現，都需要良好的資訊視覺輔助，相關專家才能夠正確的判斷，並做出正確的決策。許多細節的惡魔都隱藏在巨量的資料裡，透過正確的視覺輔助，可以大量的降低風險，提高居住以及生活的安全以及便利性。

2.4 系統環境

本節將介紹系統環境，包含 MySQL 資料庫、統計分析軟體 SPSS、網路爬蟲及網站伺服器。系統會連接到內政部實價登錄服務網下載實價登錄資料；連到各大房仲網站，經由網路爬蟲將房仲網資料擷取下來。同時開放網路伺服器供使用者以瀏覽器在各地連線，查看本系統網站：多源房價資料分析網。如 Figure 3 系統架構圖。

本系統分成網站伺服器、網路爬蟲、資料庫及統計分析四個部分，如 Figure 3 系統架構圖左下方框。網站伺服器包括以 PHP 撰寫的網站框架 CodeIgniter、CSS 介面優化的 BootStrap 及資料視覺化圖表函式庫 D3JS。網路爬蟲為以 PHP 撰寫的 HTML 檔擷取程式，搭配 linux 的 bash 檔執行。資料庫選用 MySQL，為免費但功能健全的資料庫系統。最後是統計分析軟體 SPSS，負責將收集到的資料最參數分析、模型建置，完成房價模型，達成房價預估的任務。

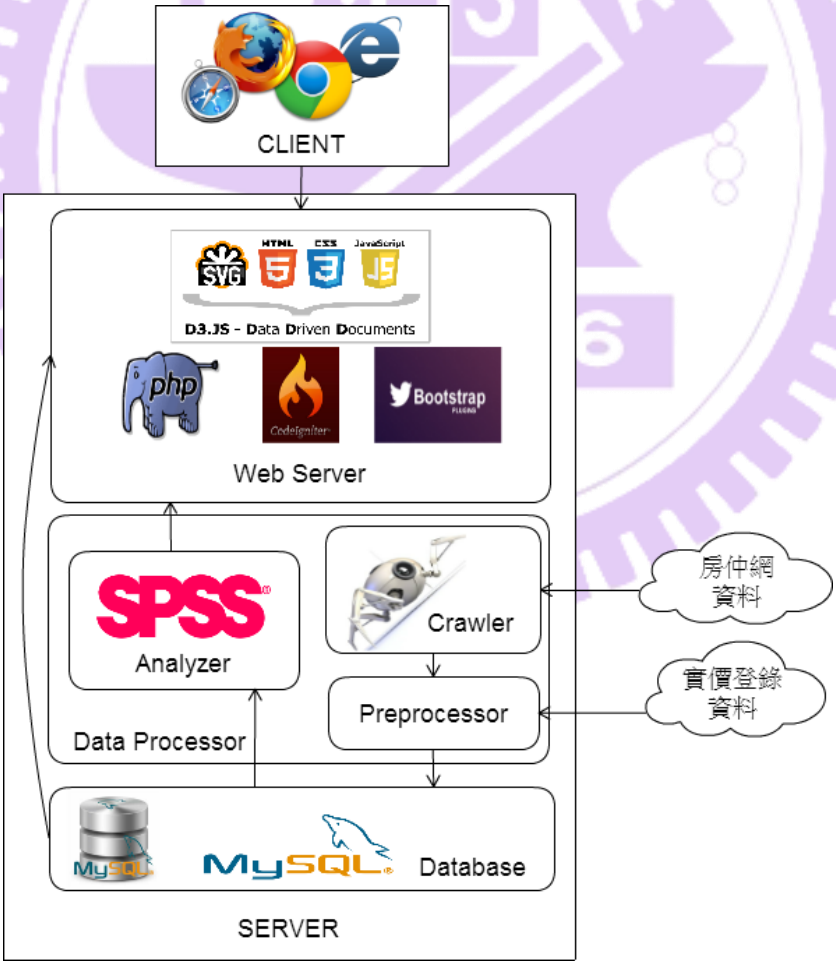


Figure 3 系統架構圖

2.4.1 Google Maps 的限制

網頁地圖呈現我們已 Google Maps JavaScript API 第 3 版來實作。Google Maps 是著名且被廣泛使用的免費地圖之一。基本的地圖呈現、標上標籤、多個標籤合併都有相對應的 API 可以使用，算是成熟度相當高的地圖，這也是我們會優先考慮使用 Google Maps 的原因。

但免費的難免有些限制，如每日最多載入 25,000 次地圖、Google Geocoding API (如地址轉座標)，每天只能要求 2,500 個地理位置。當然除了 Google Maps 之外，還有其他免費的開放地圖如 Open Street Map、MapBox 等等。

2.4.2 系統框架及工具

系統網站框架以 CodeIgniter 為骨幹，前端搭配 BootStrap 做 CSS 的美化。CodeIgniter 是一套小巧但功能強大的 PHP 框架，採用 MVC 的開發模。MVC ，model-view-controller 將網站運作分成三大部分，model 是進行資料管理和資料庫設計的地方，view 是圖形介面設計，而 controller 是負責轉發、處理請求的地方。將前後端做區分，方便日後程式的維護及管理。BootStrap 則是負責網頁美化的工具，主要與 CodeIgniter 的 view 結合。可以挑選想要的主題內部其他元件也可以很彈性的擴充。

系統會顯示許多不同的靜態的圖表、互動式資料呈現或是房價地圖在使用者介面，這些就是要靠 d3.js。d3.js，Data-Driven Documents，為一套強大的 JavaScript 函式庫，他通過使用 HTML、SVG 和 CSS 在網頁上展示數據讓我們可以很容易的將資料與文件組合起來，並加以操作。

本篇資料庫選擇開放原始碼的 MySQL，架設在本機端。搭配 phpMyAdmin 操作控制。用 PHP 銜接到系統，再用 JavaScript 做視覺化呈現。

2.4.3 統計分析軟體 SPSS

SPSS 統計套裝軟體，全名為 Statistical Package for Social Science，是 IBM 公司推出的一系列用於統計學分析運算、數據挖掘、預測分析和決策支持任務的軟體產品，本篇使用 SPSS Statistics 22.0。

SPSS 提供一個友善的使用者介面，如 Figure 4 SPSS 操作介面，可以透過滑鼠的點選和拖曳，輕鬆地完成資料的讀取、分析和產出報表，而廣泛地應用於商管、心理、教育、農業、醫學、金融界...等等。

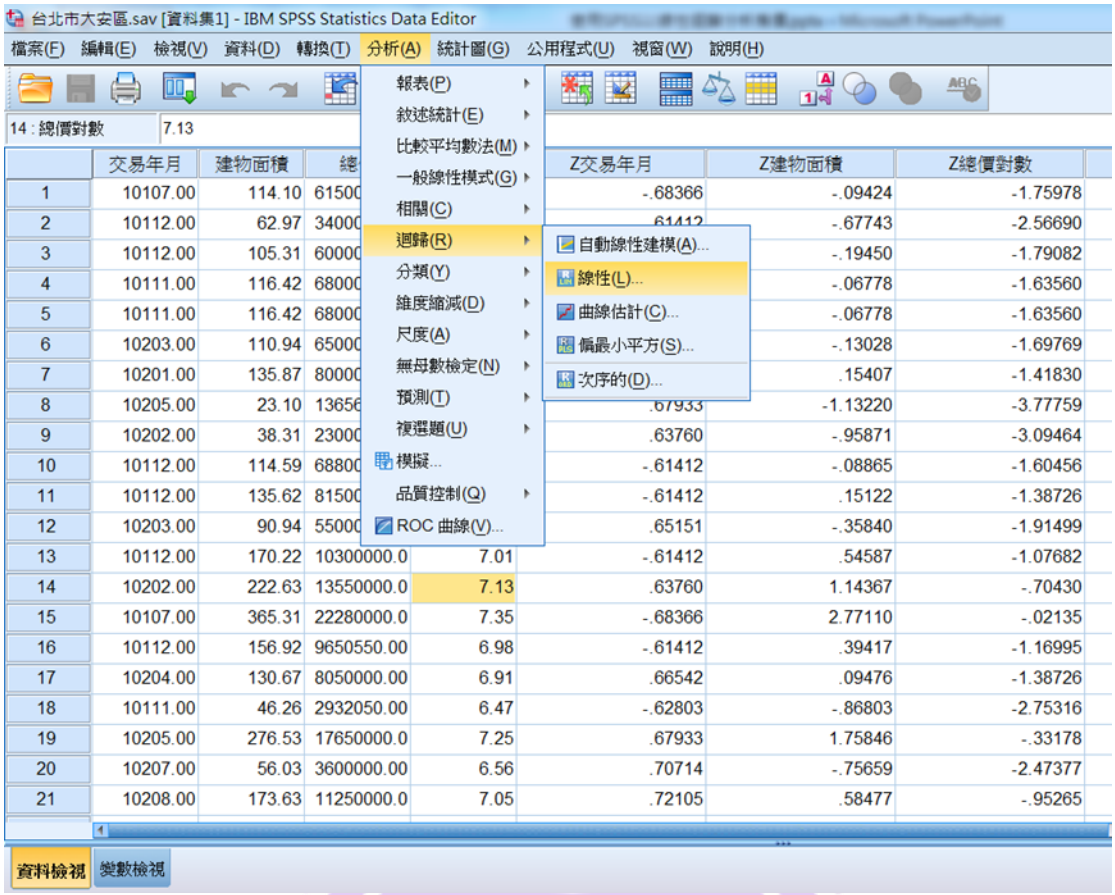


Figure 4 SPSS 操作介面

chapter 3 實價登錄資料庫設計

本章將介紹資料庫的設計理念，配合行政院內政部釋出的實價登錄資料欄位內容設計了專屬的資料庫。內容有 3.1 實價登錄的資料介紹；3.2 為實價登錄設計專屬的資料庫；3.3 資料表格式瀏覽；3.4 資料匯入流程介紹。

3.1 行政院內政部實價登錄

3.1.1 實價登錄內的資料來源

實價登錄每份檔案中包含了 28 個欄位，如表格 1 實價登錄實際資料。其中檔案名稱中有包含縣市資訊，因此對同一筆資料可參考的屬性欄位有 29 欄。當然，資料可能會有空白，要使用的話還是需要經過處理。

表格 1 實價登錄實際資料

鄉鎮市區	文山區	中正區	文山區	中山區	信義區
交易標的	房地(土地+建物)+車位	車位	土地	房地(土地+建物)	建物
土地區段位置/建物區段門牌	臺北市文山區景美街1~50 號	臺北市中正區廈門街 147 巷 1~50 號	政大段二小段 451~500 地號	臺北市中山區長安東路一段 53 巷 1~50 號	臺北市信義區松隆路 101~150 號
土地移轉總面積(平方公尺)	6.58	1.69	1.69	6.42	0
使用分區或編定	商	住	住	商	
交易年月	10207	10207	10207	10207	10206
交易筆棟數	土地 3 建物 1 車位 1	土地 0 建物 0 車位 1	土地 2 建物 0 車位 0	土地 2 建物 1 車位 0	土地 0 建物 2 車位 1
移轉層次	八層	一層		二層	三層
總樓層數	011	012		007	010
建物型態	住宅大樓(11層含以上有電梯)	其他	其他	套房(1 房 1 廳 1 衛)	套房(1 房 1 廳 1 衛)
主要用途	住家用	見其他登記		住家用	商業用

		事項			
主要建材	鋼筋混凝土造	鋼筋混凝土造		鋼筋混凝土造	鋼筋混凝土造
建築完成年月	0980116	0930514		0761223	0850925
建物移轉總面積(平方公尺)	61.7	16.8	0	36.69	182.27
建物現況格局-房	1	0	0	1	1
建物現況格局-廳	1	0	0	1	1
建物現況格局-衛	1	0	0	1	1
建物現況格局-隔間	有	有	有	有	有
有無管理組織	有	有	無	有	有
總價(元)	7000000	1160000	10000	4430000	10200000
單價(元/平方公尺)	113452		0	120741	55961
車位類別	升降機械	坡道機械			坡道平面
車位移轉總面積(平方公尺)	0	16.8	0	0	43.75
車位總價(元)	0	0	0	0	0
交易標的橫坐標	304741	302360	308671	303131	307723
交易標的縱坐標	2764927	2768275	2763996	2771388	2771087

3.1.1.1 取得途徑

由實價登錄網站[9]下載，下載的資料為兩個月前兩周內的資料，分為買賣、租賃及預售三種類型，內容包含表格 1 實價登錄實際資料的欄位。另一種途徑是直接向內政部購買，付費標準是以量計價。

3.1.1.2 資料格式

下載資料時可選擇以 xml、csv 以及 txt 格式下載。除檔案的格式還可以選擇要下載的資訊，有全國 (含不動產買賣+預售屋買賣+不動產租賃) 以及進階下載 (勾選欲下載 縣市/交易類別)。

對於申請人填寫的格式也有相關的規定，以下摘錄自實價登錄說明文件，關於特定欄位的填寫標準：

1. 建物門牌：如為房地、建物或車位成交案件，須填載登記 (簿) 謄本所載建物門牌。如成交案件僅有土地而無建物者，本欄無須填載。如成交案件之建物有多個門牌且未分開計價者，本欄填載建物面積最大之建物門牌，如面積相同填載序號在先之門牌。
2. 交易筆棟數：指買賣移轉登記實際交易之筆棟（戶）數及車位數，如成交案件為土地 2 筆、建物 1 棟、車位 2 個，依實填載。惟多筆多棟或 1 筆多棟之交易，如有個別交易價格者，應就每棟分別填載申報書。
3. 房地交易總價：房地交易總價係為土地交易總價、建物交易總價及車位交易總價之總計，倘如僅為土地或建物或車位之交易，除須於相對欄位填載價格外，本欄仍需填載價格。

   房地交易如未能拆分土地及建物之個別交易價格時，僅就該成交案件交易總價填載，如含車位則應計入車位總價，車位價格並應另行填載；但無法拆計車位價格者，則需勾選「車位未單獨計價，且已含入交易總價。」無車位交易則勾選「無車位交易」。
4. 土地交易總價：指房地成交案件內土地之交易價格或僅有土地交易之交易價格，土地未分開計價者免填本欄 (非填 0)。
5. 建物交易總價：指房地成交案件內建物（房屋）之交易價格或僅有建物交易之交易價格，建物未分開計價者免填本欄 (非填 0)。
6. 車位交易總價：指房地成交案件內車位之交易價格或僅有車位交易之交易價格，係為車位價格之加總。車位未分開計價者免填本欄 (非填 0)。

7. 車位資訊:「車位未單獨計價，且已含入交易總價」指房地、土地或建物成交案件內含車位之交易，但無法拆分車位交易價格者，請於本欄位勾選。無車位交易者勾選無車位交易。
8. 住址資料先下拉選擇縣市鄉鎮，後方空白欄位直接填寫里鄰或路名門牌。
9. 若為特殊交易無買賣價格,交易總價應填零,並於備註欄加註特殊交易種類，切勿自行填入公告土地現值或其他價格。

舉例來說，表格 1 實價登錄實際資料是在實價登錄原始.csv 檔中挑選了不同交易類型交易標的資料。此五筆資料的交易標的分別為房地+車位、車位、土地、房地及建物。可以看到土地區段位置/建物區段門牌欄位中土地的地址為地號，其餘為我們常用的地址。土地轉移欄位建物為 0。使用分區土地為空值。交易筆棟數欄位在建物部分為土地 0 建物 2 車位 1，也就是其實是建物+車位，由此可知交易筆棟數的資訊筆交易標的資訊更完整且詳細,因此再匯入我們將以交易筆棟數取代交易標的。接下來幾個欄位都是跟建物相關，土地幾乎都填 0 或空值，在建物型態填其他，而建物現況格局-隔間居然是填有，由於交易是土地，所以此欄位並不會採計。在平均單價的欄位，車位是空值，而土地是 0。車位類別欄位在交易筆棟數中有車位的交易均有值。但車位移轉總面積欄位中交易標的為房地+車位的卻填 0，而車位總價欄位均為 0。座標的欄位均有值。

觀察表格 1 實價登錄實際資料可知實際登錄的資料並不是完全照著規定填寫登錄。因此，在資料庫設計時我們以能詳細記錄資料原貌為宗旨，盡量將原始資訊完整保留。

3.1.1.3 資料更新

由實價登錄網站下載更新，兩個禮拜會更新一次，但釋出的資料內容為兩個月前的資料。由網站下載更新需在期限內下載完畢，否則更新過後就的資料無法取得。而向內政部購買，一樣兩個禮拜會有新資料。由於是以量計價，需累積至一定的資料量在購買會比較划算。

3.2 資料庫設計

配合行政院內政部釋出的實價登錄資料欄位內容，因應各屬性之間的關係，設計了七個資料表來存放並組織資料，分別為交易、房屋、建物、土地、車位、地址及座標。把原來 26 個欄位的.csv 檔適當的合併並匯入我們建置的資料庫中。

資料庫平台我們選用 MySQL。資料表的設計過程一開始先觀察實價登錄資料的屬性、欄位及架構，把相關連的欄位為抽離並獨立。建了交易、房屋、建物、土地、車位、地址及座標七個資料表，作最初步的正規化

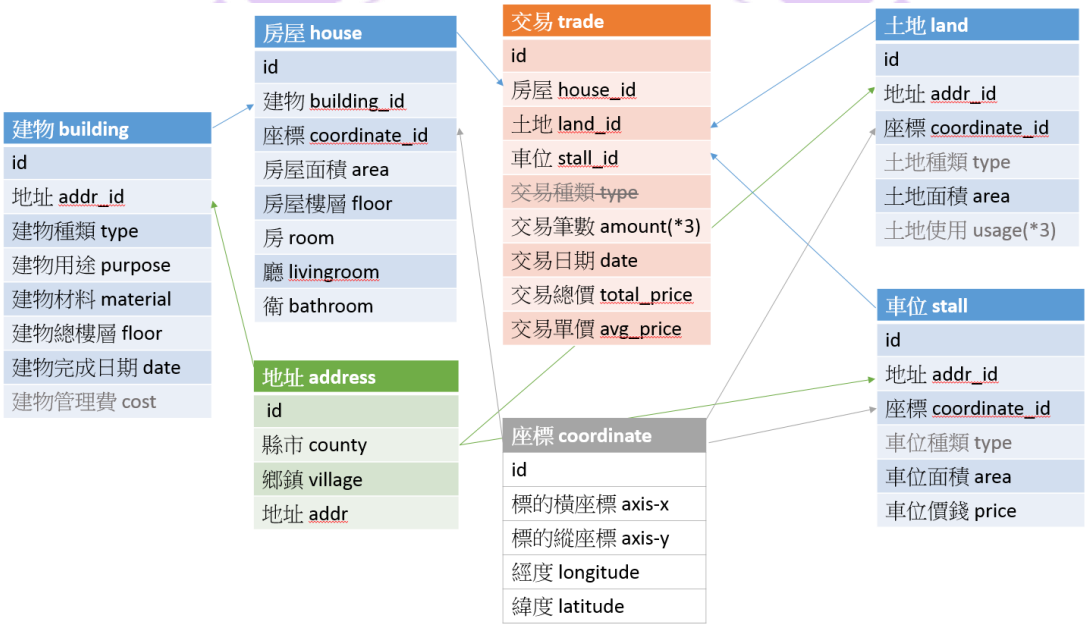


Figure 5 實價登錄資料表關係圖

但事實上並不會如此完美，因為實價登錄資料中有許多欄位可能空缺，會造成某些資料表 identify 的困難。但正規化後的資料庫可以節省一些不必要的空間浪費。

3.3 資料表呈現

參照表格 1 實價登錄實際資料分別建立此七個資料表如下：

表格 2 實價登錄-建物

建物 building			
	欄位名稱	型態	附註
	id	Int	Primary key
地址	addr_id	Int	
建物種類	type	Text/UTF-8	
建物用途	purpose	Text/UTF-8	
建物材料	material	Text/UTF-8	
建物總樓層	floor	Int	
建物完成日期	date	Int	
建物管理費	cost	Int	

表格 3 實價登錄-房屋

房屋 house			
	欄位名稱	型態	附註
	Id	Int	Primary
建物	building_id	Int	
座標	coordinate_id	int	
房屋面積	area	Float	
房屋樓層	floor	Int	
房	room	Int	
廳	living	Int	
衛	bath	Int	

表格 4 實價登錄-土地

土地 land			
	欄位名稱	**型態**	**附註**
	Id	Int	Primary key
地址	addr_id	Int	
座標	coordinate_id	Int	
土地面積	Area	Float	
土地使用	usage_urban	Text/UTF-8	
	usage_country_district	Text/UTF-8	
	usage_country_type	Text/UTF-8	

表格 5 實價登錄-車位

車位 stall			
	欄位名稱	**型態**	**附註**
	id	Int	Primary key
地址	addr_id	Int	
座標	coordinate_id	Int	
車位種類	type	Text/UTF-8	
車位面積	area	Float	
車位價錢	price	int	

表格 6 實價登錄-地址

地址 address			
	欄位名稱	**型態**	**附註**
	id	Int	Primary key
縣市	county	Text/UTF-8	
鄉鎮	village	Text/UTF-8	
地址	addr	Text/UTF-8	

表格 7 實價登錄-座標

座標 coordinate			
	欄位名稱	**型態**	**附註**
	id	Int	Primary key
標的橫座標	axis-x	Int	
標的縱座標	axis-y	Int	
經度	longitude	Double	
緯度	latitude	Double	

表格 8 實價登錄-交易

交易 trade			
	欄位名稱	**型態**	**附註**
	id	Int	Primary key
房屋	house_id	Int	
土地	land_id	Int	
車位	stall_id	Int	
交易種類	type	Text/UTF-8	
交易筆數	amount_house	Int	
	amount_land	Int	
	amount_stall	Int	
交易日期	date	Int	
交易總價	total_price	Int	
交易單價	avg_price	float	

3.4 資料匯入

資料庫建置與資料匯入整理流程包含資料前置處理及匯入資料庫兩部分。資料前置處理，包含異體字統一、異常單價及面積處理、座標錯誤的解決方案及各欄位的資料整理。匯入資料庫，資料匯入的流程包含建立資料表、及插入資料兩步驟。

3.4.1 資料的前處理

在正式進行分析之前必須先將手上的資料先做適當的整理。資料的預處理主要分為三個主要部分，依序為合併各縣市之.csv 檔、中文異體字統一及各欄位資料整理。

3.4.1.1 異體字

中文有需多異體字，指的是音義均相同，但字形不同的字。在實價登錄資料中也有出現異體字，在鄉鎮市欄位及地址欄位中出現了「台」與「臺」這組異體字，以及「市」與「市」。異體字的出現讓程式將相同縣市分成兩組，解決的方式為建立異體字表，以「台」和「市」統一。

3.4.1.2 異常單價、面積

雖然內政部的公部門已將實價登錄中交易總價最高級最低各 5%隱藏，但以平均價格來看因為建物總面積的關係還是有些異常的單價，如高雄市鳳山區有一筆交易的建物面積才 0.02 平方公尺，造成了每坪平均單價達到了 49950 萬元。

3.4.1.3 地址與坐標不一致

實價登錄資料中的地址雖然有經過內政部公布們的前置處理，地址為 50 號一個區段。我們還是發現在向內政部購買的資料當中有縣市、鄉鎮市區、地址均一致，但座標位置卻差了兩至三個縣市以上。很明顯的這是內政部再做地址與座標轉換時的疏失。目前解決方式為人工檢查，計算與縣市中心相差太遠的資料，將其隱藏不做顯示。

3.4.1.4 各欄位資料整理

實價登錄資料並非拿到後就可以立即使用，處理完上述合併檔案、解決了異體字問題，還需要將欄位的資訊轉換成我們想要的狀態。主要需整理地址及樓層這兩個欄位。

地址欄位有些有包含鄉鎮市，有些有包含縣市及鄉鎮市，有些則都沒有。統一方式為全部整理為「縣市+鄉鎮市+地址」。在過程中有發現地址中包含的鄉鎮市與鄉鎮市欄位中的值不合。推斷是在土地重劃後造成鄉鎮市變更，但舊有地址尚未變更。此情況不對地址做更動。

移轉層次欄位的問題較大。樓層資訊在實價登錄資料中分為兩個欄位，移轉層次及總樓層數。總樓層數是以數字表示因此不須另外處理。移轉層次則是以中文表示，且裡面可能有多個以中文逗號隔開的文字，除了表示單一樓層、多樓層還有其他的附屬條件，例如陽台、騎樓、車位等等。多樓層均為連續的樓層，為有效處理取其平均。其他附加條件則另成新的欄位，以 1 和 0 表示有或沒有該附加條件。

3.4.2 匯入資料庫

按照 Figure 5 實價登錄資料表關係圖建立 coordinate、address、stall、building、land、house 及 trade 七個資料表。再依序將資料按下述步驟匯入至 coordinate、address、stall、building、land、house 及 trade 七個資料表中，過程需檢查該筆資料是否已存在在資料庫中。若已存在則紀錄此 id，否則將資料 insert 進該資料表中。

chapter 4　房仲網 by crawler

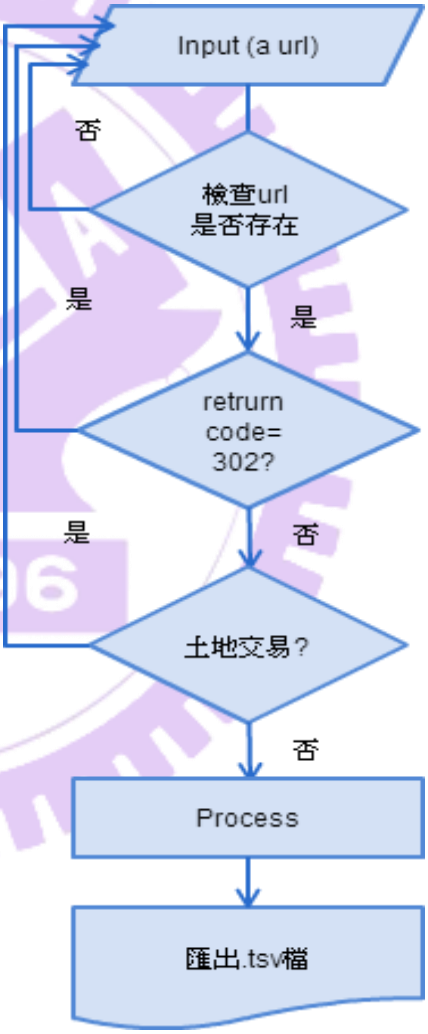
為了充分了解台灣房市，除了實價登錄資料外，房仲網也是一個了解房市資訊的管道。但不同於實價登錄為開源資料，想得到房仲網的資料得自行收集。本章將介紹我們是如何收集房仲網資料，4.1crawler 方法及流程；4.2 資料庫的設計及建置；4.3 相關問題。

4.1　Crawler 方法及流程

在著手寫 crawler 之前我觀察目標網頁 url 形式，發現房仲網的目標網頁 url 都有一個通性，就是只有末端的固定幾碼會變動，類似 index 的功能。因此以迴圈方式進行探索取代一般 crawler，配合 bash 檔自動化執行。

判斷 url 是否存在是用了 PHP 的 get_headers，並順便判斷 return code 是否為 302 redirection。在 crawling 有巢氏時若 return code 為 302，則該 url 會把網頁導至 list 頁面，而非目標網頁，因此予以排除。

確定目標網頁後就輪到 parsing 了，我是用 PHP Simple HTML DOM Parser 來分析網頁內容並將關鍵內容萃取出來。在過程中先判斷該筆資料為房地資料，若為土地交易資料則忽略不計。最後將從.html 檔中萃取出的資料寫進.tsv 檔中。再將格式進行整理，匯入資料庫。



4.2 相關問題

第一個遇到比較大的問題是有些房仲網會擋別人抓取 html 檔，如信義房屋。所以我先選擇了有巢氏房屋做資料收集，可能的解決方式可以對 http request 的 header 進行修改。

第二的問題是丟 request 的速度。由於資料量很大，希望能加快執行速度，但速度過快會造成對方 server 負擔太大，而被擋掉無法繼續，經過測試每次同一個 IP 丟的 request 間休息 0.5 秒最適當。當然不同 IP 的話不受此限。

4.3 資料庫設計及建置

有巢氏網頁資料收集主要有 20 個欄位，分別為總價、平均單價、地址、屋齡、樓別/樓高、房屋型態、主要建材、建物管理費、有無中庭、面前巷道、是否邊間、電梯數量、建物格局、主建物坪數、土地登記坪數、停車方式、車位狀況、停車管理費、車位管理方式。

為配合實價登錄資料作日後的分析比對，我們盡可能的將資料格式統一，再存入資料庫前做一些調整：樓別/樓高改為樓別與樓高兩個欄位及地址改為縣市、鄉鎮及地址。

總價	單價	房屋格局	建坪	地址	屋齡	樓層	擷取時間	url_index
2,920	34.04	7房3廳4衛	85.79	新竹縣竹東鎮杞林路	34	1～4 / 4	2014/5/21	150
498	13.69	3房2廳2衛	36.39	台中市沙鹿區保安路	9.7	3～3 / 8	2014/5/20	2012
1,480	18.27	7房2廳7衛	81	台中市沙鹿區北勢東路	20.3	-1～3 / 3	2014/5/20	2792
7,800	29.99	5房2廳6衛	260.1	台中市烏日區成功西路	4.9	-1～3 / 3	2014/5/20	4048
398	19.11	2房2廳2衛	20.83	台中市西屯區文華路	18.6	-- / 11	2014/5/20	4872
3,230	38.47	5房2廳6衛	83.97	台中市沙鹿區英才路		-1～4 / 4	2014/5/20	6035
628	17.79	3房2廳2衛	35.3	桃園縣平鎮市興華街	14.9	7～7 / 7	2014/5/20	6191
1,880	49.93	2房2廳2衛	37.65	新竹縣湖口鄉忠孝路	42.4	1～3 / 3	2014/5/20	6193
728	15.25	4房2廳2衛	47.75	台中市北屯區松和街	20	-- / 10	2014/5/20	7384
1,280	12.55	0房0廳0衛	101.97	彰化縣福興鄉彰鹿路七段	20.2	1～3 / 3	2014/5/20	7393

總價	單價	房屋格局	建坪	地址	屋齡	樓層	擷取時間	url_index
2,920	34.04	7房(室)3廳4衛	85.79	新竹縣竹東鎮杞林路	34.2	1 ~ 4 / 4	2014/7/10	150
1,480	18.27	7房(室)2廳7衛	81	台中市沙鹿區北勢東路	20.4	-1 ~ 3 / 3	2014/7/10	2792
7,800	29.99	5房(室)2廳6衛	260.1	台中市烏日區成功西路	18	-1 ~ 3 / 3	2014/7/10	4048
398	12.54	3房(室)2廳1衛	31.73	台中市太平區新平路一段	20.9	4 ~ / 8	2014/7/10	4102
628	17.79	3房(室)2廳2衛	35.3	桃園縣平鎮市興華街	15	7 ~ 7 / 7	2014/7/10	6191
1,880	49.93	2房(室)2廳2衛	37.65	新竹縣湖口鄉忠孝路	42.5	1 ~ 3 / 3	2014/7/10	6193
700	14.66	4房(室)2廳2衛	47.75	台中市北屯區松和街	20.1	-- / 10	2014/7/10	7384

由於有巢氏房屋的資料是代售房屋實價登錄資料為成交案件不同，且每個月收集的資料會有許多重複。如 Figure 8 有巢氏房屋七月資料收集及 Figure 7 有巢氏房屋五月資料收集，其中紅色的資料表示五月有收集到而七月時就消失下架了。綠色的表示五月收集時還沒有該筆資料而七月出現了。藍色表式 url_index 相同但七月實有部分欄位資料更動。

因此，在資料庫設計上多了收集資料的時間、資料狀態與網址標號，分別記錄收集該筆資料當下的時間、資料狀態：新上架、更新及已下架三種狀態、當作識別碼的網址末端六碼數字。以下是狀態介紹：

1. 新上架：出現新的建物者視為新上架
2. 更新：同建物，價格或其他條件有變動視為更新
3. 已下架：該網址標號網頁不存在者視為已下架

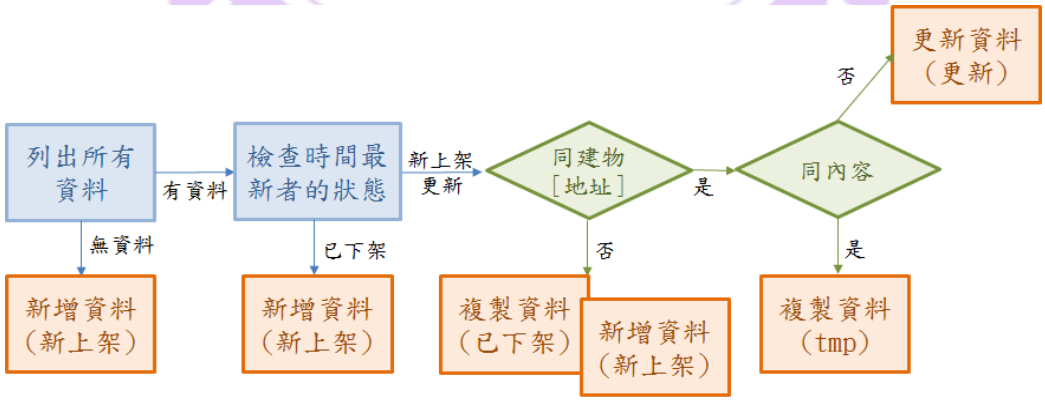


Figure 9 有巢氏房屋資料收集流程 part I

資料狀態都是以同一個網址標號當作物件來看。而同建物的判斷標準為縣市、鄉鎮、地址、面積相同者。作業流程如 Figure 10 有巢氏房屋資料收集流程 part I 及 Figure 9 有巢氏房屋資料收集流程 part I。整合完成後資料表如 Figure 11 有巢氏房屋整合後資料。

總價	單價	房屋格局	建坪	地址	屋齡	樓層	擷取時間	url_index	state
2,920	34.04	7房(室)3廳4衛	85.79	新竹縣竹東鎮杞林路	34.2	1～4 / 4	2014/5/21	150	新上架
498	13.69	3房2廳2衛	36.39	台中市沙鹿區保安路	9.7	3～3 / 8	2014/7/10	2012	已下架
498	13.69	3房2廳2衛	36.39	台中市沙鹿區保安路	9.7	3～3 / 8	2014/7/10	2012	已下架
1,480	18.27	7房(室)2廳7衛	81	台中市沙鹿區北勢東路	20.4	-1～3 / 3	2014/5/20	2792	新上架
7,800	29.99	5房(室)2廳6衛	260.1	台中市烏日區成功西路	18	-1～3 / 3	2014/7/10	4048	更新
398	19.11	2房2廳2衛	20.83	台中市西屯區文華路	18.6	-- / 11	2014/7/10	4872	已下架
398	19.11	2房2廳2衛	20.83	台中市西屯區文華路	18.6	-- / 11	2014/7/10	4872	已下架
3,230	38.47	5房2廳6衛	83.97	台中市沙鹿區英才路		-1～4 / 4	2014/7/10	6035	已下架
3,230	38.47	5房2廳6衛	83.97	台中市沙鹿區英才路		-1～4 / 4	2014/7/10	6035	已下架
398	12.54	3房(室)2廳1衛	31.73	台中市太平區新平路一段	20.9	4～ / 8	2014/7/10	4102	新上架
628	17.79	3房(室)2廳2衛	35.3	桃園縣平鎮市興華街	15	7～7 / 7	2014/5/20	6191	新上架
1,880	49.93	2房(室)2廳2衛	37.65	新竹縣湖口鄉忠孝路	42.5	1～3 / 3	2014/5/20	6193	新上架
700	14.66	4房(室)2廳2衛	47.75	台中市北屯區松和街	20.1	-- / 10	2014/7/10	7384	更新
1,280	12.55	0房0廳0衛	101.97	彰化縣福興鄉彰鹿路七段	20.2	1～3 / 3	2014/7/10	7393	已下架
1,280	12.55	0房0廳0衛	101.97	彰化縣福興鄉彰鹿路七段	20.2	1～3 / 3	2014/7/10	7393	已下架

Figure 11 有巢氏房屋整合後資料

chapter 5　資料視覺化

資料視覺化對資料分析者來說是相當重要的。有了資料之後第一步就是要適當的呈現，如以長條圖比較數量的多寡、以折現圖觀察趨勢、從圓餅圖看出比例分布。本章將以更多視覺化方式來呈現實價登錄資料，主要分成三大呈現主軸：5.1 靜態圖表呈現、5.2 房價地圖及 5.3 互動式視覺化呈現。

5.1　靜態圖表呈現

傳統的資料呈現方式如長條圖、折線圖、圓餅圖等，為靜態圖表的一部分。這類圖表可以初步表現資料的數量、趨勢及比例，如 Figure 12、Figure 13。除了傳統圖表外，d3.js 還能畫出更多視覺化圖表。如 Figure 14　台灣房價漲跌圖。房價漲跌圖用顏色代表漲跌的程度，讓使用者可以一目瞭然，以最直觀的方式得到房價漲跌的資訊。

Figure 12 新竹市成交量與房價走勢圖

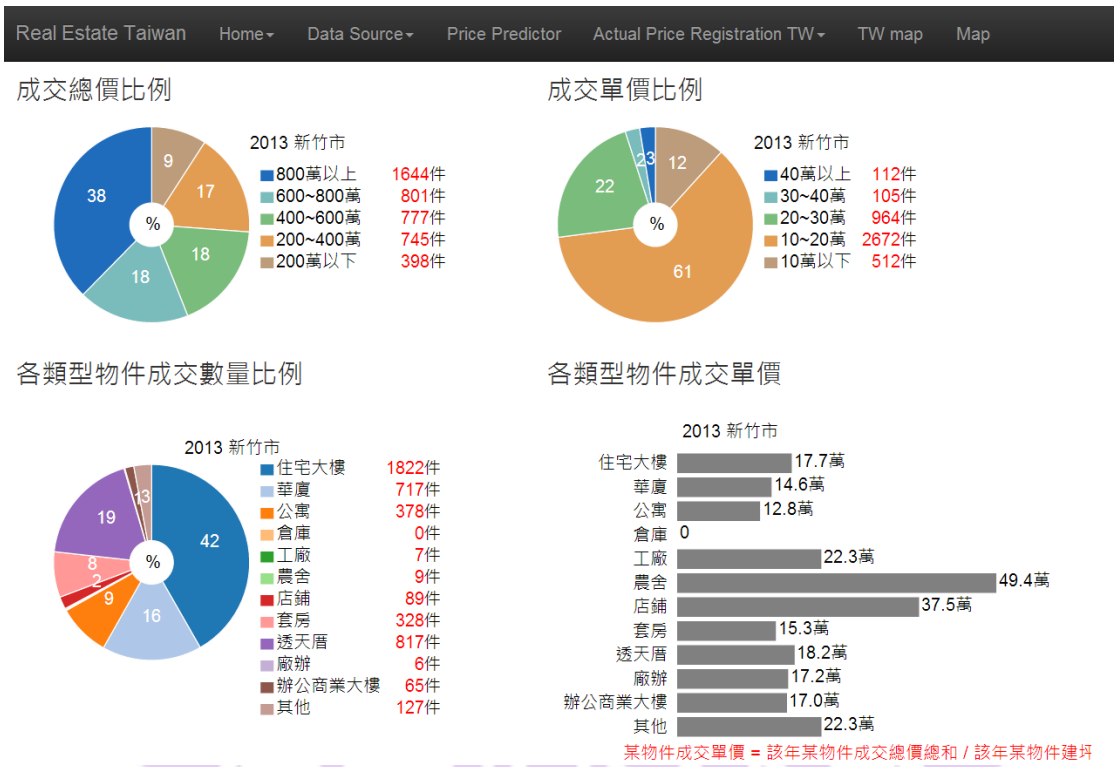


Figure 13 新竹市成交房價比例圓餅圖

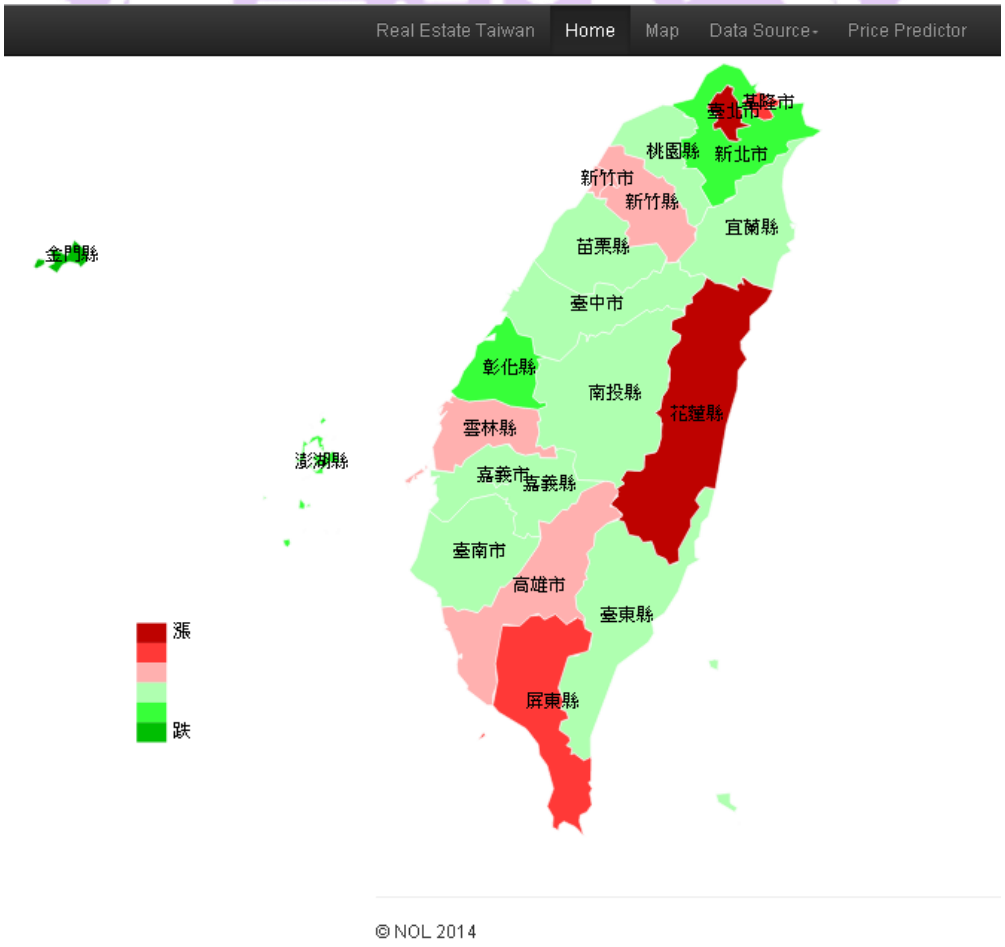


Figure 14 台灣房價漲跌圖

除了單一縣市的詳細資訊本篇還提供多元資料統計呈現比較，目前針對實價登錄及有巢氏房屋的資料做個別以及綜合的統計呈現。

實價登錄的資料在前面已經有各區詳細的資料介紹呈現了，在多源比較這裡是已總覽的方式做呈現，讓使用者可以一眼看出各縣市之間的差異，如 Figure 15 實價登錄資料總覽。

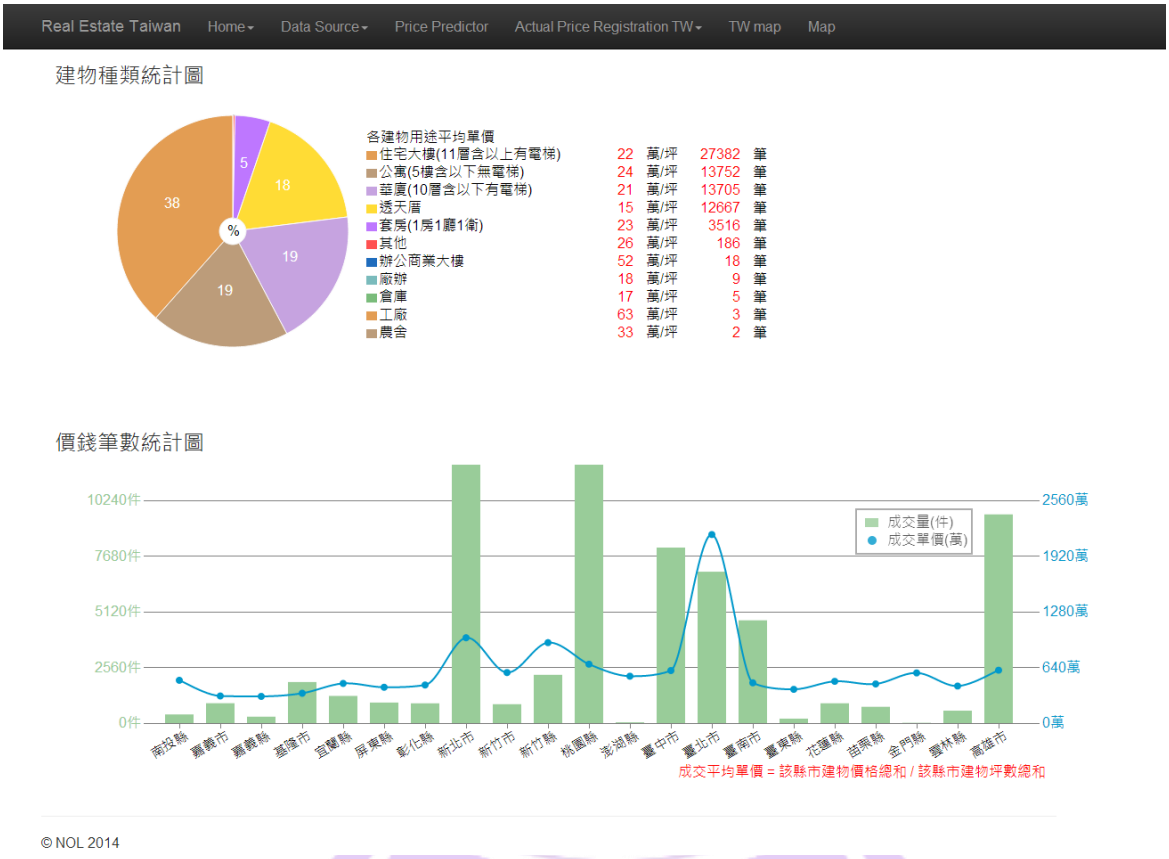


Figure 15 實價登錄資料總覽

有巢氏房屋的資料方面則有建物用途統計圖及價錢筆數統計圖。可以從中觀察有巢氏房屋銷售的建物型態種類是已住宅為重，其次是店面。而價錢筆數統計圖可看出台北市價錢最高，此現象與實價登錄資料一致。但交易筆數方面是台中市、新北市跟桃園縣較多，如 Figure 16 有巢氏房屋資料總覽。

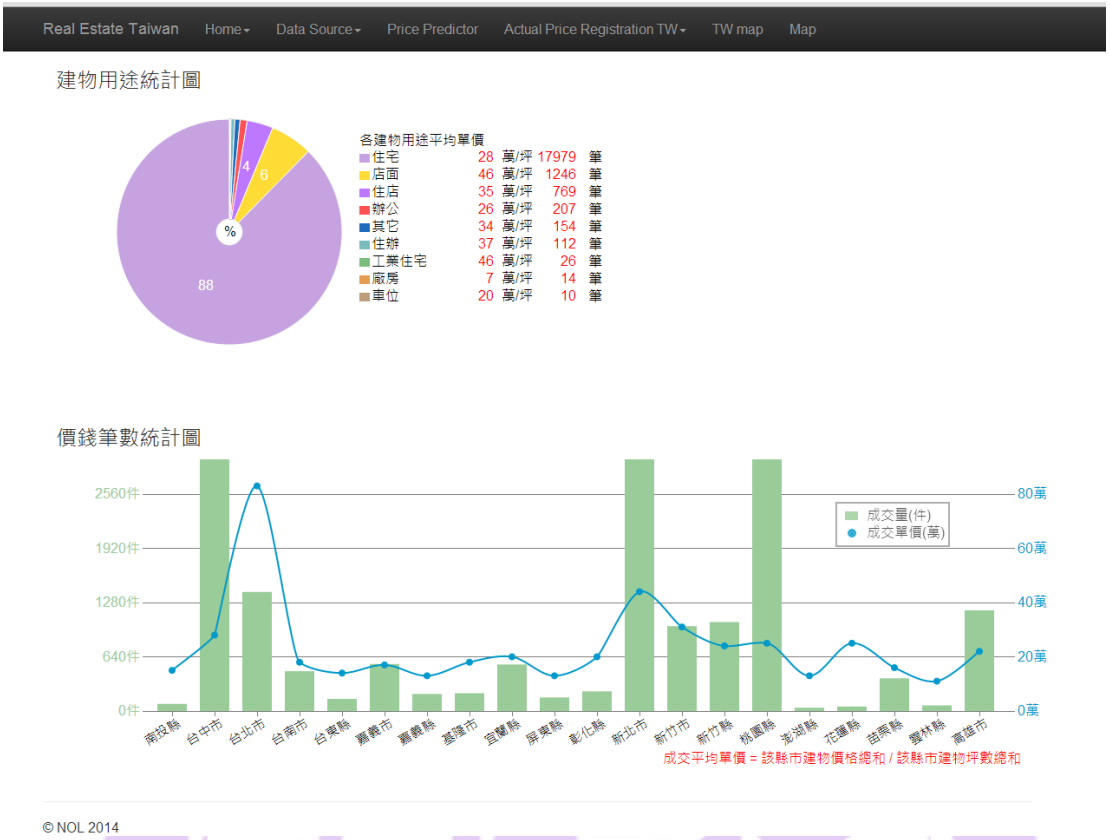


Figure 16 有巢氏房屋資料總覽

綜合比較有總價、單價及交易筆數的比較，可看出有巢氏房屋在總價及單價方面都比實價登錄來的高一些，推測是因為有巢氏房屋的價錢是欲售的價格，與實價登錄的實際成交價格有所不同。較高可能是因為要留一些殺價的緩衝。如 Figure 17 多源資料比較圖表。

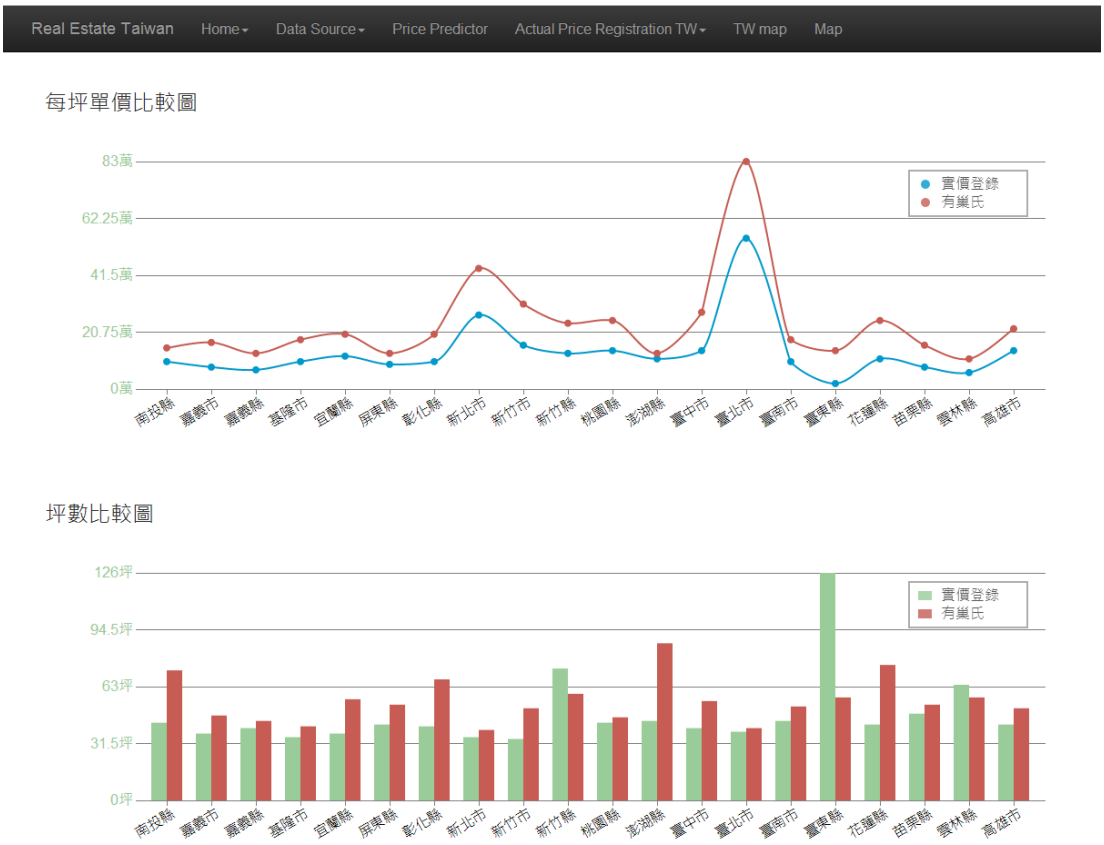


Figure 17 多源資料比較圖表

5.2 房價地圖

以 Google Maps 方式呈現台灣各區的價錢，讓使用者可以針對自己有興趣的區域更深入了解價錢的分布情形。同時，點擊可查看該筆資料的詳細資訊。如 Figure 18 系統房價地圖。

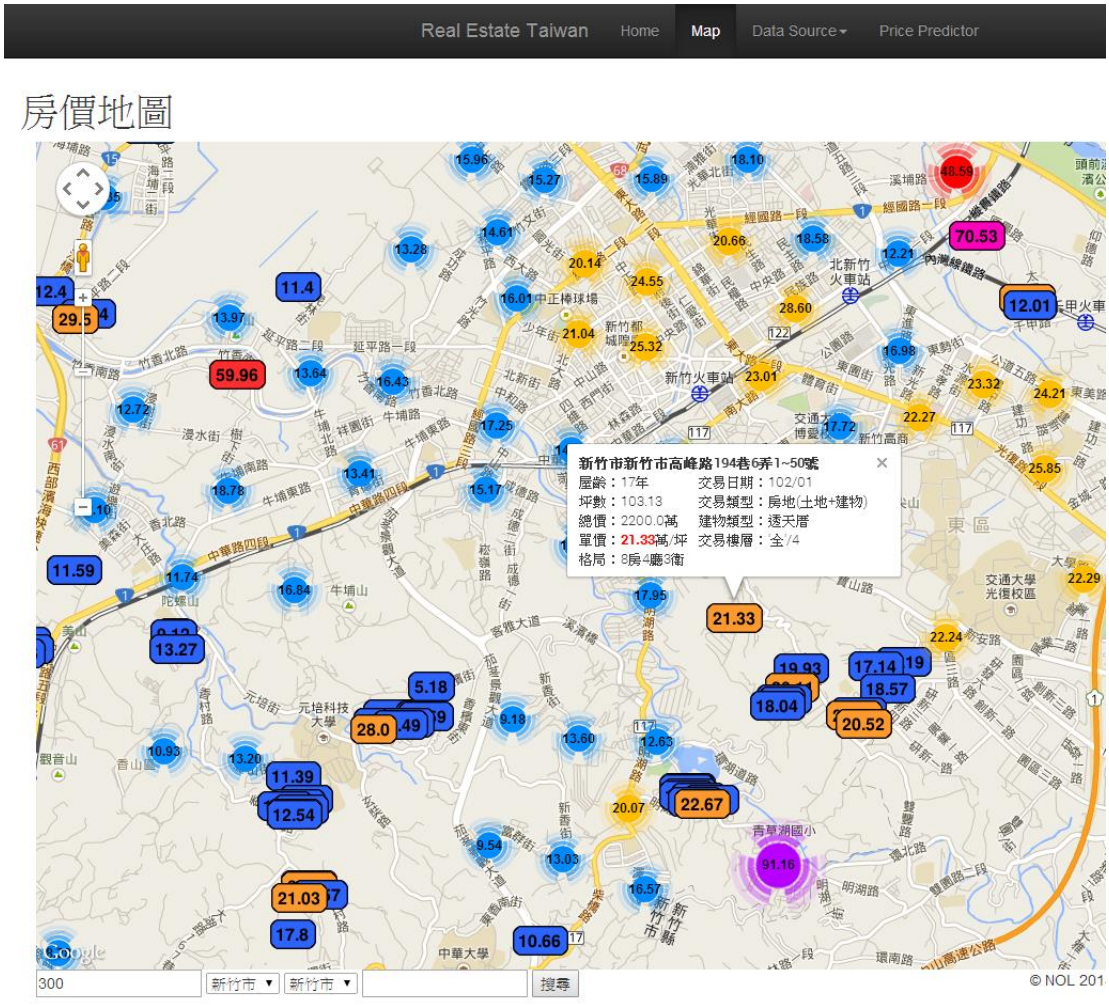


Figure 18 系統房價地圖

5.3 互動式視覺化呈現

互動式呈現方式將實價登錄資料有價錢筆數累積圖、價錢坪數散布圖、房屋種類圓餅圖和屋齡/樓高鳥瞰圖。

價錢累積圖不但可以看出各縣市價錢分布、交易量的多寡，還可以非常容易看出不同縣市的房價差異，如 Figure 19 價錢分布堆疊圖。

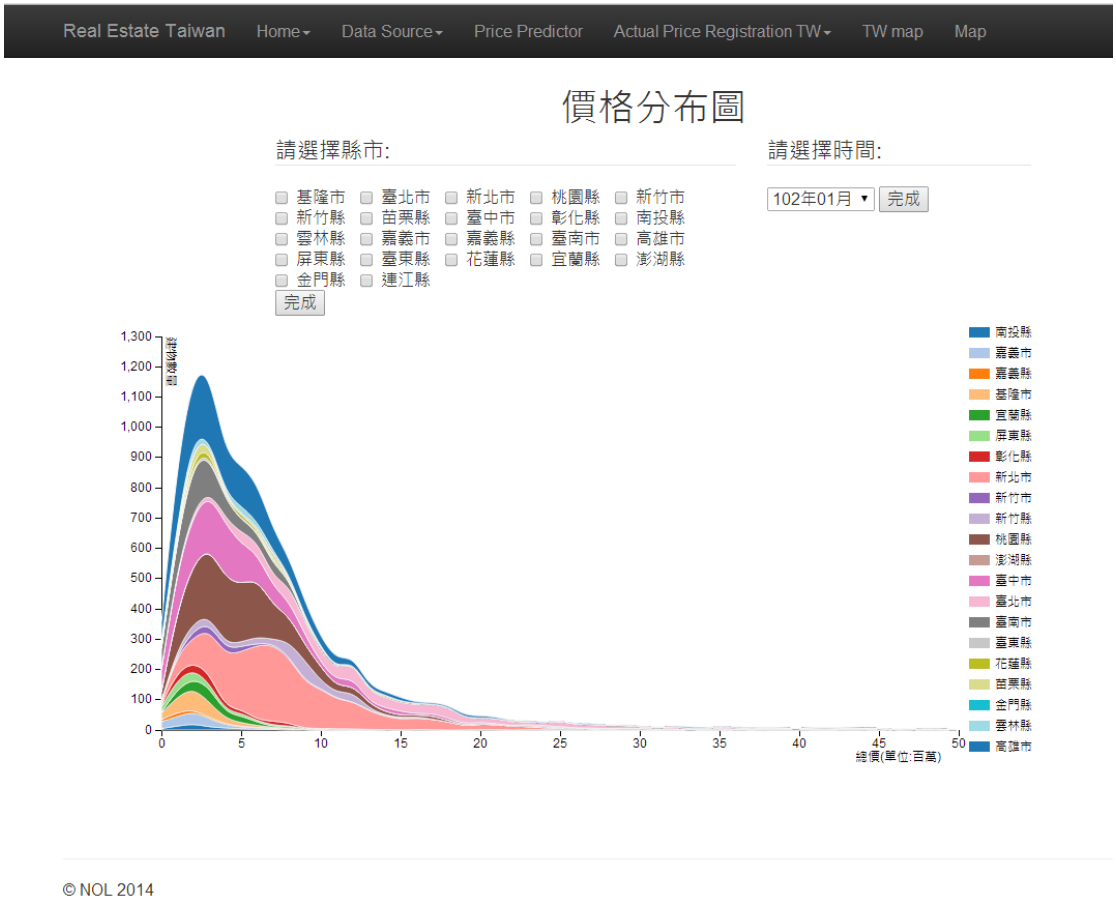


Figure 19 價錢分布堆疊圖

價錢坪數散布圖呈現不同縣市價錢與坪數的分布，能夠輕易看出城市高單價小坪數而鄉村單價低坪數大，如 Figure 20 價格坪數散布圖。

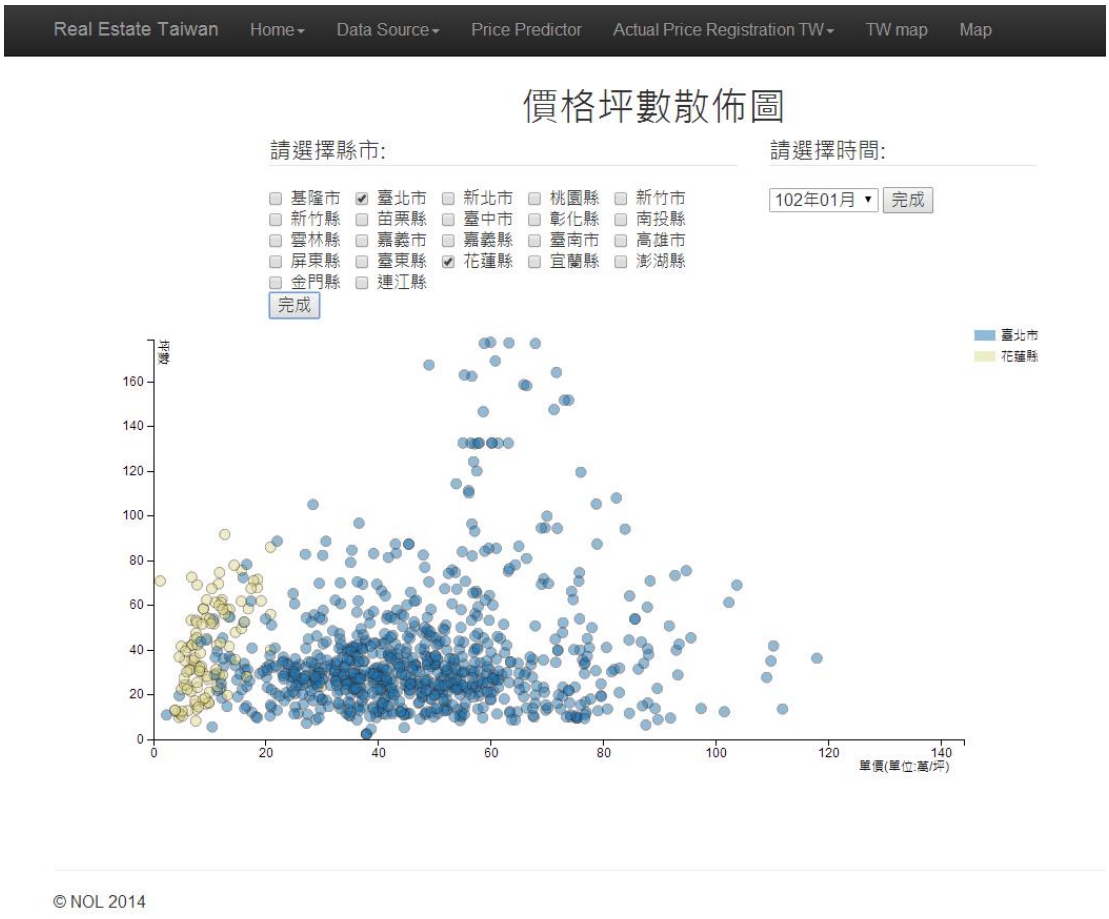


Figure 20 價格坪數散布圖

價錢累積圖及價錢坪數散布圖可以下拉式選單選擇想了解的時間，同時也可以觀察出各地區時間對價錢及坪數的影響。

房屋種類圓餅圖也是一個能夠輕易區分出城市與鄉村差異的呈現方式，城市多大樓、公寓，而鄉村多透天厝，如 Figure 21 台北市房屋種類圓餅圖及 Figure 22 台東縣房屋種類圓餅圖。

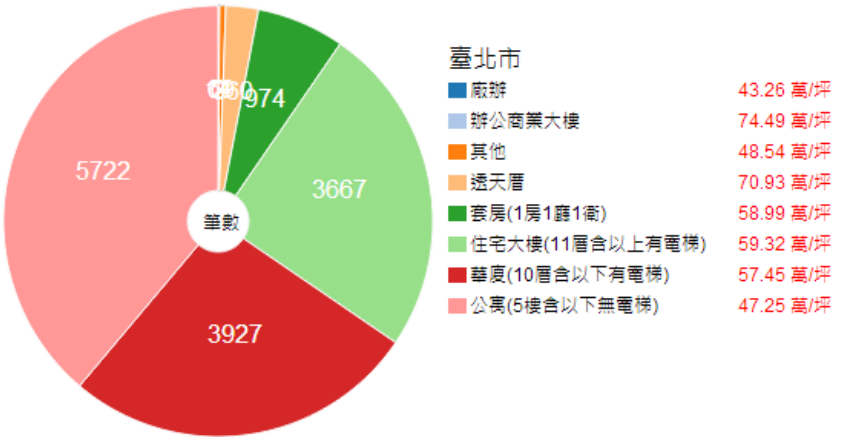


Figure 21 台北市房屋種類圓餅圖

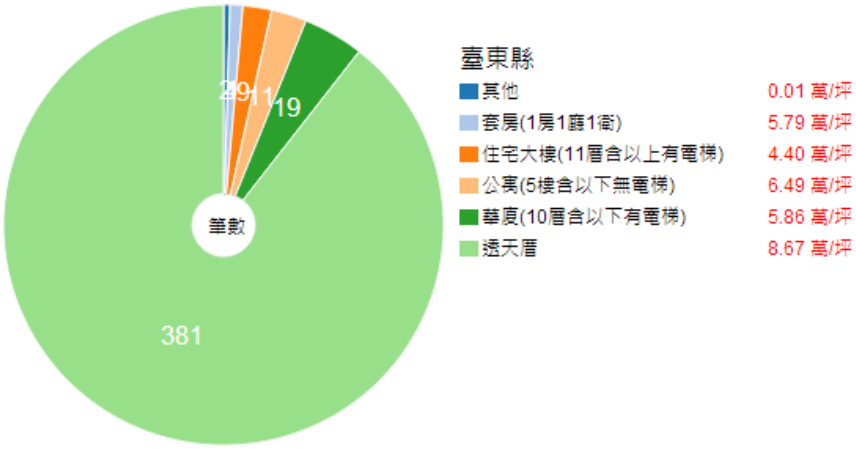


Figure 22 台東縣房屋種類圓餅圖

屋齡/樓高鳥瞰圖主要想在台灣地圖上以顏色分布的方式呈現樓層及屋齡的資訊，讓人可以一眼就看出台灣哪裡的房屋比較老，可能是下一個都更的目標，做為投資房地產的參考，如 Figure 23 屋齡預覽地圖。

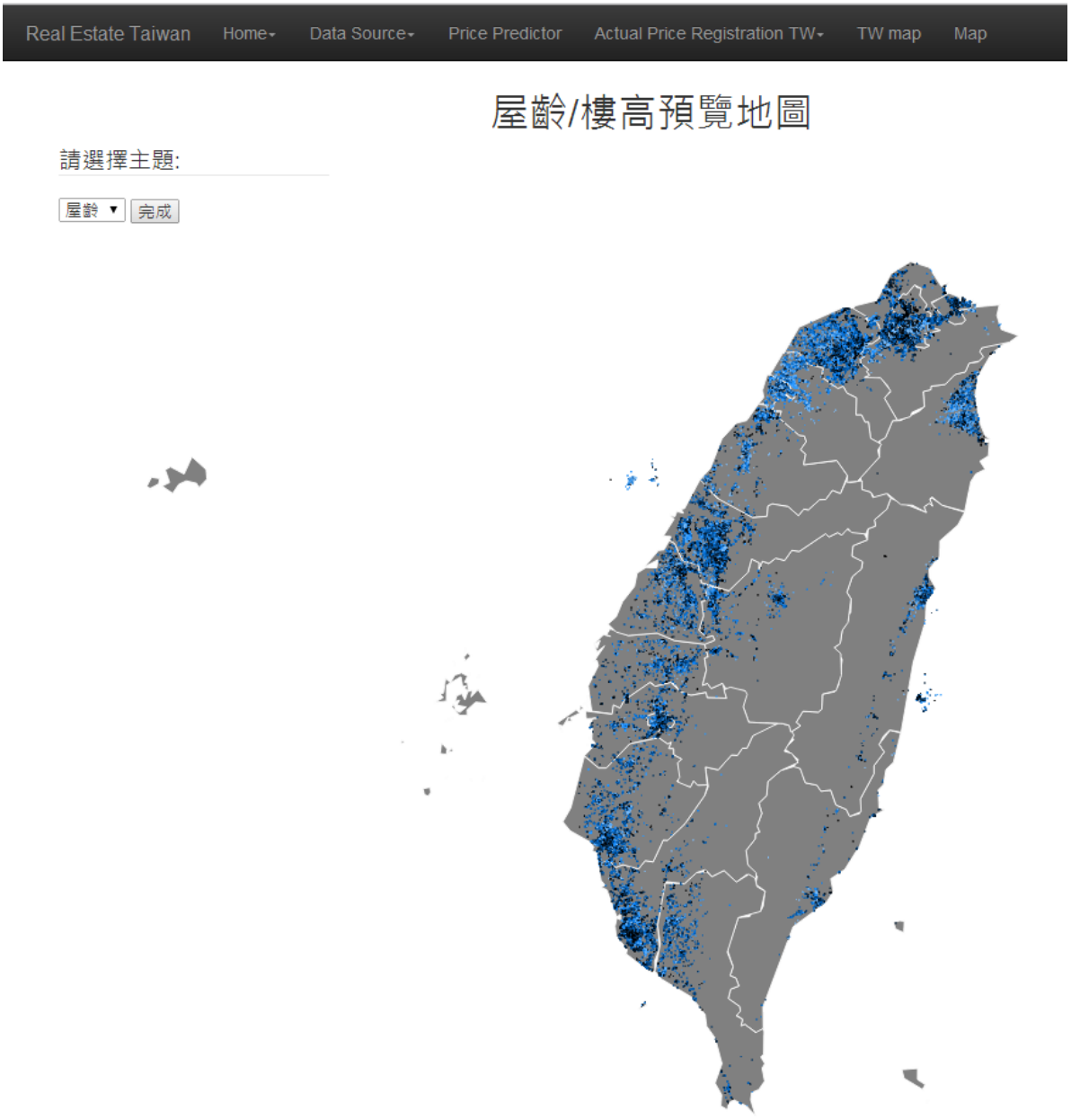


Figure 23 屋齡預覽地圖

chapter 6 平均單價迴歸分析

迴歸分析可以對資料進行初步的分析，幫助我們了解資料中各屬性之間的關係，以及屬性之間是如何互相影響的，而影響的程度又是如何。

6.1 迴歸分析方法介紹

參數估計：$y=\beta_0+\beta_1 x$，我們使用「最小平方法」，求迴歸線 L，找出參數 β_0、β_1。

最小平方法是令迴歸線 $L=\beta_0+\beta_1 x$ (預估數據)，使得所有點 P(實際數據)到 L的距離平方和最小。我們採用到迴歸線 L 的最短距離平方和最小，來求解：

$$\sum_{i=1}^{n} d_i^2=\sum_{i=1}^{n}\frac{(y_i-(\beta_0+\beta_1 x_i)^2}{1+\beta_1^2} \tag{1}$$

由於分母 $1+\beta_1^{\,2}$ 不影響最佳化結果，故將分母移去：

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2 \tag{2}$$

分別用 β_0、β_1 對 Q 偏微後令其等於 0：

$$\begin{cases}\frac{\partial Q}{\partial \beta_0}=-2\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]=0\\ \frac{\partial Q}{\partial \beta_1}=-2\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]x_i=0\end{cases} \tag{3}$$

整理後得正規方程式：

$$\begin{cases}\sum_{i=1}^{n} y_i=n\beta_0+\sum_{i=1}^{n} x_i\beta_1\\ \sum_{i=1}^{n} x_i y_i=\sum_{i=1}^{n} x_i\beta_0+\sum_{i=1}^{n} x_i^2\beta_1\end{cases} \tag{4}$$

解 β_0、β_1:

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=r\cdot\frac{S_y}{S_x} \tag{5}$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x} \tag{6}$$

得迴歸線，也是我們的迴歸式： $y=\beta_0+\beta_1 x$ (7)

6.2 參數重要性分析

本節將介紹參數的重要性分析報表的解讀，說明由 SPSS 所產生的模型分析表中的判定係數及 ANOVA 表中的顯著性代表意義。

6.2.1 殘差分析

我們做變異數的分解：

y 的實際值　　迴歸式推估的 y 值 (擬合值)

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

$$SSTO = SSR + SSE$$

總變異量
Sum of Square Total

解釋變異量 (組間變異)
Regression Sum of Square

未解釋變異量 (組內變異)
Sum of Square Error

$SSTO = S_{yy}$　　$SSR = SSTO - SSE$　　$SSE=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

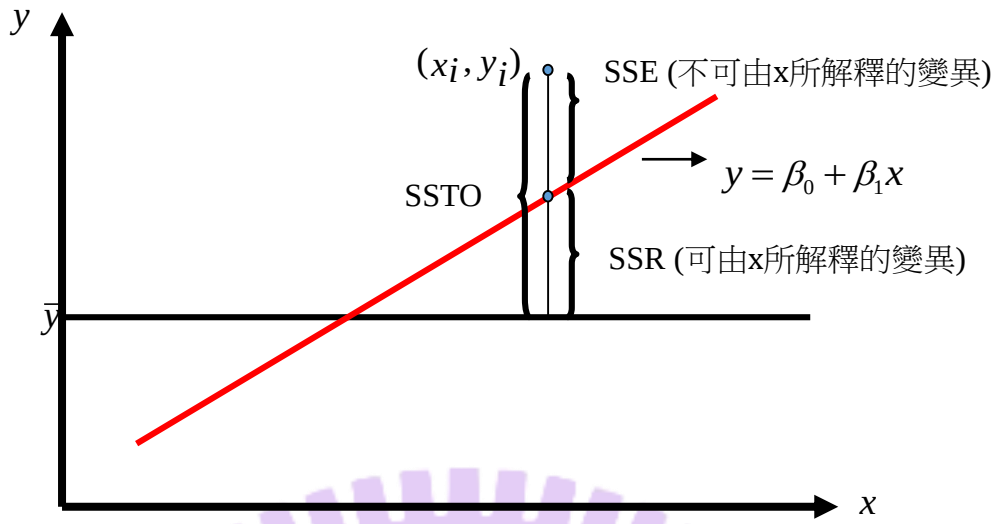


Figure 24 變異數分解

這裡我們介紹一個新名詞：判定係數 (Correlation Coefficient) R^2，表使用 x去預測 y時的解釋力，即依變數 y被自變數 x所解釋的比率，反應了由自變數與依變數所形成的線性迴歸模式的配適度(goodness of fit)：

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{462.28}{1098.93} = 0.5793 \quad (8)$$

判定係數R^2 有以下性質：

- $0 \le R^2 \le 1$
- R^2越大，代表 x 對 y 的解釋力越強

那 R^2 究竟要多大，才能決判別自變數 x 對依變數 y 是否有解是能力呢？我們可以透過變異數分析模型 (Analysis of variance，簡稱 ANOVA) 來探討。ANOVA 分析表的目的為檢定變異數分析模式是否顯著，也就是判定在實驗中，自變數 x的不同組別對依變數 y是否有顯著差異。運算流程如表格 9 ANOVA 分析表，由左至右算出 F值，其中均方和為平方和除以自由度。

表格 9 ANOVA 分析表

變異來源	平方和	自由度	均方和	F 值	P 值
迴　歸	SSR	k	MSR	$\frac{MSR}{MSE}$	$\frac{MSR}{MSE}$
殘　差	SSE	$n-(k+1)$	MSE		
總	$SSTO$	$n-1$			

而最後我們利用F值所計算的虛擬假設機率值，也就是P值，來判定顯著性。

P值代表的意義就是此判定的出錯的機率：

如果$P<0.05$，就是代表此迴歸式可以解釋x與y的因果關係出錯機率不到 5%；反之就是 x 無法有效解釋與 y 的因果關係，也就是顯著性不夠。

延續以學生的數學成績預測微積分成績的例子，我們將 15 位學生的數學成績及微積分成績，在軟體 SPSS 跑線性迴歸：此例中有 1 個自變數～數學成績 (k=1)，以及 15 個樣本個數 (n=15)。

模式摘要

模式	R	R 平方	調過後的 R 平方	估計的標準誤
1	.761[a]	.579	.547	5.963

a. 預測變數:(常數), 數學成績

Anova[a]

模式		平方和	df	平均平方和	F	顯著性
1	迴歸	636.655	1	636.655	17.904	.001[b]
	殘差	462.278	13	35.560		
	總數	1098.933	14			

a. 依變數: 微積分成績

b. 預測變數:(常數), 數學成績

係數[a]

模式		未標準化係數		標準化係數	t	顯著性
		B 之估計值	標準誤差	Beta 分配		
1	(常數)	57.160	4.771		11.981	.000
	數學成績	.428	.101	.761	4.231	.001

a. 依變數: 微積分成績

由 Figure 26 可看到，ANOVA 表格為顯著性 0.001<0.05，代表預測變數對依變數有顯著的影響，也就是數學成績 x 可以有效預測 (解釋) 微積分成績 。我們可以利用統計軟體：例如 SPSS，來做線性迴歸的分析，比起直接代入公式計算，有更高的效率以及方便性。在下一小節，我們會透過統計軟體 SPSS 對內政部實價登錄的房價資料，做線性迴歸的分析。

6.3 迴歸模型建立

我們希望透過線性迴歸，分析內政部實價登錄的房價資料。而資料可以從內政部不動產交易實價查詢服務網下載每期最新的房價登錄資料 (如果需要非當

期的資料，需要付錢向內政部購買，購買資料會比下載的實價登錄資訊更為詳細)。

6.3.1 例 1 - 分析房價與坪數、交易年月的關係

我們希望了解房價是否會因建物面積、交易年月而波動，我們取向內政部購買的實價登錄資料(從實價登錄上線至2013/11)，選台北市大安區的無車位住宅，依房屋價格(總價元)排序，去除頭尾各 10%的資料，只留下「交易年月」、「總價元」、「建物移轉總面積平方公尺」 (建物面積) 這三個欄位，最後由總價元取對數新增一個欄位，當作這次我們所要的分析資料，附幾行參考數據。

接下來我們列出取過對數的房價(總價對數)與交易年月、建物面積的關係式：

$$\text{總價對數} = \beta_0 + \beta_1 \text{交易年月} + \beta_2 \text{建物面積} \quad (9)$$

總價對數是為依變數 y，交易年月和建物面積為自變數 x。接著我們開啟 SPSS (可從交大校園軟體授權服務網下載)，將整理好的資料以 csv 檔匯入到 SPSS 的資料欄位中：

	交易年月	建物面積	總價	總價對數
1	10107.00	114.10	6150000.00	6.79
2	10112.00	62.97	3400000.00	6.53
3	10112.00	105.31	6000000.00	6.78
4	10111.00	116.42	6800000.00	6.83
5	10111.00	116.42	6800000.00	6.83
6	10203.00	110.94	6500000.00	6.81
7	10201.00	135.87	8000000.00	6.90
8	10205.00	23.10	1365610.00	6.14
9	10202.00	38.31	2300000.00	6.36
10	10112.00	114.59	6880000.00	6.84
11	10112.00	135.62	8150000.00	6.91
12	10203.00	90.94	5500000.00	6.74
13	10112.00	170.22	10300000.0	7.01
14	10202.00	222.63	13550000.0	7.13
15	10107.00	365.31	22280000.0	7.35
16	10112.00	156.92	9650550.00	6.98
17	10204.00	130.67	8050000.00	6.91
18	10111.00	46.26	2932050.00	6.47
19	10205.00	276.53	17650000.0	7.25
20	10207.00	56.03	3600000.00	6.56
21	10208.00	173.63	11250000.0	7.05

資料檢視 變數檢視

Figure 26 SPSS 介面

接著在「變數檢視」中，修改欄位「名稱」及「測量」屬性：在此例的分析資料中，交易年月、建物面積、總價、總價建築都屬於『線性』關係，所以「測量」屬性要選擇『尺度』。

(文後在例 2 會說明『名義』的「測量」屬性)

	名稱	類型	寬度	小數	標記	值	遺漏	欄	對齊	測量	角色
1	交易年月	數字的	8	2		無	無	8	靠右	尺度(S)	輸入
2	建物面積	數字的	8	2		無	無	8	靠右	尺度(S)	輸入
3	總價	數字的	8	2		無	無	8	靠右	尺度(S)	輸入
4	總價對數	數字的	8	2		無	無	8	靠右	尺度(S)	輸入

Figure 27 SPSS 變數檢視

接下來就要使用 SPSS 的分析功能，幫我們直接做線性迴歸的分析了，不用一個個公式慢慢帶入計算：選擇「分析」>「迴歸」>「線性」

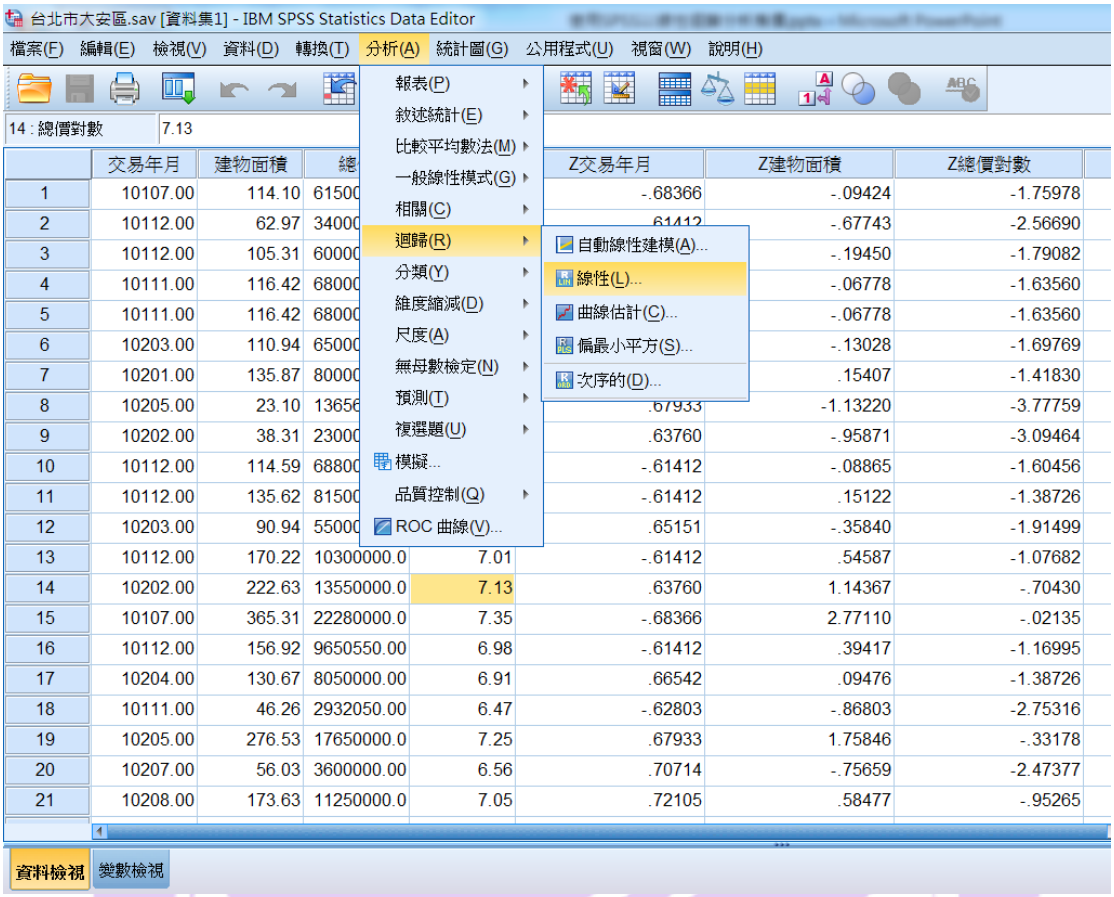


Figure 28 SPSS 分析介面

接著如同在此例中，我們剛開始建立的關係式：

$$\text{總價對數} = \beta_0 + \beta_1 \text{交易年月} + \beta_2 \text{建物面積} \tag{10}$$

將總價對數加入「依變數」，交易年月、建物面積為「自變數」加入。

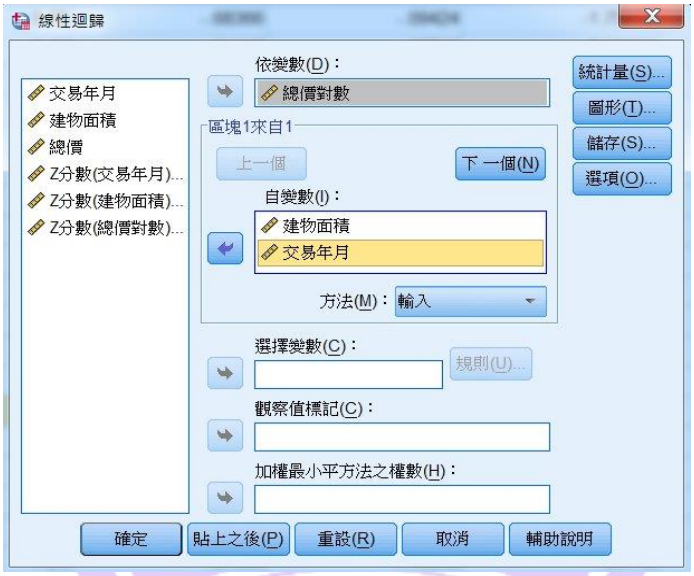


Figure 29 SPSS 變數選擇

按下「確定」，就可以看到分析結果：

模式摘要

模式	R	R 平方	調過後的 R 平方	估計的標準誤
1	.800[a]	.640	.639	.19354

a. 預測變數:(常數), 交易年月, 建物面積

Anova[a]

模式		平方和	df	平均平方和	F	顯著性
1	迴歸	93.243	2	46.621	1244.689	.000[b]
	殘差	52.551	1403	.037		
	總數	145.794	1405			

a. 依變數: 總價對數

b. 預測變數:(常數), 交易年月, 建物面積

係數[a]

模式		未標準化係數		標準化係數	t	顯著性
		B 之估計值	標準誤差	Beta 分配		
1	(常數)	2.129	.796		2.674	.008
	建物面積	.003	.000	.836	47.935	.000
	交易年月	.000	.000	.107	6.117	.000

a. 依變數: 總價對數

Figure 30 SPSS 模式摘要

我們透過最下面的表格來對分析的資料做解讀：

首先可以由「標準化係數」看到，建物面積的標準化係數為 0.836 > 交易年月的標準化係數 0.107，因此建物面積對於房價的影響比起交易年月更為明顯。

再來由「顯著性」看到，三個變數顯著性<0.05，代表這三個參數可以解釋房價(總價對數)，因此接下來我們可以透過「B 之估計值」，來完成我們房價的關係式：

$$\text{總價對數} = 2.129 + 0.003\text{建物面積} + 0.000\text{交易年月} \quad (11)$$

係數[a]

模式		未標準化係數 B 之估計值	未標準化係數 標準誤差	標準化係數 Beta 分配	t	顯著性
1	(常數)	2.129	.796		2.674	.008
	建物面積	.003	.000	.836	47.935	.000
	交易年月	.000	.000	.107	6.117	.000

a. 依變數: 總價對數

Figure 31 SPSS 係數分析

6.3.2 例 2 - 分析房價與建材、行政區域、交易年月的關係

我們希望了解房價是否會因建材、行政區域以及交易年月的不同而變化，我們取向內政部購買的實價登錄資料 (從實價登錄上線至 2013/11)，選台北市的無車位住宅。每一個行政區域都為一欄，該欄位值為是否為該行政區域，是為 1，否則為 0；建材分為鋼筋混凝土、磚造，也是同行政區域，每種建材為一欄，是為該建材值就為 1，否則為 0。最後由成交單價 (每平方公尺單價) 取對數新增

一個欄位，當作這次我們所要的分析資料。(可見附檔 avgA_台北.xlsx)

我們列出取過對數的成交單價與交易年月、建材、行政區域的關係式：

$$\begin{aligned}\text{成交單價對數} = \beta_0 + \beta_1 &\text{交易年月} \\ + \beta_2 &\text{是否為磚造} + \beta_3 \text{是否為鋼筋混凝土} \\ + \beta_4 &\text{是否為士林區} + \ldots\ldots + \beta_{15} \text{是否為萬華區}\end{aligned} \quad (12)$$

成交單價對數是為依變數 y，交易年月、行政區域、建材為自變數 x。

接下來我們使用 SPSS 來幫我們用線性迴歸做分析，我們將整理好的資料複製貼上到 SPSS 的「資料檢視」的欄位中，接著在「變數檢視」中，修改欄位「名稱」及「測量」屬性：在此例的分析資料中，交易年月、成交單價屬於『線性』關係，所以「測量」屬性要選擇『尺度』；而是否為某建材或行政區域的欄位為『類別』關係，所以「測量」屬性要選擇『名義』，這裡應該要改成附幾行參考數據。接下來就要使用 SPSS 的分析功能，幫我們直接做線性迴歸的分析：

一樣選擇「分析」＞「迴歸」＞「線性」(此例沒有先做標準化)，依變數為成交單價對數，交易年月、行政區域、建材為自變數。按下「確定」後的分析結果：

係數[a]

模式		未標準化係數		標準化係數	t	顯著性
		B 之估計值	標準誤差	Beta 分配		
1	(常數)	4.540	.131		34.569	.000
	士林區	.013	.006	.019	2.234	.025
	大同區	-.059	.010	-.045	-5.986	.000
	大安區	.197	.006	.300	34.671	.000
	中山區	.072	.006	.109	12.593	.000
	中正區	.109	.007	.126	15.644	.000
	文山區	-.088	.005	-.157	-17.237	.000
	北投區	-.094	.005	-.160	-17.857	.000
	松山區	.078	.007	.092	11.331	.000
	信義區	.073	.006	.096	11.565	.000
	南港區	.023	.006	.032	3.818	.000
	萬華區	-.155	.008	-.146	-18.648	.000
	磚造	.014	.006	.017	2.341	.019
	交易年月	6.097E-005	.000	.034	4.717	.000

a. 依變數: 成交單價對數

Figure 32 SPSS 係數分析 (台北市各區)

我們透過分析後的表格解讀：

表格上的參數，顯著性皆<0.05，所以都能解釋成交單價對數，而其中內湖區以及鋼筋混凝土並沒有出現在此分析表格中，因為分析後把這兩個參數歸為『排除變數』，原因可能為資料不足或無法解釋房價。

我們透過「B 之估計值」，來完成我們房價的關係式：

$$\begin{aligned}\text{成交單價對數} = {} & 4.540 + 0.000061\text{交易年月} - 0.014(\text{是否為磚造}) \\ & + 0.013(\text{是否在士林區}) - 0.059(\text{是否在大同區}) - 0.197(\text{是否在大安區}) \\ & + 0.072(\text{是否在中山區}) + 0.109(\text{是否在中正區}) - 0.088(\text{是否在文山區}) \\ & - 0.094(\text{是否在北投區}) + 0.078(\text{是否在松山區}) + 0.073(\text{是否在信義區}) \\ & + 0.023(\text{是否在南港區}) - 0.155(\text{是否在萬華區})\end{aligned} \tag{13}$$

我們可以從 SPSS 報表中的標準化係數，看到台北市房價中大安區的房價是最高的，而北投區、文山區以及萬華區的房價都屬於偏低的行政區域。

用我們的房價公式，推估台北市士林區 102 年 12 月的鋼筋水泥土建築，每

坪單價 (1 平方公尺= 0.3025 坪)，約為 50 萬：

$$10^{(4.540+0.000061*10212+0.013)}/0.30205=496424.446135 \quad (14)$$

與信義房仲網的成交資料比較：102 年 12 月台北市士林區房價約為每坪 53.7 萬。

Figure 33 信義房屋士林區截圖

chapter 7　資料收集作業流程

資料收集的過程相當繁瑣，因此本章將介紹資料收集的流程，包含資料收集的過程、資料更新方式以及使用與呈現。7.1 資料收集與系統建置、7.2 資料更新及 7.3 系統重建機制。

7.1　資料收集

資料收集主要分成實價登錄及房仲網兩部分。流程有資料的取得、資料整理及匯入資料庫。

1. 資料取得

實價登錄資料是於 102 年 11 月向內政部購買 101 年 8 月至 102 年 9 月 15 日的租賃、買賣及預售資料。而房仲網方面是有巢氏房屋，自行撰寫 crawler 搭配 bash 檔執行。

2. 資料整理

實價登錄資料雖說有政府的把關，但資料空缺、格式不一的情況還是層出不窮。觀察表格 1 實價登錄實際資料以下是我們做的資料整理：

A. 「土地區段位置/建物區段門牌」依實價登錄之規定應該填入除去縣市及鄉鎮名的地址區段，但資料中建物地址大多數還是保留了縣市及鄉鎮名，因此需將其除去。

B. 面積的計算方式在收集儲存時是以實價登錄原始的型態儲存，也就是以平方公尺作為度量單位，而呈現時需將其轉換成以坪當單位。同時平均單價也需作變動。

C. 「移轉層次」裡除了包含交易的標的樓層外，還有許多其他資訊，如：陽台、騎樓、夾層、電梯間等等，相當混亂。因此在整理過程中需將其主要資訊，也就是交易的標的樓層轉換成數字儲存，方便日後計算或呈現使用。

D. 「交易標的坐標」是 TWD97 的座標系，但在許多資料呈現都是已經緯

度作參考，因此再匯入資料庫前也需將其座標轉換成經緯度一起存入資料庫中。

將上述問題整理完後匯出成.csv 檔，就完成實價登錄資料整理的部分。有巢氏的部分如下：

A. 當初crawler將資料匯出成.tsv而非.csv的原因是有巢氏在價錢的欄位有加上千分位的逗點，若以.csv 紀錄會造成藍位區分上的混淆。因此，在整理時要將千分位的逗點移除。

B. 格局的欄位需將「x 房(室)x 廳 x 衛」形式切割成房、廳、衛三個不同的欄位儲存。

C. 有巢氏房屋的資料不像實價登錄資料有縣市、鄉鎮、地址的獨立欄位，而是全部都寫在地址一個欄位中。為了資料的結構化及統一性我們將其切成三個欄位，與實價登錄資料庫一致。

D. 單位去除。如：面前巷道及管理費的資料有含單位，需將其去除後在存入資料庫中。

3. 匯入資料庫

第三步為匯入資料庫，以 phpMyAdmin 的介面，勾選.csv 檔匯入，並將換行字元改為'\n'。

7.2 資料更新

資料更新主要也是分兩個步驟，資料收集與網頁更新。資料收集分成實價登錄與房仲網。

由實價登錄資料須購買，計費方式是以資料量作為標準，100M 以下是 2000 元，100M 到 500M 是 4000 元，以學術研究名義可半價購買。經過衡量半年購買一次較為恰當。同樣的，在資料更新也包括資料收集和資料整理。如 7.1 的資料整理部分，我們需要做地址路段化；面積單位轉換，同時平均單價也需作變動；「移轉層次」轉換成數字儲存；「交易標的坐標」轉換成經緯度一起存入資料庫

中。

房仲網是自行抓取較不受此限，約兩個月抓取一次。抓取方式參照資料收集流程及可。資料收集完成後同樣也需要做整理。將價錢資料中千分位的逗點移除；格局的欄切割成房、廳、衛三個欄位儲存；將地址資訊切成三個欄位儲存；特定欄位的單位去除。

第二部分是網站更新，需要再資料收集後產生新的檔案供網頁顯示。產生完相對應的檔案後，網頁內容也需要作為調，如下拉式選單需新增幾個月分等等。

7.3 系統重建機制

意外隨時可能會發生，像是停電、主機硬體故障，甚至電腦中毒或是操作不當，造成硬體故障或是資料損毀而無法讀取都有可能造成系統毀壞。因此，適時的備份是很重要的。而除了備份資料外，系統要能運作還需要一些繁瑣的設定何調整。本篇制訂了一套重建機制來解決未來可能發生意外的狀況。

當主機發生意外或是系統需要移植到另外一台主機上時，就需要作系統重建。在討論重建之前要先討論備分。當資料更新過後，資料庫以及網站的資料夾需要備份至雲端，以確保之後的系統有最新的版本能復元。

在需作系統重建時，首先將資料庫備份還原至最新備份版本。第二，將網頁資料夾擺放置新的位置。內部需調整地方有網頁顯示及連接資料庫兩部分：網頁顯示部分，房價趨勢地圖的部分，網址 url 需依據檔案路徑作調整；連接資料庫的部分需更改 IP、帳號密碼。IP 須依照新的資料庫所在主機 IP 設定，而帳號密碼也須變更成擁有能夠存取新設置資料庫權限的帳號密碼。

chapter 8　結論

本章將總結上述的問題、目標貢獻、解決方法。本研究收集了多源的房地產買賣資訊，除實價登錄資料外，還有其他房仲網的待售房屋資料。用視覺化的方式呈現在網頁上，幫助觀察、分析。利用實價登錄資料以線性迴歸的方式來建立房價模型，分析了解房市變化。

8.1　結論

本篇收集了多源的房地產，除了內政部公開的實價登錄資料，還有非公開的房仲網房屋待售資料，並制定一套資料收集流程以利日後資料收集作業。觀察分析需要將資料視覺化，除了一般的圓餅圖、折線圖外，我們有實作其他互動式的圖表讓使用者能更容易從資料中獲得想要的資訊。藉由房價的線性迴歸模型的建立，分析排除異常資料，做房地產價格的推估。建立時間模型，觀察房市走向。

要做資料呈現之前必須要有穩定資料來源是重要的。因此本篇制訂了一套資料收集的標準作業流程，讓系統能呈現最新的資料。本篇制定的資料收集機制包含資料庫建置、定期資料更新、資料整理呈現這三大部分。

資料視覺化部分本篇除了將資料做簡單的統計，透過網頁的方式呈現視覺化的資料。能輕易從中獲得想了解的資訊、看到一些重大事件對房市交易的影響或是房市的整體走勢。

房價模型建立能夠幫助買賣雙方在交易過程更順利、政府合理課稅。由於買方為了不想買貴會想了解自己中意的房屋價位在哪，而賣方若也能充分了解自己房屋的特性及價格不但可以讓交易更快速順利，也不怕自己開價太低而少賺一筆。除了買賣雙方，政府若能掌握房地產交易的金額，對未來實價課稅也有一定的幫助。

本篇將經過前置處理的實價登錄資料當作來源，對每個區域建構出該區的房價迴歸模型。使用者可以輸入自己對房屋的需求，如坪數、樓層、格局等，經過 SPSS 建模型，推算出房價。

8.2 未來研究

本篇對實價登錄及其相關資料的蒐集、分析和視覺化呈現有了初步的成果。除了分析呈現之外，還可以對資料做分群，再對各群進行分析，可以找出相似條件下的交易中價格偏高級偏低的異常資料。

有了好的自動化蒐集資料流程可以對資料做長期的觀察分析，建立時間模型，能確實掌握整體的變化趨勢。

以迴歸模型分析房價，著重用統計方法初步找出對於房價影響的因子 (例如: 坪數、交易年月、行政區...等)，在預測結果上還不是非常準確，未來納入更多房價特徵 (例如屋樓層、格局、房屋屬性、屋齡...等) ，對於類別資料建立交互變異參數項，最後以逐步迴歸法排除可能產生共線性的因子，挑選對於房價公式能有更精準預測的參數。

未來研究可能會採取以資料探勘技術：利用分類器將誤差過大的資料篩選出，以不同價格區間當作類別、分群法將不同房屋特徵做分群；或採取改良型的線性迴歸模型，以求在房價預測上能有更精準的數據，能有更好的實務應用。

參考資料

[1] 迪歐地圖, http://dio.idv.tw/

[2] 樂屋講座, http://www.rakuya.com.tw/hnews/hnews_list/4226/2/0/8/0

[3] 好房網新聞, http://news.housefun.com.tw/gloria/article/52176639176

[4] 今日新聞, http://www.nownews.com/n/2014/03/13/1146297

[5] 中時電子報,
http://www.chinatimes.com/newspapers/20140120000299-260102

[6] MBA 智庫百科, http://wiki.mbalib.com/zh-tw/特征价格法

[7] Desiring Clicks 色彩與視覺,
http://dclick.fourdesire.com/2013/03/01/【資訊視覺化】資訊視覺化與工程的應用?ref=category

[8] “住宅價格指數之研究－以台北市為例 Housing Price Index in Taipei”，林秋瑾、楊宗憲、張金鵬，住宅學報 4 期，1996，p1–30

[9] 內政部不動產成交案件實際資訊資料供應系統,
http://plvr.land.moi.gov.tw/DownloadOpenData

[10] “不動產自動估價與估價師個別估價之比較－以比較法之案例選取、權重調整與估值三階段差異分析”，江穎慧(Ying-Hui Chiang)，住宅學報 1 期，2009，p 39-62

[11] “半參數法於國內不動產大量估價之可行性評估”，彭建文、楊詩韻、林財川，都市與計劃-住宅與不動產，2010

[12] “台灣地區特徵性房價函數估計係不一致問題之探討 台灣地區特徵性房價函數估計係不一致問題之探討”， 林素菁，中華民國住宅學會第十一屆年會論文集，2002，p 268-277

[13] “房價指數模型建構之研究--以桃竹地區市鎮交易資料為例”，劉錦龍、王恭棋，產業經濟研究所碩士在職專班碩士論文，2006

[14] ”影響房價變動因素之探討－以高雄市區爲例”，毛麗琴，Journal of Commercial Modernization，Vol.5，2009

[15] ”都市房價變動影響因素之系統動態模擬 —台北市之實證研究”，陳幸宜，成功大學都市計劃學系學位論文，2003