# Automated Mass Appraisal System with Cross-City Evaluation Capability: A Test Development in China

by

**Yihan Guo**

A thesis submitted in partial fulfillment of the

requirements for the degree of

**Doctor of Philosophy in Engineering**

**University of Warwick, WMG**

September 2018

# Contents

# List of Tables

# List of Figures

# Acknowledgments

During my journey of PhD, there are many people supplied help and advice. I would like to thank them all for the support they gave me in their own and unique ways.

I would like to express my sincere gratitude to my supervisor, Professor Jay Bal, for his continuous support, endless patience, guidance and advice of my PhD study, developing my critical reasoning, and most importantly for being a friend through all these years.

My sincere thanks also go to my second supervisor, Dr Xiao Ma, who provided me with his guidance, advice and friendship that inspiring me through the journey.

I would like to express my appreciation to my parents, Qingping Guo and Lixia Nie for their supports and encouragements during all these years that I have been away from them.

Last but not the least, I would like to thank all my friends and colleagues who supported me during this journey.

# Declarations

I declare that except where acknowledged, the material contained in this thesis is my own work and that is has neither been previously published nor submitted elsewhere for the purpose of obtaining an academic degree.

-Yihan Guo

# Abstract

The appraisal of property value is extremely important in a modern economy. For example, developers and end-consumers use appraisals for their investment decisions. Governments use it for taxation purposes, while banks rely on appraisals to update their risk profile when managing mortgage and credit application activities. With fast developing economies, quickly valuing new cities and suburbs as they get built becomes particularly difficult. Globalisation has also increased the need for common international valuation standards and automated methods.

This research investigates the present mass appraisal systems and the role of automated valuation models. Financial institutions and institutional investors are increasingly more concerned about constantly updating their present portfolio value especially in a dynamic market. Trends of significant peaks and troughs need to be accounted in a faster cycle time with short bursts of pricing adjustments. The problem poses a challenge because property transactions are infrequently traded unlike other commodities such as securities. Hence, there are not many recent transactions for the same property to receive an updated value with a simple adjustment based on economic conditions.

The study proposes a method that solves both large-scale mass appraisal with an ability to search across cities to discover properties with similar char-

acteristics for its update and comparison scheme. This research advances the automated valuation model for the residential property market with a test development performed in China. In particular, the resulting model was tested with data from Chinese Tier 1, 2 and 3 Cities to evaluate property values. This research performs several major accomplishments. First, it demonstrates the efficacy of reducing human cognitive effort in the mass appraisal exercise. Second, by applying Artificial Neural Network capabilities in the automated valuation model, pricing of residential properties are able to draw upon knowledge from more mature cities with greater number of transactions and apply to newer developments in less developed cities. Third, the proposed mass appraisal system shows the reliability and robustness that matches the rapid development of Chinas real estate market that had been verified by a real application.

Finally, the approach developed provides a valuable new method for property valuation that reduces the possible bias, increases consistency and lowers the effort required by current manual methods, with a lower data requirement.

# Chapter 1

# Introduction

*"The role of appraisal cannot be overemphasized because appraisal values are used as a basis for lending and investing" Fisher (1906)*

Property appraisal is the assessment of the market value of the real-estate industry, encompassing the buying, selling, lending, taxation and a whole range of external support functions. This introductory chapter states the underlying basis of the thesis, and provides the necessary background information on property appraisals and explains the motivation for pursuing a mass-appraisal system that is of great significance to various stakeholders. For example, developers and end-consumers used appraisals for their investment decision, governments use it for taxation purposes, while banks relied on appraisals to update their risk profile when managing mortgage and credit application activities. This chapter begins with the research area (section 1.1), followed by section 1.2 explaining the motivation for this research. Subsequently, the research objectives are set in section 1.3. Finally, section 1.4, outlines the remainder of the thesis.

## 1.1 Research Area

The property appraisal process was traditionally been considered as a process of rational judgment of the appraisers to produce an estimate of the current property value. Thus, it was an accepted practice to view property appraisal as adjustments made over time introspectively, which led to the moving average measurement movement in the literature (Geltner et al., 2001). In this stream of literature, it has been acknowledged that in a sparsely traded market, rational outcomes follow a smoothed or lagging appraisal trend (Quan and Quigley, 1991). The optimal value of current appraisal was a function of a weighted average price based on establishing some comparables, and the potential pricing movement from the previous appraisal. This form of approach establishes a weighing index ranging from 0 to 1, such that they account for the expected price changes from previous appraisal, and the estimation of the reliability of the market based on a comparable basket of objects.

The implicit assumption made in property appraisal was that using information from the previous appraisal would result in the improvement of current appraisal due to the uncertainty of market signals are. Nonetheless, this view solely focuses on an appraisal process from a task-oriented standpoint that considers market highly uncertain. There were a few aspects that were raised previously in order to provide further insights into this view. First, there was no solution as to whether an appraiser is acting in a consistent manner in the valuation exercise (Brown and Matysiak, 2000). This research avoids the controversy of this issue, as the focus will be on the efficiency of a large-scale mass appraisal system and not on the pricing or the valuation accuracy. Second, there were suggestions to evaluate appraisal by pursuing empirical

approaches of the property markets (Geltner, 1989). The stream of thought sought to address the factors that could contribute to the value changes of the property. Again, although this research covers the issues surrounding this topic, the focus of this research is on creating efficiency for the large-scale mass-appraisal system.

In the behavioural real estate literature, valuers were perceived as information consolidators but did not necessarily conform to normative expectations (Diaz and Hansz, 2002). This was because bias might appear as a result of the agreed sale price and comparing acceptable adjustment figures with peers rather than an analysis on market forces or supply and demand. This was seen as anchoring on the decisions by ones own network which contributed to the smoothing of the appraisal. Since there were many appraisals that had been challenged in the courts especially after the recession in 1990, the UK market revealed that appraisers were more reliant on past values and hard evidences to justify the present value (Crosby et al., 1998). One would think that having hard evidences is a good step in valuing a property but this is not completely accurate in sparse markets. Properties are commodities that transact at a longer interval unlike the buying and selling of securities. Hence, relying on actual transacted information might be too old for present day valuation. To bridge the value gap between a transactions that happened a decade ago with the expected prices today, requires the ability to reduce estimation uncertainty.

Therefore, literature on mass appraisals have been heavily centred on investigating the smoothing effects, where one is concerned about naturally arriving at an appraised value such that it also mirrors the measurement of property investment performance. Smoothing inaccuracies, in the context of

appraisal mathematics is referred to as an under or over-measurement of the true variance. Hence, the extent to which market change is underestimated is believed to be affected by the inability to gauge market movement.

## 1.2 Research Motivation

Performance on any appraisal system needed to reflect the ability to filter through big data of prior over/under-pricing into the appraisal process. The challenge was for the fund manager to manage periodic performance appraisals at a high service rate, which was a huge undertaking effort of an appraisal service. In addition, research in this area has shown that the lack of appraiser effort contributed more to the inability to account for past over/under pricing than the poor quality or lack of market information(Galal, 2001). High human cognition effort needs to be maintained for a high-level performance of appraisal. In a large fund, taking into account a surge in appraisal workload would increase the uncertainty in judgment. Hence, the motivation of this research is not in establishing an empirical regime to confirm this view as there were already many studies confirming this problem. Therefore, this study seeks to reduce the mental effort in the valuation process by valuers by introducing the role of machine learning capabilities in undertaking the repetitive behaviour in judging the smoothing of the valuation in real estate markets. This research is a direction closer to avoiding the self-fulfilling prophecy for valuers as market participants and intermediaries who are influenced by their own historic appraisals in the pricing decisions. This study contributes to the smoothing literature, where one can impartially evaluate the extent to which appraisal characteristics are accounted in the appraisal process, while avoiding

the issue of market inefficiency. These issues are critical when property trading activities and fund management of real estate requires a mass appraisal system that accounts for the characteristics of actual trading prices in an efficient, consistent and reproducible manner.

## 1.3 Research Objectives

The main research objective of this study is to demonstrate the characteristics necessary to develop a mass appraisal system that has the following capabilities:

1. An ability to reduce the cognitive effort of human appraisers in the context of high-volume appraisal activities.

2. An ability to account between the characteristics that are unique to the location of the appraised properties and the characteristics that are common to separate locations.

3. An ability to avoid the human bias that may cause the inaccuracies from smoothing of the appraised properties.

4. Reduce uncertainties that are involved in the need to estimate for present value when the last transacted prices had been achieved long ago.

## 1.4 Outline of the thesis

This section of the chapter outline the remaining structure of the thesis as follows: Chapter 2 provides the literature review on mass appraisal system and how it is linked to the automated valuation model, including coverage on

the China market. Chapter 2 also reviews the various machine learning and traditional approaches that are applied in real estate appraisals. Chapter 3 provides more in-depth review on the type of ANN which sets the basis for our choice in the automated valuation model for the proposed mass appraisal system. This section also covers the various algorithms in ANN. Chapter 4 discusses the research design and the approach of data collection as well as the variables used in the study. Chapter 5 outlines the technical development of proposed algorithm in the cross-city appraisal evaluation capability. Chapter 6 discusses the experimental results, while Chapter 7 concludes the study by matching the findings with the objectives of this study. The implications to the stakeholders as well as policy makers including the suggestions for future research area also discussed in this concluding chapter.

# Chapter 2

# Mass appraisal and Automated Valuation Method-Literature Review

## 2.1 Introduction

Standardised procedures and methods for property valuation and mass appraisal of property worldwide have become a strong requirement across many nations. The International Accounting Standards (IAS), the International Financial Reporting Standard (IFRS) and the International Valuation Standard (IVS) have pointed out that it was vital to have consistent approaches to the valuation of real estate for accounting, banking activity, stock exchange listing and leverage lending purposes (Mansfield and Lorenz, 2004; Lützkendorf and Lorenz, 2005). The controlling body of IVS even attempted to propose how the method should be developed - The Development of Mass Appraisal Systems for property taxation should follow recognized scientific standards in

statistical applications. (International Valuation Standards Council, 2017). It contributes to a surge on the development of statistics based approaches to property valuation, for example, the use of multiple regression analysis (MRA) in property valuation.

Kauko and d'Amato (2008) stated Mass appraisal might be defined as a systematic appraisal of groups of properties using standardised procedures. Typically Mass appraisal methods applied to a large number of grouped properties, as in a region or country and not a single property. The reliable judgement of the values of a group of properties and an individual property within that group indirectly, using a standardised approach to achieve specific objectives (e.g. property taxation, insurance valuation etc.), is the main goal of these methods (McCluskey et al., 1997; González et al., 2002; Fu et al., 2015; Tian and Yang, 2014). Computing powers has advanced these methods into automated systems for example:

> *"...software systems that use one or more mathematical techniques to estimate the Market Value of a given property at a specified date. They automatically select the relevant market information and the appropriate method of aggregation by means of one or more proprietary algorithms that typically perform a statistical analysis of available property market information. They utilise databases of public and private property records, historical appraisal reports and multiple for sale listing services that provide property characteristics and sales information. They offer the competitive advantage of speed and low cost. Their methodology aims to guarantee an objective valuation without appraiser subjectivity or bias." (Downie*

*and Robson, 2008)*

> *"a mathematically based computer software program that pro-duces an estimate of market value based on market analysis of location, market conditions, and real estate characteristics from information that was previously and separately collected. The distinguishing feature of an AVM is that it is a market appraisal produced through mathematical modelling. Credibility of an AVM is dependent on the data used and the skills of the modeller producing the AVM" (IAAO, 2015)*

For mass appraisal, researchers need to understand the relationship between the property value, the property attributes and urban social and economic setting in the region. The market behaviour is influenced by factors such as the property prices, the condition of the property asset and the geography of the location (Robinson, 1979; Harvey, 2016). Hedonic price modelling defines an econometric relationship between the price and the property characteristics, particularly in a residential context. Often the standard MRA (Multiple Regression Analysis)-based hedonic price models are not suitable for accessing and evaluating all of the data required for the value assessment model. Research on developing a tool to perform this more robustly and accurately is important in markets where data is available. However, in many regions, standardised data is not available or sufficient, parameters are not available or sufficiently robust. particularly for newly formed regions or regions undertaking major regeneration. Then Automated Valuation Models (AVMs) requiring a low deterministic relation between property values and attributes may be necessary. MRA-based approaches remain the key theoreti-

cal framework currently represent the mainstream approach to mass property appraisal. New approaches have emerged thanks to the advances in Artificial Intelligence. Some authors such Kauko and dAmato (2008) called those approaches as heretic because of divergence from model based approaches such as MRA. These heretic estimation techniques are not based on statistics models and include approaches such as artificial neural networks (ANNs), fuzzy logic, Rules-based reasoning, Self-Organising maps (SOM) and case based reasoning (CBR). These methods help elicit the core estimation rules of thumb used by experienced human values from datasets but maintain a mathematical basis to uncover the heuristic rules from property value datasets. Such a valuation model can be more accurate and quicker than its formal, regression-based competitor. Fundamentally these new methods are about Pattern recognition in a non-graphical way, though graphical pattern recognition through the application of computer aided design (CAD) methods such as Parametric CAD with geographical information systems are another relatively untested method within this field. A number of authors propose interesting but not totally causally clear methodology (Jenkins et al., 1999; Mccluskey and Anand, 1999*b*). In this research the author explores the opportunities within this under-theorized problem area (Kauko and dAmato, 2008). Specifically, this study investigates the current state of mass appraisal approaches from two different points of view: the theory and the practical application.

## 2.2 Automated Valuation Models

Another contributing factor to the growth of mass appraisal was the advances in computing hardware, better data collection, storage and the availability of

user-friendly software. Hence, this study will be focusing on the aspect of mass appraisals in automated valuation models.

### 2.2.1 Worldwide use of AVMs

Majority of Automated Valuation Models (AVMs) apply statistical models on recorded transaction of the properties. This will give the estimation value of the type of residential property. Although some argue that automated appraisals may be less accurate, the benefit is cost efficiency. There are several aspects where stakeholders will forego accuracy of prediction for cost. For example, in less risky appraisals, such as providing some additional loan advance or refinancing packages, banks used AVM services in their underwriting (Fitch Ratings Structured Finance, 2012). Due to the large number of such activities, AVM is cost-effective for tracking collateral values underlying the banks portfolio of mortgage loans by risk managers. This tracking or risk profiling was required by banking operations and regulation in an internal basis.

For an external basis, AVM is used in securitising mortgage loans, where it forms a fund which is pooled. This AVM process updates the loan-to-value ratios, and this information is passed onto rating agencies. It addresses the probability of defaulting overall risk of loss for the loans (Moodys Investors Service, 2012). AVMs can also help potential property buyers to anticipate the expected price for their purchase, or existing property owners in a fast appraisal to decide on relocation decisions. Government agencies who are dependent on public funds benefit from this cost-effective appraisals for taxation, planning and land use (IAAO, 2015). In other words, AVMs are fast and less costly and it also provides visibility of the residential property markets.

AVM services are usually provided by external parties and the methods in them are proprietary information. In the UK, examples of AVM services are Hometrack and RightmoveData. Despite the mechanism of such AVMS are not detailed in the literature, the basic building blocks are still discussed. Academic papers demonstrate cover property characteristics and how they are linked to marginal valuations (Hill et al., 1997).

There are papers covering statistical models in price valuations (Bin, 2004; Case et al., 2004).There are some problems when statistical models are used in AVMs for predicting property prices because mostly they are transforming the data using log prices. It may fit the statistical model but not the actual scenario. Such studies are in the line with the examination of out-of-sample performance focusing on selecting the best statistical model for a given data-set, but it still lacks on how to implement them successfully in an AVM.

### 2.2.2 AVMs in The Chinese Housing Market

The rapid development in Chinese property market provides an ideal avenue for the study of appraisal systems. The adoption of electronic platforms for appraisal to integrate property market services, legal services as well as mortgage and bank lending to increase efficiency began in late 1980s in China (Sun, 2013). Since then, the appraisal industry has grown much rapidly and by 2016 there was the formal China Appraisal Law was passed to standardise the burgeoning appraisal industry. Overall, the appraisal industry in China is valued at 12 billion RMB with more than 3000 appraisal firms employing approximately 100,000 individuals in the country (Sun, 2013). The application of mass appraisal for property in Shenzhen was conducted in 2003 and Beijing

soon followed suit by establishing a cost method using the automatic valuation model 2005. In 2009, Nanjing, Dandong and Hangzhou developed their mass appraisal system for housing transaction tax pricing models based on indirect, direct and hedonic price model respectively; whereas in Shanghai and Chongqing, the automatic valuation model was implemented in 2011 (Wang and Liu, 2014). Hence, it is clear that the rapid development of Chinese property market along with the appraisal industry provides an excellent avenue for scholars to examine the data and determine the best methods possible to study AVM service in China.

From the literature, it seems that the study of tax base valuation system was established sometime during 1999 (Ji, 1999), however the model did not incorporate the essential market environment components that are essential in real estate appraisal method and procedures. This led to the study of mass appraisal system (Ji and Fu, 2005) as well as the research in the application of AVM in mass appraisal system by Ji et al. (2006). The introduction of the application of artificial neural networks in the field of real estate valuation occurred in 1988 (Sheng and Tang, 1998) with the application of back propagation neural network as part of the urban property rental valuation system (Yang, 2002). Further research to improve the algorithm continued to grow within this field. For example, Yang and Liu (2004) presented the Euclidean distance to improve the recognition method of back propagation artificial neural network; while Zhang (2011) incorporated the rough set theory within the back propagation artificial neural network context to remove redundant information that might otherwise occur between real estate price factors. In 2014, there were two studies that explored a combination approach with the first being a study conducted by Mo (2014) that combined particle

swarm optimisation along with neural network algorithm; while the second study was conducted by Lu (2014) explored real estate valuation with the use of neural networks and genetic algorithm. In 2015, Wu and Deng explored the valuation and verification of real estate tax base system of ten districts located within Wuhan by modifying and improving the particle swarm optimisation algorithm introduced by Mo (2014).

It clear that based on the empirical research above that the application and improvement of the artificial neural networks models (e.g., fuzzy mathematics, expert scoring, back propagation artificial neural networks, genetic algorithm, support vector machine and particle swarm optimisation) appear to be an ideal avenue to explore the real estate tax base valuation theory for improving the robustness as well as accuracy of the appraisal of real estate. The current issue with real estate taxes either residential or commercial is that the tax is applied within the holding part of the property or organisation. With the continuous reform in taxation system, the automatic valuation model proves useful when examining mass appraisal of real estate especially with the constant development of Chinese real estate market in pilot cities that ensure further standardisation of tax behaviour and enhancing credibility of the real estate tax base valuation system.

## 2.3 Traditional Econometric Approaches

### 2.3.1 Standards Approaches

Traditionally real estate valuation has been a qualitative field with surveys conducted by experts, but the need for consistency has led to more struc-

tured and quantified approaches being recommended through international and national trade bodies. The question of whether more quantitative technique from computer processing would be better suited to this complex field has been raised by researchers such as O'Roarty et al. (1997) and McCluskey and Anand (1999*a*). The identification and consistent application of rules derived from human experts (heuristic) may provide a better approach. The two main models explored in research and application of real estate appraisals both utilise MRA, which are the statistical drive approach and the hedonic approach.

Hedonic models are the most common applied in property valuation, and for understanding trends in the housing market. These models utilise the property and plot attributes (size, state, material etc.) as well as situational aspects such as the local environment (e.g. access to bus and train services). Additional variables such as inflation adjustment and regional planning strategy (Miller, 1982; Kanojia et al., 2016) are often considered. Hedonic price model makes econometric analysis possible with large databases of price, property characteristics and regional characteristic, which describes the property and its surroundings. Theoretical statistical approaches, often regression-based analysis provide tools for a diverse set of applications in many countries that are digitally enabled, requiring minimal effort to add many spatially aggregated data.

MRA-based house price studies have been devised, tested and applied in a multitude of studies as reviewed by Ball (1973) and Lentz and Wang (1998), which are not solely limited to hedonic price theory. These studies mostly test performance based on model fit (i.e. the R-squared value), and the model significance as shown by the F-test.

Hedonic modelling has been mostly applied on constant-quality price indices (Hoesli, Thion and Watkins, 1997; Kanojia et al., 2016); determining rental values (Hoesli, Giaccotto and Favarger, 1997; Stevenson, 2004); estimating the problems associated with traffic (Wilhelmsson, 2000; Theebe, 2004); and estimating prices (Laakso, 1997) with some success, though the linearity of hedonic functions must be questioned due to the non-linear behaviour of some hedonic functions. Des Rosiers et al. (2000) explored valuation variability with effects from power transmission pylons and with middle schools, and found severe non-linear behaviour in valuation. For example, the visibility of pylons was not linear, and the trade-off between easy accesses but the traffic generated by the school caused non-linear valuations (Des Rosiers et al., 2000)

In hedonic modelling research proxies for location variables have a few interpretations. For example accessibility measures were addressed in surveys by the following authors (Pavlov, 2000) and the more recent work of Des Rosiers et al. (2000). Kauko (2003) provided research on locational proxy variables influencing the price formation of residential property. Kang and Reichert (1991) devised a variable termed location quality based on the price per $m^2$ of the living area. Similarly, a solution that assigned the location by value was achieved through the ward-variable. This was devised by Mccluskey and Anand (1999*b*) with the values based on the average transaction prices for the ward. Using real estate data from 351 properties within a city, Jansen et al. (2011), concluded that the success of a valuation algorithm dependent on the end application, and showed that only a few variables out of the 97 measures were useful in determining price. Some examples (Rossini et al., 2008) were:

1. For rating and taxation authorities: Weighted mean and geometric mean

ratios, Coefficient of Variation (COV), Standard Error, Price Related Differential (PRD),Coefficient of Dispersion (COD)

2. Financial institutions: Forecast Standard Deviation (FSD)

3. Academic papers: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE)

Hedonic model often lacks suitable variables in practice and data to capture value changes according to the environment changes and are also affected by the issue of spatial resolution. Modern geographic information systems (GIS) have provided new tools for the complex analysis of value differences in regions. GIS-aided analysis has been explored to address these issues with additional benefits in data visualisation and storage and spatial analytics. GIS aided approaches are also inherently more user friendly in many cases because they provide a visualisation capability in their analysis. A GIS addresses spatial discontinuity when a spatial dimension is applied, better than in normal hedonic model (Orford, 2017; Wyatt, 1996*a*,*b*).

To improve valuation methods, combining different approaches is common. LIU et al. (2011) combined a GIS system with a BP neural network to create a hedonic valuation model for mass appraisal. Transaction data drives the hedonic price model; the BP neural network uses spatial data; attribute data; social data; economic data etc. A three layers perceptron ANN network with one hidden layer was used to train the network from the sample data. Claims of better prediction though, were not backed up with data. Four years later, Bohari et al. (2015) demonstrated an improvement in the OLS (Ordinary Least Square) property valuation model in association with a GIS for the state of Perlis, Malaysia. They provided a property value map for visualization

though the GIS which aids understanding and planning. Visualisation via a GIS also helps in identifying variable that may have not been considered or collected in the original data.

Tian and Yang (2014) explored combining a GIS with Hedonic modelling for the Chinese real estate market. They discussed advantages of GIS system such as the ability to easily include spatial analysis, storage and management functions. They used residential buildings information from 270 units in Xian (Tier 1 city in China), with visible buffer zones in the GIS to represent property less than 1000 meter from a service facility, i.e. hospital. They claim that this increased the accuracy of the appraisal model. This method may bring better results but is crude in its use of value enhancing features. The work of Des Rosiers, F, Theriault, M, (2008) showed that such buffer zones were usually non-linear internally. Close to the hospital would be disturbance from ambulances etc, further away may cause access problems and suffer from less convenience. In between there was a sweet spot where these balance and property values are higher.

Fu et al. (2015) proposed and tested a geographic data based method by valuing according to the local neighbourhood area (termed individual), adjoining estates (who they term peers) and business resources (zones). They proposed a ranking and clustering system (ClusRanking) to exploit geographic dependencies between the three defined zones in a probabilistic ranking model. Their key innovation was the use of taxicab trajectory data to capture neighbourhood popularity and local business zone rating to model the influence of the business zone onto valuation.

Borst and McCluskey (2007) used geographically weighted regression (GWR) in the mass appraisal environment reviewed location based technol-

ogy studies with market segmentation. They claimed that improvements in segmentation have positive impact on performance. GWR was proposed to detect market segments in their studies, which the GWR produced a regression equation for every point in the dataset filled with transactions. However, they comment that retesting on other data set is necessary, especially for different regions and different economic markets.

GIS systems are not designed to incorporate complex spatial statistical methods. Thus such analysis is often undertaken in software that is loosely coupled to the GIS. The GIS can for many datasets be used to store data and for visualisation though this may be problematic for very large datasets such as the ones available for this research (whole regions or countries). A common practice in processing GIS data, is to import it into a statistical software package where methods such as standard or multi-level regression analysis can be undertaken before transfer back to the GIS for visual inspection and analysis. Outputs from the GIS represent a surface showing the effect of specific (or collection) variable on price. Studies by many authors show in different ways that the importance of location is high (Orford, 2017; Lake et al., 1998; Rodriguez et al., 1995; Wyatt, 1996*a*; LIU et al., 2011; Tian and Yang, 2014). GIS-aided property value analysis is feasibly linked to standardised and broad information databases in the literature. Though whether this continues to be the case in future is an interesting question.

A number of more established spatial techniques provide options for addressing hedonic-based house price analysis with location features. For example, the evaluation of sub-clusters within a region and calculating spatial changes may aid improvement significantly (Orford, 2017). This involves proper methods for accounting for non-linearity, considering the dynamics of

topography. That is, novel processes containing hierarchical specifications with spatial models that could be scaled. Kriging and its variants are considered to be higher-order spatial statistical methods that was built on parametric MRA.

Bourassa et al. (2007) compared spatial statistical models and ordinary least squares (OLS) model using neighbourhood dummy variables. They used a database which contained residential sales data in Auckland, New Zealand. In multi-level specifications, each attribute and its price effect is assessed at an appropriate level. In property valuation, the appropriate levels may be neighbourhood, street and property. Thus the effects of spatial variations in the housing stock and of the place itself can be differentiated (Orford, 2017).

In a spatial expansion model, the effect of housing attributes on the calculated price can change over distance. This creates a set of interrelated sub-markets with varying boundaries. Many reported studies utilise such a feature sensitivity analysis were performed to parameters in order to test the heterogeneity of the property market (Geoghegan et al., 1997; Wilhelmsson, 2002). Riddel (2001) enhanced the hedonic framework, catering for the interaction between space and time, since neighbourhood munificence features may create market disequilibrium. Francke and Vos (2004) used Kalman filter to study the trend of property markets in a hierarchical manner.

Many techniques consider a lack of knowledge of the independent variables that cause the inability of results to be generalised to a wider population (Michaels et al., 1990). Segment resolution is not sufficient to accommodate the environmental factors present. This may be access to shops, travel hubs, scenery etc. However, the conventional approach to hedonic analysis still considers market segmentation (Simonotti, 2009). A hedonic regression cannot detect articulated boundary conditions. It can cater for the significance of the

sign and the weighting effect of the value factors. The reliability and strength of association within the total valuations are also calculated. One way of addressing the problem is to use dummy variables (Laakso, 1997) or computing hedonic quality ranks for each valuation (Rothenberg et al., 1991). Another approach. as adopted by the authors Chinese data partner, is to built many separate model versions by partitioning data on a separate hedonic expression. Then models are tested against the latest data and the closest fit model is chosen to evaluate prices for that subset. Hence data is clustered into different partitions, which are predefined or synthesised. If segmentation in general is not conducted, the partitioning approach may still be applied (Needham et al., 1998).

### 2.3.2 Flexible Regression Approaches

Ripley (1993) used the term flexible regression instead of non or semi parametric regression. Flexible regression methods lie between newer heretic or radical methods and orthodox approaches. Verkooijen et al. (1996) confirmed and developed the term for:

- Local approximations E [Y X = x] with splines or kernel, which are typically weighted.
- Additive models that can handle both parametric and non-parametric terms.
- Computation with adaptive capabilities such as metaheuristics and neural networks.

For estimating values of real estate, the flexible regression approach has

been explored by many authors such as Pace (1995) and Mason and Quigley (1996). Coleman and Larsen (1991) however were critical towards those approaches for value estimation. In extending the approach some researchers attempted to extend this to the spatial aspect explicitly. Pavlov (2000) worked apply space-varying coefficients, which combined the two spatial techniques, the error method or the lagged method, while incorporating location as a non-parametric factor.

There is no strong theoretical support on which functional form to select, giving the empirical researchers some form of flexibility. In this case, non-parametric and semi-parametric approaches offer benefits for the case of estimating non-linearities. The generalised additive model (GAM), which was developed by Mason and Quigley (2007) smoothed the price-distance curve. A key advantage of this method is the comprehensibility that it affords in generating insights.

Colwell and Munneke (2003) applied a similar approach but from a non-parametric aspect, which derives price functions in non-urban areas. The shape of a spatial price curve can take many forms which is a feature for non-parametric estimates. The official name for this approach is the piecewise parabolic multiple regression analysis (PPMRA), assuming no curves between valleys and peaks. Grid lines on a map takes up the spatial units, which is a continuous function. Colwell and Munneke (2003) applied bary-centric coordinates as independent variables a procedure that applies standardisation and weighted averages to convert absolute values into relative values. This adds efficiency to the method. The regression coefficient of the points represents the vertex relative to the constant term. This means sacrificing between increased continuity versus a loss in flexibility. Their work was considered novel

at the time as it supports a sufficiently comprehensive tool for general real-estate markets, and applicable in mass-appraisals if the data is suitable. A key advantage is it avoids extremely smooth point estimates that may hinder applications that require semi-parametric support.

Wang et al. (2014) introduced a Unity valuation approach by providing the sales comparable method as suggested by the China Institution of Appraisal and Agent. Their investigation is mainly about characteristics of the property market in Shenzhen. They claim that the Unity Valuation model would be acceptable and reliable for mass appraisal in China.

Flexible estimation as an alternative model based on parametric methods is less efficient than strict parametric regression but avoids model based problems as it estimates the distribution from the data itself. Some assumptions that flexible regression methods are based upon are open to debate:

1. Property price variances caused by taxation, zoning or by physical land features are smoothed by arbitrage, which is obviously true over time but not necessarily over land surfaces.

2. Or that models can incorporate such discontinuities.

However property prices are not continuous in space, and this is a problem with the flexible regression method as it is still based on an orthodox framework. Standard real estate models and city/land economics understanding does not cater to spatial non-linearity readily. Time may provide reasonable linear smoothness in a model, however distance rarely does.

## 2.4 Machine Learning Based Approaches

Developments largely driven by advances in computer technology and artificial intelligence, in particular, have caused researchers to create a number of new approaches to the challenges of mass valuation. Traditionally these approaches are termed heretic as they do not derive directly from multiple regression based methods. From a number of areas, these techniques create a new experimental approach to research on mass property appraisal. Though seemingly very different in philosophy, language and method they do have a partial overlap with the historical methods discussed earlier.

### 2.4.1 Artificial Neural Networks

Artificial Neural Networks (ANN) attempt to solve problems by replicating the neural network structure in the human brain. Verkooijen et al. (1996) and others such as Pace (1995) and Mccluskey and Anand (1999*b*) were part of the first wave of researchers in this domain who viewed ANNs conceptually as a type of flexible, non-model based regression. A neural network can be considered as a black box, with no clear response relationship between inputs and corresponding outputs. The relationship behaviours are generated by training the ANN with example data. The basic structure in an ANN is neurones that act as nodes which connect inputs to other neurones according to weightings. Taken together, neurones process external inputs, or internal inputs from other neurones to generate an output value. The inputs signals are numerical values. Resulting output values are adjusted iteratively, until the system translates input combinations to the right output combinations with the desired accuracy. ANNs have three core types of structure, feed-forward, feedback and compet-

itive architectures. The most common is the feed-forward architecture. A feed-forward network known as a multi-layer perceptron (MLP) has a similarity to multiple regression, but lacks the clear mathematical relationship between inputs and outputs, though there are exceptions (White, 1989).

McCluskey et al. (2012) compared the performance of an artificial neural network (ANN) and several multiple regression techniques in terms of their accuracy and capability in the mass appraisal industry. They used the data collected by Lisburn District Council, Northern Ireland (The data contained 2694 use-able residential property sales data during the period 2002-2004).

ANNs did offer a simplified modelling approach, but ANN was not sufficiently transparent to provide predicted values and maybe subject to severe objection. Nonetheless, many research studies have shown ANNs to be effective for classification in fields such as finance and economics. Overall their results have been similar to more traditional statistical methods such as logistic regression or linear discriminant analysis (Verkooijen et al., 1996). Scholars in mid 1990 identified problems with the ANN approach (Worzala et al., 1995) because of cocern over how prices were formed in the blackbox, which directed mass appraisal research away from this technique. However, later Nghiep and Al (2001) among others have shown that this approach is efficient for large heterogeneous data sets.

Opinion on the value of ANNs is divided. Tay and Ho (1992), Borst (1995), Worzala et al. (1995) have all written on the subject. Objections are largely due to the lack of robustness viewpoint as the ANN is still a black box, though producing comparable predictions. The work of Nguyen and Nghiep and Al (2001) compares a Multi-Layer-Perceptron (back propagation ANN) and MRA, using the mean absolute percentage error of the model and forecast-

ing error with thresholds of 5 and 15. This measures how many predictions fall with a percentage value of the actual sale value. The MRA model improves with improvement in the functional specification improves. The MLP model improves as the training data set size is increased. Kauko (2002) compared SOM and MRA in a similar way, with the ANN-based model being superior to MRA models with a small number of data samples in recognising a hypothesised impact between externality and price. Conversely, González et al. (2002) found accuracy performance for ANN was marginally better than with MRA models.

### 2.4.2 Self-Organizing Maps (SOM)

Self-organizing maps (SOM) sometimes known as Kohonen maps are a type of ANN that generate clusters of features, where each feature is a characteristic combination of attribute levels. A SOM generates a surface where areas with similar groups of variables can be viewed but also compared with surfaces showing differing combinations of variables. The SOMs also generate a typical value at any node for a given surface feature. Visual approaches such as SOMs have the benefit of allowing for qualitative analysis in support of the quantitative analysis.

There are many similarities between the SOM and the flexible regression methods. Colwell and Munneke (2003) for example compared their SOM (known as PPMRA) method with kernel estimation, a type of flexible regression method which was very similar to the SOM. In essence, the same criticism may be applied to SOMs as to flexible regression. There are a few differences of note. One such difference is that though the kernel and the SOM force

a priori lattice like surface structure, this is smoothed from the valleys and peaks in data values. Normally in the piecewise parabolic method the peaks and valleys are used as vertices of the grid. With SOMs, this can be corrected by creating a sufficiently large resolution that allows the high and low values to be fitted with the response.

The accuracy of SOM and MLP are both affected by the decisions made on issues such as:

1. Selection of field ranges for the input variables, to avoid domination by one factor.

2. When should the run based on trial and error be stopped?

3. Its hard to know when an optimum solution has been found.

4. The number of variables to be considered the optimal solution.

5. How to identify the causality behind the correlations generated by the black box?

Geographically Weighted Regression (GWR) and Self Organising Map (SOM) techniques have similarities. In both models, proximity and intensity, determines the response. With GWR for any data point the output is heterogeneous and for a SOM is heterogeneous for any cluster of points in the data. With a GWR models, this influence is in two dimensional geographical X, Y space and for a SOM the input variables (n) define it is in the n-dimensional space. Thus a SOM with only two inputs with much higher weightings will produce an output like a GWR.

### 2.4.3 The Genetic Algorithm (GA)

The genetic algorithm (GA) is a stochastic search approach from the machine-learning toolset designed to explore wide and complex spaces. Genetic algorithms have been effective in evaluation of large spaces by performing the transformation of population by replacing them from a previously generated population, simulating natural inheritance of attributes from parents to children. The GA search algorithm consists of a few phases: initialisation, selection, crossover, and mutation.

Applications of Genetic algorithms are scarce the property market domain. In terms of combining GA with other models, Ahn et al. (2012) used ridge regression with GA to improve real estate appraisal for the Korean property industry. Ma et al. (2015) proposed a combination approach of hierarchical GA with the least squares method to optimize an ANN for predicting property index. Manganelli et al. (2015) tested the relationship between real estate prices and property location through GA. De Mare et al. (2012) applied GA to investigate how high speed railways influence the real-estate market. Genetic algorithms (GA) may have practical value in real estate valuation as they have advantages highlighting movements in sub-clusters within an overall property market (Del Giudice et al., 2017). Mccluskey and Anand (1999*b*) coupled GA with ANN and termed it as an intelligent hybrid system, which combined the augmented nearest-neighbour algorithm used as a simple planning technique.Though data mining analysis is helpful in the search for patterns in a search space, González et al. (2002), and subsequent work, maintained that those techniques were very dependent on the data you used in them.

Del Giudice et al. (2017) used GAs to identify the effect on real estate

rental values from the location of housing units in the Naples, Italy. They tried to assess their reliability for real estate appraisals, They used a multiple regression analysis (MRA) to compare results from the GA and MRA for estate rental values. They found that the results obtained by the GA were excellent (absolute mean average percentage error equal to 10.62%). However, they did not significantly improve on traditional parametric approaches, as forecasting results for GA was 7.65% above that for MRA.

### 2.4.4 Rule-based Expert Systems

Rule-based Expert systems have emerged as a quantitative approach and as a counter paradigm to ANNs and other numerical techniques. This approach tries to capture from human experts, structure and then apply rules that model human judgement. Capturing the heuristics that humans use was their biggest drawback: human experts could if willing formalise their explicit rules, but often they are unaware of implicit rules that they apply subconsciously. Thus rule extraction from human experts could be drawn out but costly (Verkooijen et al., 1996). The resulting models can be incomplete and inconsistent, thus they may not offer robust solutions. Nevertheless researchers have explored their value in property domain. Scott and Gronow (1990) explored the components of valuation expertise and suggested the production of a property valuation expert system. Mccluskey and Anand (1999*b*) saw knowledge elicitation and simulation of human expertise as strengths, but the resulting models were weak and lack robustness.

### 2.4.5 Case-Based Reasoning (CBR)

Case-based reasoning (CBR) identifies similar past cases to the one to be assessed. For property valuation Case Based Reasoning (CBR) is based on historical transactions. It does not directly capture heuristic knowledge, but infers it from the relationships embedded in previous cases. Thus variable selection relies on the data and is inherent in each case. The method works from a case library of previous similar task cases, it then finds the most similar cases and draws conclusions about the new case from them (Tan et al., 2006). O'Roarty et al. (1997) stressed that the real challenges were in the process of choosing similar cases and in making judgements from them to the new situation. According to Wyatt (1997) Case-based reasoning was considered to be one of the more promising methods, because it considered how to mine and extract and update new knowledge. Objectivity and justification are the strong points of this method. It eliminates the problem of obtaining information from experts associated with expert systems. However, CBR requires large reliable datasets to be effective as described by Mccluskey and Anand (1999*b*) has shown.

### 2.4.6 Fuzzy Logic

Fuzzy logic helps address the imprecision of the present. Conceptually it is different from probability based methods which address the uncertainty of the future. Fuzzy logic helps define for an entity a measure of the degree of membership in a set. Fuzzy logic tries to capture and represent the uncertainty in evaluating alternatives and making decisions that exist in the real world. also, define it which can be used in computerised applications. Since human

real estate valuation is full of uncertainty and mass property valuation requires computational resource Bagnoli and Smith (1998) suggested it can be applied to the domain. Fuzzy Logic is about computing with words (Bagnoli and Smith, 1998) and can reduce the subjectivity caused by the human appraiser and provide flexible adjustment of key factors (Lee et al., 2003). They proposed fuzzy linguistic functions with overlapping membership to replace the absolute values used in other methods. They argued by replacing crisp memberships functions with a fuzzy function, the basis for the valuation was improved. This was because normal numerate methods lose information when dealing with imprecise information (Sui, 1992).However for Sui (1992), fuzzy set theory had problems in "how to define the correct membership function for spatial purposes, as the membership function is derived ad hoc."

Guan et al. (2014) tried address real estate valuation through a Fuzzy Neural Approach. This is in response to the limitation of MRA method: non-linearity, multi-collinearity and heteroscedasticity, where they offered an adaptive neuro-fuzzy inference system (ANFIS). They stated that there was a lack research with ANFIS modelling though they proved the ANFIS approach might be useful for classification and estimation in a range of fields, including mass appraisal. They had access to a massive data set, whereas a small data set is more common in most studies in this domain. They demonstrated their method with tax assessment office data for a large mid-western city of the USA. Their results showed the superiority of ANFIS over MRA in mass property appraisal for tax valuation purposes in all their test scenarios.

### 2.4.7 Rough Set Theory (RST)

Rough set theory (RST) seems ideally suited to real estate valuation as it extracts Boolean rules from real market data like CBR and did not consider expert knowledge as in the Rule-based expert systems (D 'amato, 2007). Nonetheless, it still retained the qualitative strengths of expert systems and the CBR strength though requiring more real data to model the rules with improvements in its performance.

## 2.5 Conclusion

Having reviewed the methods that potentially could be applied to real estate mass evaluation, this study finds that ANN possess the flexibility necessary for scaling up for larger cases and has great adaptability. As with MRA and flexible regression methods, ANN has the ability to be combined with other techniques, thus forming hybrid systems that is core to various fields in artificial intelligence (AI). It is therefore natural to consider ANNs as a mean to be used in Automated Valuation Models (AVM) for mass appraisal. The next chapter will discuss in greater detail the mechanics of ANN and why it would be suitable in the context of this study.

# Chapter 3

# ANNs and Modelling

## 3.1 Introduction

The main topic for this chapter will cover some common knowledge and theory regarding the background and modelling of ANNs. It is important to analyse the ANNs modelling under a complex real-world problem especially in the field of residential property appraisal. Hence the possibility outline for ANNs modeling is necessary to define the network topology, activation functions between layers, and the training algorithm of ANNs.

## 3.2 ANN Topology

An ANN topology includes the three main layers which are the input, hidden and output layers. There is direct relationship between the number of neurons and the number of the input variables. The number of layers for the hidden layer could increase to more than one when there are a number of neurons. However, the number of neurons in the output layer equals to the number of

outputs (the output size is equal to one for this research case). The performance and design of ANN is affected by the neurons in the input and hidden layers. This is supported by Zhang et al. (1998) who stated that by comparing the number of neurons in the input layer and the number of neurons in each hidden layer, the former had a stronger presence in terms of developing a forecasting model. To ensure that a neutral network model performs an out put with better accuracy, the sufficient number of input will be the key. As both too little or too much inputs variables would cause the performance of a neural network model to produce an unreliable and unsuitable output.

### 3.2.1 ANN Basics

It is better to simplify a neural network topology by using a number set notation rather than an actual diagram. Under the number set notation, in a neural network, inputs represent the first number, outputs refer to the last number, all other middle number in every hidden layer could refer to the number of neurons. Multi-layer Perceptron (MLP) is a supervised learning algorithm which can be taken as an example to explain the description, a MLP(8;16;1), can be defined as total of eight inputs, 16 neurons in a hidden layer and one output for this neutral network.

Figure 3.1 will explain in details of general number set notation for a neural network topology. For example, While for a MLP with (5;4;2;1) as shown in Figure 3.2, it represents a neural network with four inputs, eight neurons in the first hidden layer, two neurons in the second hidden layer and three outputs.

Besides input, hidden and output layers, ANN model also consists the

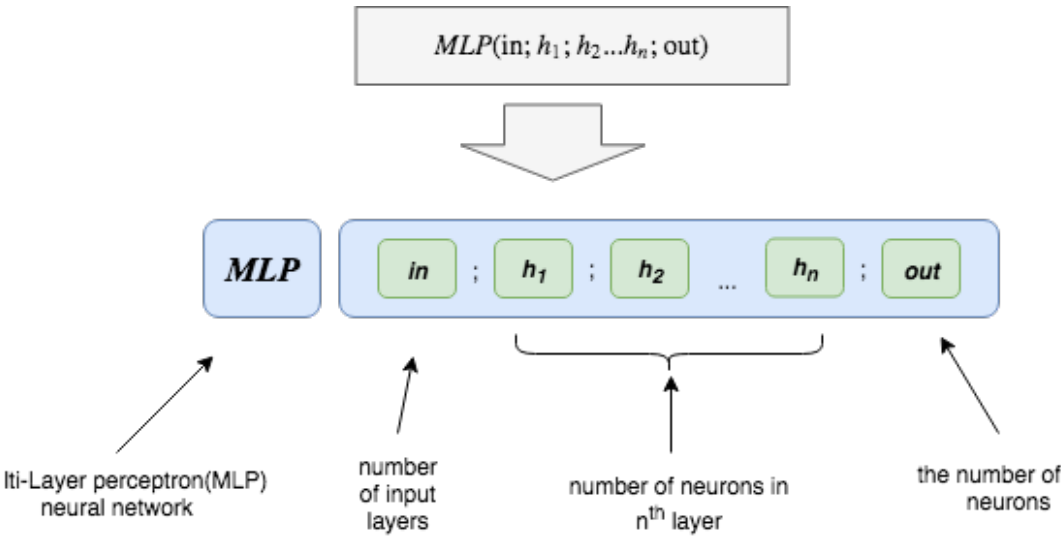


Figure 3.1: Notation of A Neural Network Topology

type of training required by the network to learn. In section 3.4, the two training types, which are supervised and unsupervised learning are explained in details. On the other hand, there are differing ANN models which have no hidden layers. SOM neural network is an example of that. The MLP network topology shown in Figure 3.2 indicated the connections and layers of the neurons. The unprocessed neurons in an input layer will carry over the inputs to the following layer based on the number of input variables. Under the hidden layers, the weights of each processing neuron are adjusted by using the error. The calculation is done at the end of every iteration under the training set through the comparison between the estimated output and the ideal output. A trained neural network could take hundreds of decisions. Supported by Jeff (2011) that neural network would be more effective having considered the main role of a bias neuron, which shares the similar function of hidden neurons. However, a bias neuron never receives input from prior layer since it outputs a constant of one, which is not the same as other neurons in

| Layer Name | Neural Number |
|---|---|
| Input Layer | 5 |
| Hidden Layer 1 | 4 |
| Hidden Layer 2 | 2 |
| Output Layer | 1 |

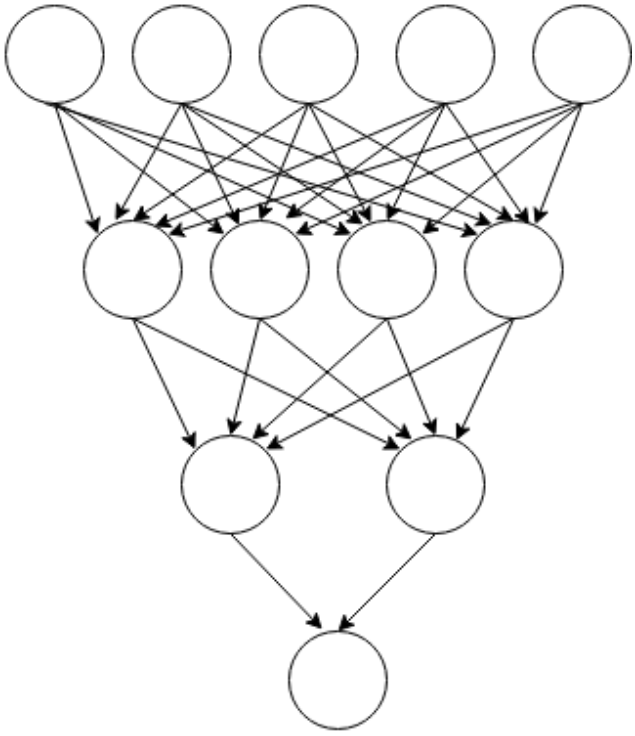

Figure 3.2: An Example of Neural Network Topology

a neural network.

### 3.2.2 Input Layer Neurons

Since the number of input layer and variables could provide extra flexibility in terms of model design, it is noteworthy to extract as much information in order to keep the number of neurons in the input layer to be the same as the number of input variables. Furthermore, Iit is important to firstly remove all the outliers by using the z-score method or any other outlier detection method. The advantages of this process is to improve the accuracy of ANN model by selecting the most important variables, hence to detect and eliminate the least

sensitive ones.

### 3.2.3 Hidden Layer Neurons

Even though it was proven by many studies that a single hidden layer might be enough for a neural network to approximate any complex nonlinear relationships (Do and Grudnitski, 1992; Yeh and Hsu, 2018), a neural network still could manage to obtain as much hidden layer as the model requires. The disadvantage to determine the systematic trial-and-error for every hidden layer with large number is very time consuming taking up to several weeks of time.

### 3.2.4 Output Layer Neurons

It is easy to ensure the number of neurons required in the output layer as it matches the number of output variables. For baseline model and all proposed models, the number of neurons is just one (for the output layer), hence, it only requires one output.

## 3.3 Activation Functions

The tools to evaluate relationship between the input neurons and the output neuron are called the activation functions. When functions are added into the output end of the neural network, it is known as the transfer functions. There are many other activation functions which are most commonly applied would be explained in the following sub-sections. In this study, 10 fold Cross validation used to select to better performance activation function.

### 3.3.1 Sigmoid Function

$$\text{Sigmoid}(x) = \frac{1}{1 + \exp(-x)} \tag{3.1}$$

As shown in Equation 3.1 and Figure 3.3, one advantage of Sigmoid function that it has bounded output between 0 to 1, which could use to represent probability. however, is perceived to be the appropriate selected to apply in this research work study. The derivative of sigmoid activation function (used to adjust the weights) is simpler to calculate.

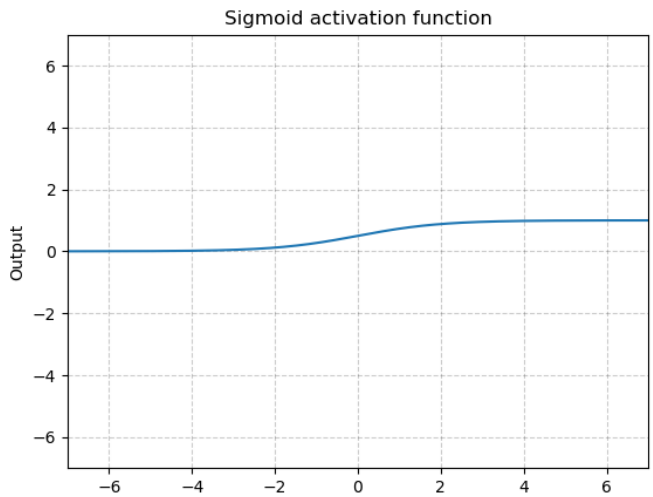


Figure 3.3: The response of a Sigmoid non-linearity

### 3.3.2 Tanh Function

Hyperbolic Tangent (Tanh) Function is similar to sigmoid function. Its range outputs between -1 and 1 as shown in the Figure 3.4 and Equation 3.2.

$$\text{Tanh}\,(x) = \frac{2}{1 + e^{-x}} - 1 \tag{3.2}$$

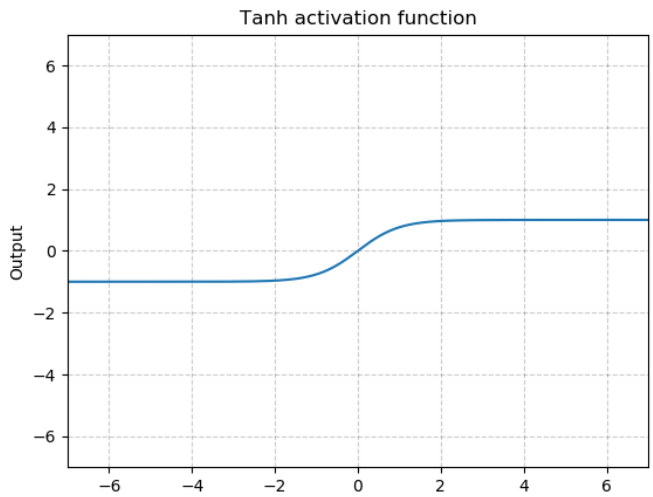


Figure 3.4: The response of a Tanh non-linearity

### 3.3.3 HardTanh Function

The HardTanh Function (as shown in Equation 3.3 and Figure 3.5) is variation from Tanh Function, where somtimes prefeerd over the Tanh function, beacuse of computational cost of HardTanh is considered be cheapter.

$$\text{HardTanh}(x) = \begin{cases} 1 & \text{if } x > 1 \\ -1 & \text{if } x < -1 \\ x & \text{otherwise} \end{cases} \tag{3.3}$$

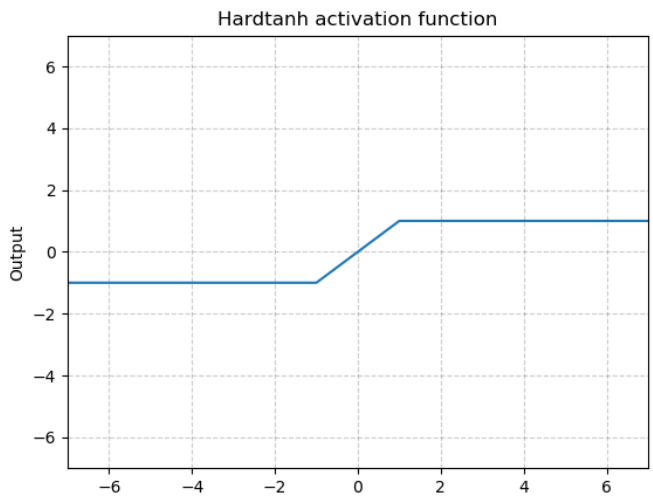


Figure 3.5: The response of a HardTanh non-linearity

### 3.3.4 ReLU Function

Apart from sigmoids, rectified linear units (ReLUs) is most recent deep learning networks for hidden layers. If the input of the rectified linear is less than 0, then output is 0. linear with a slope of 1 when x is greater than zero. (As shown in Equation 3.4 and Figure 3.6)

$$\mathrm{ReLU}(x) = \max(0, x) \tag{3.4}$$

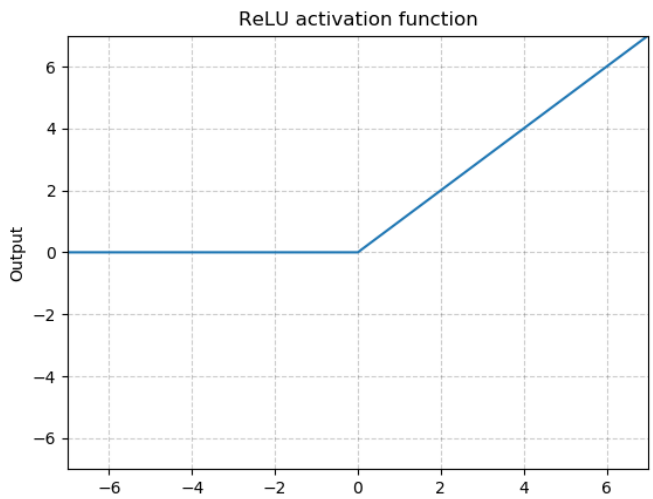


Figure 3.6: The response of a ReLU non-linearity

### 3.3.5 Leaky ReLU Function

In ReLU activation function, linear units hava a problem of "dying", if the input to a ReLU with its weight is negtive, the output will be 0. Leaky ReLU overcomes this difficulty, which demonstrate in equation 3.5 and figure 3.7.

$$\text{LeakyRELU(x)} = \begin{cases} x, & \text{if } x \geq 0 \\ negative_slope \times x, & \text{otherwise} \end{cases} \tag{3.5}$$

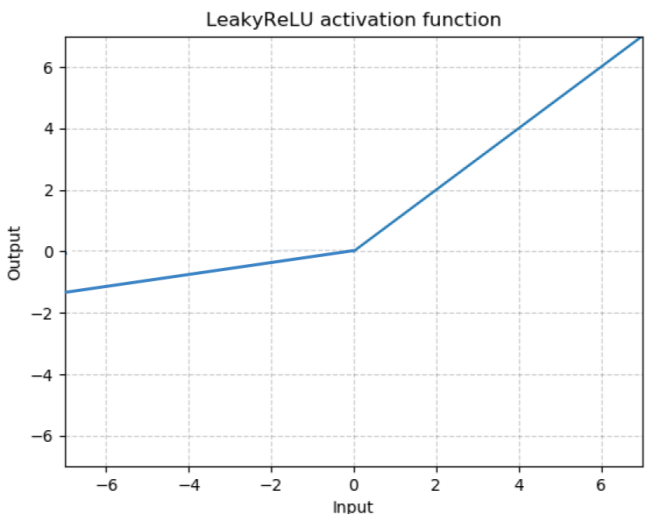


Figure 3.7: The response of a LeakyReLU non-linearity

## 3.4 ANN training Algorithms

ANN Training Algorithms are used for weight adjustment in a systematic manner on every hidden neuron within a neural network, which provides desirable outputs to a set degree of accuracy. After going through every epoch of the training process, the neural network will improve its knowledge of the surrounding. The Propagation training algorithm is an aspect of supervised training(LeCun et al., 2015). The algorithm goes through several iterations and every iterative loop, which passes the training data will affect the improvement in terms of the internal error rate or the RMSE. The training set determines the RMSE in terms of difference in percentage points between the computed output from the neural network and the ideal output. In terms of the training data, the weights are then re-calculated for each Asymptote Crossover point.

As stated by LeCun et al. (2015), the propagation training algorithm indicates how machine react on its internal parameters, as activated function. This is because the activation functions are used to compute the gradient for every neuron connection in the neural network. In the following sections, the training algorithms are explained into two section, supervised Algorithms and unsupervised Algorithms.

### 3.4.1 Supervised Algorithms

In supervised algorithms, the algorithm learning process takes starts from the training dataset. This is akin to a supervisor monitoring the actions of his or her subordinates. The supervisor knows the correct outcomes, and keeps track of the output produced by the subordinates, and corrects them if needed.

This process takes an iterative form until an acceptable performance level is achieved, thus learning stops.

### Backpropagation

Backpropagation is a form of Supervised learning because it provides feedback if the computed outputs of the network are correct. It is one of the first method used to train neural network. Backpropagation training algorithm needs a learning rate to be specified. This is a percentage that captures the degree of the gradient of an activation function that manages to apply the hidden neuron weights, which is also a rate for momentum allowing the neural network to overcome the local minimal points.(Rosenblatt, 1962)

### Mahattan Update Rule

Backpropagation training algorithm requires a careful adjustment of the hidden neuron weights. The training process oscillates when the learning rate is too much and it may fail at global minimum when the coverage is too little. The step value is the learning rate, which can be discovered by only the use of systematic trial and error. The Manhattan update rule only uses the sign of the gradient Jeff (2011), which is inherited from the gradient. The weights are adjusted by user assigned step values.

### Quick Propagation

The quick propagation algorithm updates information in batches by approximating the curvature of the error surface once all cases have their gradients aggregated. This is in contrast to the back propagation method where each case updates its network weights in an iterative manner.

### Perceptron Rule

There are many types of perceptrons which are explained by Rosenblatt (1958) and Widrow and Lehr (1995). When the existing weight required to respond to all the training patterns are correct, there is a need for the algorithm to adjust the weight for the correct value and then allow the net to respond to all of the training patterns appropriately within a set number of steps.

### Levenberg-Marquardt Algorithm

The Levenberg-Marquardt algorithm considers to be one of the fastest training algorithms as it is very efficient method for training the neural network Moré (1978). The Levenberg-Marquardt algorithm acts as a hybrid algorithm for the strength integration of both Quasi-Newton method and gradient descent (Backpropagation).Therefore, gradient descent is guaranteed to converge to a local minimum, albeit slowly. However, there are few restrictions when applying the Levenberg-Marquardt algorithm. First, it can only be applied to a single output network. Second, it can only be applied for situations with limited space proportion since the total number of neurons should be less than one hundred.

### Resilient Propagation

The difference between Resilient propagation (RPROP) and all other training method was that it could apply without the training parameters (learning rate and momentum) and compute the optimal parameters automatically. Therefore, Resilient propagation training algorithm was very user friendly and had been recognised as one of the best training methods available(Jeff, 2011).

### 3.4.2 Unsupervised Learning

Unsupervised learning algorithm could be associated with self-organisation and do not require a set of ideal output as data presentation is organised for the network and their emergent similar properties can be detected. The success of unsupervised learning depends on the effort of the user to complete the task when an appropriate neural network is designed that can be used to learn independent criteria when the neural network is required to learn. Take Hebbian rule and Self-organising rule (Hassoun, 1995) as examples, the neural network weights can be optimised for tasks with independent criteria.

#### Hebb Rule

Hebb rule is the earliest and simplest among the ANN unsupervised training algorithms (Herz et al., 1988). The basic theory for Hebb Rule is that once both neurons are activated, the synaptic strengths between the interconnected neurons will increase. Also McClelland Rumelhart (1988) had made some improvement to the original training algorithm, which allows the synaptic weight of the connected neurons to increase when they are shutting down.

#### Self-Organising Map

The Self-organising map is another form of ANNs (Fu, 2003). It applies the un-supervised learning as a training clustering algorithm (Kohonen, 1982). Compared with other ANNs, its topological properties for the input space are retained by using neighborhood functions. Hence, in the neural network, every neuron will compete with the input patterns. The comparison idea allows the input vector to compete with the weight vector and only one neuron will be

produced as the winner. The winning neuron will work together with connection weights for a better direction adjustment. Therefore, self-organising map is broadly applied in pattern recognition, such as image recognition, and cluster analysis.

# Chapter 4

# Research Design

## 4.1 Introduction

The purpose of this chapter is to explain the research design of the mass-appraisal system through the lense of automated valuation model (AVM). Hence, this chapter provides the research design for the study of AVM, it covers the data collection and sampling strategy, including the justification of selection of cities investigated. This is important because this study is solving the cross-city problem in mass-appraisal, where it seeks to use the information from mature cities to form valuation prices in non-mature developments. The type of variables and measures are also outlined in this study.

Hence, this study defines the boundary of the residential property market in china, where the rise of property development and transactions necessitates a novel implementation framework on mass appraisal system. Hence, this helps the study to clarify the purpose of the research so that we can select a suitable strategy for the study. This is particularly true in the present time and the feature of socialist market economy with Chinese characteristics.

## 4.2 The Automated Valuation Model (AVM) research design process

According to Schulz et al. (2014), There were four activities involved in the development and execution of an AVM service.

- First, there must be a continuous access to reliable data and database warehousing.
- Second, there must be a model to evaluate the data, and the outcome should be validated.
- Third, based on the results, there should be a roll-out and service provision.
- Fourth, the fourth activity is backtesting the results.

In establishing the first stage in this research, the data collected include listing information of the Chinese property market, the recorded transaction data from real estate agent, with some of the information confirmed by external local solicitors and banks who had acquired them during the mortgage underwriting process. Most of this data is itself proprietary, and the owner of the data are interested in setting up its own AVM. This research is part of a project called JinZheng intelligence appraisal system. This project is a collaboration between Cityre data specialist ltd, JinZheng real estate appraisal ltd and supported by China Institute of Real Estate Appraisers and Agents (CIREA). The aim of this project is to establish a tool for the mass evaluation of property. It is also supported by the Chinese government due to the

recent agenda on the concept of Internet Plus and using Big Data to create sustainable living standards(Wang et al., 2016).

In the process of establishing access to data sources, the main task is ascertaining that the data provision is valid and current. There was initial confusion regarding listing information and transaction data, caused by among other things, time delays between the two. In a rapidly growing region there could be significant differences between the two. The former although current, covers only asking prices. The latter is considered better due to it being closer to true market values, but there is a delay in getting information.

The second stage of an AVM begins with a data cleansing process. Due to large volumes of data, the process has to be automated. Data cleaning is followed by the identification and selection of variables that are meaningful for the residential property market in China. Hence, a complete understanding of the different regions of China and the variables and their detailed definitions is important. Most studies select variables based on statistical significance levels but in this research, for the pre-processing work, the inputs from expert knowledge from JinZheng and CIREA were used. This is because this study wants to avoid a brute force data driven approach and prefers to apply some intelligent pre-processing heuristic rules to incorporate relevant knowledge before ANNs are applied to help identify the more complex relationships in the data.

In the model specification stage, there is a need for an appropriate market value function to be considered before the analysis. This study did some initial research into the design of semi-parametric and spatial models that may provide more flexibility, but they remained more difficult to compute. For example, a non-parametric function may manage to get the location value

with high flexibility but it will be time consuming for the prediction values to be generated due to its reliance on interpolation.

Conversely market value prediction with an estimated parametric model is more direct, because the functional form itself supports interpolation. Once a suitable model (or a set of suitable models) has been established, there is an out-of sample validation process. This corresponds to a dry-run of the AVM before the service execution. The validation step also helps to differentiate and test between suitable vs unsuitable models.

The third stage of an AVM is on the implementation of the service from a technical standpoint. The real life appraisals are undertaken in real-time on desktop personal computers and an effective technical implementation strategy is important for commercial application from this research. It should be noted that in this research, the data considers the street address of residential properties and the property types, and also city location information. However, it does not consider the condition or the state of repair of the property.

The fourth stage consists of back-testing the AVM for validation. This is achieved by holding some samples for testing purposes. Back-testing requires the appraisal errors to be detected and used for realigning the model. For example, if appraisal errors in one local area has the tendency to be negative, then correction adjustments need to be identified to flexibly fit an alternative model for this area.

## 4.3 Data Collection and Sample Selection

In the last section, the research design of the AVM method was outlined. This section explains the in-depth process of the data collection method and sample selection strategy, including the variables and the descriptive statistics.

### 4.3.1 Data Collection Methods

To develop the proposed model of the research, the data had been collected form multiple sucrose for a Chinese real estate market. The raw data are collected from four main sources in the industry:

1. Real Estate Developing Company
2. Real Estate Agency
3. Real Estate Appraisal Company
4. Real Estate Management Company

The database was formed as a result of the collaboration between 2 partners, JinZheng real estate company and Cityre. This particular database became the biggest Real Estate Database in China.JinZheng real estate appraisal company is a top level real estate appraisal Company in China, which cover all regions in China. Also, Cityre is a leading company of national real estate property information and data-service provider in China. The company is the first countrywide endeavor specializing in real estate data collection, processing, analyzing and application. Throughout the years of effort between the two partners, one of the largest national real estate market databases had been created which naturally has evolved into a real estate Data Ecosystem.

Also with scale of over 300 million pieces of supply information of 283 cities. It is currently one of the biggest database of real estate in China.The sample data set was selected from this large database. For ethical consideration, companies who supplied the data for this research, did not influence the research and the purpose of the research, also had no intention to manipulate the result of the research.

.

### 4.3.2 Sample Selection Strategy

According to the Urban Planing Law in 1989, vast number of chinese cities has been classfied into four tiers, Super-large Cities (Tier 1): Cities with population over 1,000,000; Large Cities (Tier 2): Cities with population between 500,000 and 1,000,000; Medium-sized Cities (Tier 3): Cities with population between 200,000 and 500,000; Small Cities (Tier 4): Cities with population less than 200,000.(Wei et al., 2016) Due to the rapidly growth of population, plus the increase of internal migration, this classification no longer represent the current situation of Cities.(Cao, 2015) In 2014, a new system has been introduced by Chinese Academy of Social Sciences to divide cities into 5 categories, Mega-large Cities (Tier 1): Cities with population over 10,000,000; Super-large Cities (Tier 2): Cities with population between 5,000,000 and 10,000,000; Large Cities (Tier 3): Cities with population between 1,000,000 and 5,000,000; Medium-sized Cities (Tier 4): Cities with population between 500,000 and 1,000,000; Small Cities (Tier 4): Cities with population less than 500,000. (Cao, 2015; Wei et al., 2016)

This study has adopted the new city tier system and selected 6 cities

from the 283 cities in the database such that it remains representative of the economic income levels and residential property types in China. Beijing and Shanghai from City tier 1 were selected since Beijing is the Capital of China and Shanghai is the economic centre for China. Jinan and Qingdao were selected from city tier 2, which Jinan is the capital of Shandong Province, and Qingdao is a strong economy city in Shandong Province. Hohhot and Baotou from City tier 3 were selected, which Hohhot is the capital city of Inner Mongolia, an autonomous region of northern China.

**Data**

As China categorise their cities into tier types for the purposes of governance and investment strategy. This is a very strong economic factor in China. This study selected six cities from the 283 cities in the database such that the sample remained representative of the economic income levels and residential property types in China.

- Beijing and Shanghai are City tier 1 examples and were selected since Beijing is the Capital of China, and Shanghai is a major economic center for China. The Tier 1 includes 127,441 raw sales records (sale tag price) for Beijing, 243,222 raw records for Shanghai

- Jinan and Qingdao were selected from city tier 2, which Jinan is the capital of Shandong Province, and Qingdao is a major economy city in Shandong Province. Tier 2 data includes 47,124 raw records for Jinan and 67,611 for Qingdao

- Hohhot and Baotou from City tier 3 were selected, where Hohhot is the capital city of Inner Mongolia Province, and Baotou a major economic

center in the region. They are also examples from an autonomous region of northern China. Tier 3 data includes 5,609 raw records for Hohhot, and 9,614 raw records for Baotou.

The details of the dataset are shown in Table 4.1, Table 4.2, Table 4.3 and Table 4.4.

| Tier | Tier 1 | | Total |
|---|---|---|---|
| City | Beijing | Shanghai | |
| No. of District | 18 | 18 | 32 |
| No. of Residence | 1,241 | 2,581 | 3822 |
| No. of Raw Data | 278,371 | 580,211 | 858,582 |
| No. of Processed Data | 127,441 | 243,222 | 370,663 |

Table 4.1: The Data set for Tier 1 Cities

For the tier 1 cities as shown in Table 4.1, Beijing have 278,371 records for 18 different district, which includes 1,241 residence community areas. The raw data includes missing data and wrong value. After pre-processing stage, there are 127,441 valuable data left for the data set. Similarly, for the other tier 1 city, Shanghai has 18 districts and so is Beijing. 580,211 records twice as Beijing and 2,581 residence community areas, also as twice as Beijing. In total, Tier 1 cities have 32 different districts and 3,822 residential communities containing 858,582 data records for all of Tier 1 cities.

| Tier | Tier 2 | | Total |
|---|---|---|---|
| City | Jinan | Qingdao | |
| No. of District | 9 | 12 | 21 |
| No. of Residence | 443 | 580 | 1023 |
| No. of Raw Data | 104,149 | 133,988 | 238,137 |
| No. of Processed Data | 47,124 | 67,611 | 114735 |

Table 4.2: The Data set for Tier 2 Cities

For Tier 2 cities as shown in Table 4.2, Jinan is the capital city of the Shandong Province, which has 104,149 raw data for sales records for 443

residence communities within 9 districts. Qingdao has 3 more districts in the data set, more community areas and more raw data on properties in the data set (580 residence and 133,988 raw data).

| Tier | Tier 3 | | Total |
|---|---|---|---|
| City | Hohhot | Baotou | |
| No. of District | 5 | 4 | 9 |
| No. of Residence | 84 | 110 | 194 |
| No. of Raw Data | 19,334 | 25,407 | 44,741 |
| No. of Processed Data | 5,609 | 9,614 | 15,223 |

Table 4.3: The Data set for Tier 3 Cities

Table4.3 shows the data set for Tier 3 Cities. both of Hohhot and Baotou have less districts than Tier 1 and Tier 2 cities, forming 5 districts and 4 districts, respectively. Hohhot records 84 residential communities and Baotou records 110, the raw data volume of Hohhot is 19,334 which is lower than the City of Baotou (25,407).

| Tier | Tier 1 | Tier 2 | Tier 3 | Total |
|---|---|---|---|---|
| No. of District | 32 | 21 | 9 | 66 |
| No. of Residence | 3822 | 1023 | 194 | 4,997 |
| No. of Raw Data | 858,582 | 238,137 | 44,741 | 1,141,460 |
| No. of Processed Data | 370,663 | 114,735 | 15,223 | 500,621 |

Table 4.4: The Data set for City Tiers

As table above shows that Tier 1 Cities have significant more data compare to Tier 2 and Tier 3 cities, also Tier 2 has more data than Tier 3 cities. total add together as 1,141,460 sales records as raw data in the data set.

The dataset used for this study is larger, more extensive and probably more reliable than the datasets used in most previous studies on property valuation. Brown and D'amato's work only use 725 and 390 dwellings receptively

Brown and Uyar (2004); D 'amato (2007). The recent works from Garca and Arribas only use 591 and 2,149 records Cervelló et al. (2011); Arribas et al. (2016). Each of the six cities in the dataset for this research have more data compared to all of the previous research identified.

### Variables and Measures

Each residential real estate record in the dataset consisting of 15 variables were collected for each subject property: three describe the location, 10 describe the characteristics, two describe the building in which it is sited. The snapshot of a sample dataframe are provided in Figure 4.1. The detailed descriptions of the variables are shown in Figure 4.2:

| | city | district | residence | total_price | price_per_sqm | date | area | bedroom | livingroom | kitchen | bathroom | structure | year | decoration | building_type | floor | direction |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 青岛 | 平度市 | 昌泰花苑 | 70.0 | 7000.00 | 2017-02-04 16:32:14 | 100.00 | 3.0 | 2.0 | 1.0 | 1.0 | 带车库 | 2013.0 | 精装修 | NaN | 4.0 | 南向 |
| **1** | 青岛 | 平度市 | 昌泰花苑 | 53.0 | 6385.54 | 2017-02-21 00:00:00 | 83.00 | 2.0 | 2.0 | 1.0 | 1.0 | 带阁楼 | 2005.0 | 精装修 | 多层 | 3.0 | 南向 |
| **2** | 青岛 | 平度市 | 昌泰花苑 | 98.0 | 7656.25 | 2017-03-17 00:00:00 | 128.00 | 3.0 | 2.0 | NaN | 2.0 | 带阁楼 | 2004.0 | 精装修 | 多层 | NaN | 南向 |
| **3** | 青岛 | 平度市 | 昌泰花苑 | 70.0 | 4895.10 | 2017-03-12 00:00:00 | 142.90 | 3.0 | NaN | NaN | 2.0 | 带阁楼 | 2006.0 | 其他 | NaN | NaN | 南向 |
| **4** | 青岛 | 平度市 | 昌泰花苑 | 65.0 | 5909.09 | 2017-02-16 09:06:40 | 110.00 | 3.0 | 1.0 | NaN | 1.0 | NaN | 2013.0 | 简单装修 | 多层 | 5.0 | 南向 |
| **5** | 青岛 | 平度市 | 昌泰花苑 | 98.0 | 7656.25 | 2017-03-23 15:13:22 | 128.00 | 3.0 | 2.0 | 1.0 | 1.0 | 带车库 | 2005.0 | 精装修 | 多层 | 4.0 | 南向 |
| **6** | 青岛 | 平度市 | 昌泰花苑 | 82.0 | 6456.69 | 2017-01-19 00:00:00 | 127.00 | 3.0 | 2.0 | NaN | 1.0 | 带阁楼 | NaN | 精装修 | 多层 | 5.0 | 南北 |
| **7** | 青岛 | 平度市 | 昌泰花苑 | 62.0 | 4960.00 | 2017-02-19 11:33:00 | 125.00 | 3.0 | 2.0 | NaN | 1.0 | 带车库 | NaN | 未装修 | NaN | NaN | 南向 |
| **8** | 青岛 | 平度市 | 昌泰花苑 | 65.0 | 6770.83 | 2017-02-12 00:00:00 | 96.00 | 3.0 | 2.0 | 1.0 | 1.0 | 带阁楼 | 2003.0 | 精装修 | 多层 | 3.0 | 南向 |
| **9** | 青岛 | 平度市 | 昌泰花苑 | 66.0 | 6600.00 | 2017-01-21 00:00:00 | 99.30 | 3.0 | 2.0 | 0.0 | 1.0 | 带阁楼 | 2013.0 | 豪华装修 | NaN | 1.0 | 南向 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **70798** | 北京 | 朝阳区 | 清荷里南山园 | 152.0 | 23576.90 | 2017-03-07 15:10:00 | 64.47 | 1.0 | 1.0 | NaN | 1.0 | NaN | 2016.0 | 简单装修 | NaN | NaN | 南向 |
| **70799** | 北京 | 朝阳区 | 清荷里南山园 | 206.0 | 24819.30 | 2017-03-12 00:00:00 | 83.00 | 2.0 | 1.0 | 1.0 | 1.0 | NaN | 2016.0 | 毛坯房 | NaN | 9.0 | 南向 |
| **70800** | 北京 | 朝阳区 | 清荷里南山园 | 165.0 | 25000.00 | 2017-03-12 00:00:00 | 66.00 | 1.0 | 1.0 | 1.0 | 1.0 | NaN | 2016.0 | 毛坯房 | NaN | 10.0 | 南向 |
| **70801** | 北京 | 朝阳区 | 清荷里南山园 | 262.0 | 24952.40 | 2017-03-12 00:00:00 | 105.00 | 3.0 | 1.0 | 1.0 | 1.0 | NaN | 2014.0 | 毛坯房 | NaN | 6.0 | 南北 |
| **70802** | 北京 | 朝阳区 | 清荷里南山园 | 243.0 | 30000.00 | 2017-03-12 00:00:00 | 81.00 | 2.0 | 2.0 | 1.0 | 1.0 | NaN | 2017.0 | 毛坯房 | NaN | 16.0 | 南向 |
| **70803** | 北京 | 朝阳区 | 清荷里南山园 | 202.0 | 24938.30 | 2017-03-12 00:00:00 | 81.00 | 2.0 | 1.0 | 1.0 | 1.0 | NaN | 2016.0 | 简单装修 | NaN | 6.0 | 南北 |
| **70804** | 北京 | 朝阳区 | 清荷里南山园 | 205.0 | 24698.80 | 2017-03-12 00:00:00 | 83.00 | 2.0 | 1.0 | 1.0 | 1.0 | NaN | NaN | NaN | NaN | 17.0 | 南北 |
| **70805** | 北京 | 朝阳区 | 清荷里南山园 | 262.0 | 24952.40 | 2017-03-12 16:47:00 | 105.00 | 3.0 | 2.0 | 1.0 | NaN | NaN | 2016.0 | 毛坯房 | NaN | 18.0 | 南北 |
| **70806** | 北京 | 朝阳区 | 清荷里南山园 | 336.0 | 32000.00 | 2017-03-14 13:57:41 | 105.00 | 3.0 | 1.0 | 1.0 | 1.0 | NaN | 2018.0 | 简单装修 | NaN | 15.0 | 南北 |
| **70807** | 北京 | 朝阳区 | 清荷里南山园 | 220.0 | 25000.00 | 2017-03-14 17:39:33 | 88.00 | 2.0 | 1.0 | 1.0 | 1.0 | NaN | 2016.0 | 毛坯房 | NaN | 2.0 | 南北 |

Figure 4.1: Snapshot of a sample data frame

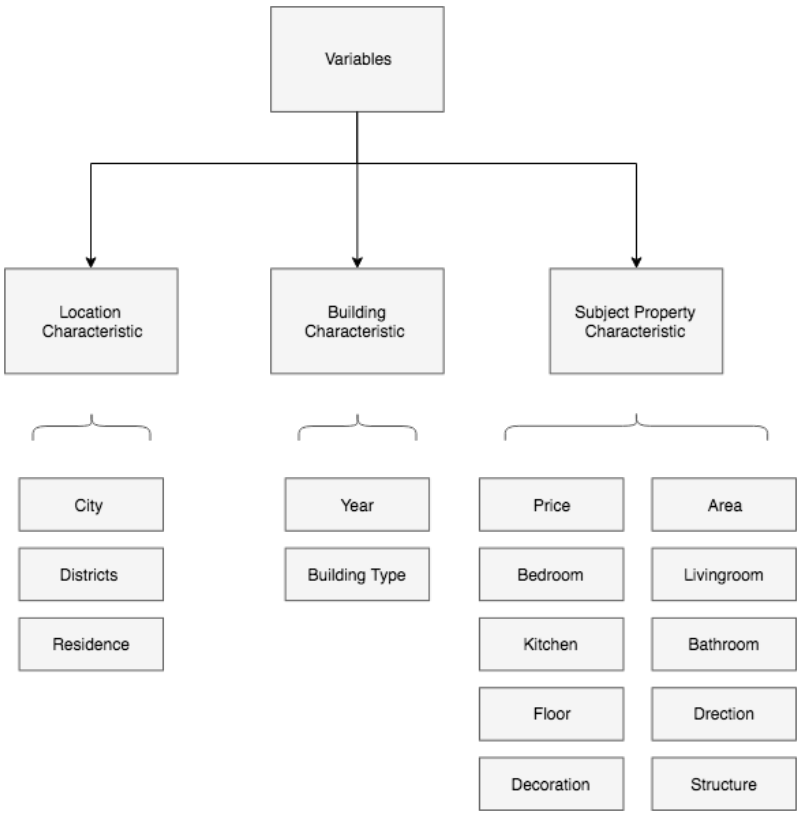


Figure 4.2: Variables Selection

- **Location Characteristics**

  Location Characteristics for the data set, includes city name, district , residential community areas.

  1. City: Six cities (Beijing, Shanghai, Jinan, Qingdao, Hohhot, Baotou).

     City as a variable in the data base, has the Feature of different the dataset from the first layer of location. This study will train the ANN for each city, therefore, the variable city not taking into account for training process.

2. District: District the subject property is located

   District is the geo-location for each property located within the city, different district could contain different infrastructure or transportation system and so on, which influence the price of the subject property. Therefore, the district variable will be include into the ANN training process.

3. Residence: Name of the residential community area

   Residence is the name of the residential community area, which is important for training. because of different community area contains different environment, convenient level and level of service inside the area. all of thees factors significantly affect the property price thus play an important factor for ANN training.

- **Building Characteristics**

  Building Characteristics will includes year and building type. this characteristics only describe the building of the subject property, which the building condition will have affect on the price of particular property in the building.

  1. Year: Year the building is completed

     Year as variable is the year of the building completion. This demonstrate the building as belonging to a particular period in terms of construction technology. This also defines value from the time left from the 70 year Lease Hold in the Chinese property market.

     - Mean 2003.95, Median 2005, Min 1900, Max 2019, Std 7.90

2. Building_type: Building types of properties

   There are 8 different building types, includes, Bungalow, High-rise, High-level, Multi-storey, Entire Block, Semi-detached House, Detached House, Siheyuan. Different building type may affect property price by many form. For example, density of population of the community area for Detached house most likely have less population density than High-level community area.

- **Subject Property Characteristics**

  Subject Property Characteristics are focused on feature of the property itself. there are 10 different variable in this category, includes, price of the property, area size, number of bedroom, number of livingroom, number of kitchen, number of bathroom, floor level, property structure, decoration level and facing of the property(direction).

  1. Price: Price per square meter

     Sale price is the most important variable in any property valuation models, because of the taxation is calculated from the selling price, to prevent the truth of the data, tag sale price had been adopted, because it represent the market expectation of the property selling price.

     - Mean 46,878, Median 44,666, Min 2,075, Max 288,690, Std 28,810

2. Area: Total area of the Subject Property

   The area variable is the total property floor area measured in square meters. the floor area in china includes partly public area, in some situation that floor area is not the area for usage, the variable include in the model training is the floor area in usage only.

   - Mean 99.89, Median 88, Min 10, Max 2900, Std 59.87

3. Bedroom: Number of bedrooms

   Number of Bedroom played an important role in residential property valuation. In genral, more bedrooms will increase the property price relative to property have fewer bedroom within similar community area.

   - Mean 2.29, Median 2, Min 0, Max 9, Std 1.03

4. Livingroom: Number of livingrooms

   Number of Living room have similar effect on price as the number of bedrooms, more living room generally will increase the property price correspondingly.

   - Mean 1.56, Median 2, Min 0, Max 7, Std 0.56

5. Kitchen: Number of kitchen

   Number of Kitchen also have positive relationship with property price in general, normally, each property only require one kitchen, however there are property type like courtyard(siheyuan) require more than one.

   - Mean 0.97, Median 1, Min 0, Max 5, Std 0.19

6. Bathroom: Number of bathroom

   Research from Garrod and Willis (1992) and Guntermann and Norrbin (1987) shows that extra bathroom corresponded influence on price compared with similar condition property in same community area. The number of bathrooms also have positive relationship with sale price.

   - Mean 1.32, Median 1, Min 0, Max 9, Std 0.64

7. Floor: Floor number of the Subject Property

   Floor number of the property is a very important variable for Chinese real estate market. because of the majority of the residential areas are building types of High-rise, High-level, Multi-storey, which makes the floor number of property is very important for different properties within same block.

   - Mean 5.36, Median 4, Min -10, Max 63, Std 4.94

8. Structure: Subject Property Structure.

   Inside of the property, can be different to flat structure, with atic, with basement and Duplex, it could affect the price.

9. Decoration: The level of the decoration. Total 6 Categories (None, Partial, Simple, Mid-range, Deluxe, Luxury)

   Level of decoration has been divided into 6 categories, (None, Partial, Simple, Mid-range, Deluxe, Luxury). better the decoration level in second hand property market will address higher tage sale price.

10. Direction: Direction of the property, 10 in total (North, South, East, West, NorthEast, NorthWest, SouthEast, SouthWest, North-South, EastWest)

    Direction variable represent the facing direction of the property, which will affect the sun coverage over the daytime. this will affect the price. for example, facing south is the default good property facing in Chinese real estate market.

### 4.3.3 Pre-processing Data

The database is in the Chinese language. In order to use Pytorch as a tool for programming, all the data has been translated from Chinese to English. This was to reduce the possibility of language differences affecting the model performance. The pre-processing of the data also removed outliers and dealt with missing data for every single variable.

## 4.4 Conclusion

This chapter has addressed the research design for this study including defining the building blocks for the framework of a mass-appraisal system (and largely the automated valuation model). The characteristics of the dataset have been explained for the automated valuation model (AVM), which highlights the suitability of addressing this in the Chinese context, given that rapid development necessitates high volume property appraisals.

# Chapter 5

# Technical Development

## 5.1 Introduction

This chapter describes and discusses the design of the Homogeneous Feature Transfer and Heterogeneous Location Fine-tuning (HTF+HLF) cross-city property appraisal framework. Through transferring partial neural network learning from a source city and fine-tuning, this study develops show a method of that less amount of location information for a target city to arrive at a better performance for a mass appraisal system.

There are two main significant contributions through the HTF+HLF design compared to the other models listed in the literature.

- This research is the first to focus on the possible value of cross-city transfer problem in property appraisal. Because the datasets collected from different cities contain entirely different sets of location features (District names, Subject Property Community Area names, etc.), most of the existing transfer learning methods cannot be directly applied to cross-city property appraisal models due to these heterogeneous location

features. Hence, the cross-city property appraisal model transfer is a new approach, the application of which is addressed in this research

- This research proposed a novel **H**omogeneous **F**eature **T**ransfer and **H**eterogeneous **L**ocation **F**ine-tuning **(HFT+HLF)** framework. By transferring part of our neural network and semi-supervised fine-tuning the remaining part, our model can surpass the fully supervised single-city ANN model by using only 20% to 30% of the available training data.

## 5.2 HTF+HLF Design

### 5.2.1 Homogeneous and Heterogeneous Features

The real estate datasets applied contains three feature categories: Location Characteristics, Building Characteristics and Subject Property Characteristics. The location characteristics tell where the buildings are located. The building characteristics describe the condition of the entire building and its neighborhood area. The Subject Property Characteristics are the specific property's internal construction attributes such as the number of bedroom or living room. The variables of building characteristics and subject property characteristics are usually homogeneous for different cities. For example, the decoration level of two subject property of city A and city B can both be labelled from the same set of categorical variables: None, Partial, Simple, Mid-range, Deluxe, or Luxury. However, different cities have different sets of location information: different district names, different residential community names, etc. Hence, the location variables are heterogeneous for different cities.

The heterogeneous location features and homogeneous property features are demonstrated in Figure 5.1

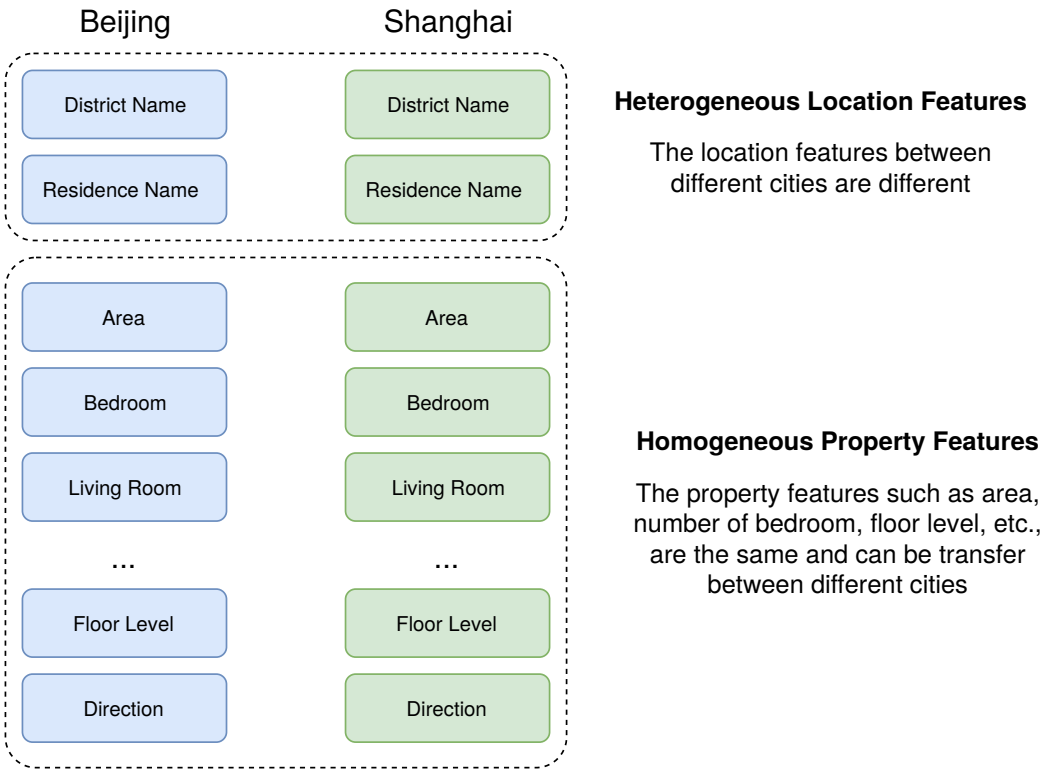


Figure 5.1: Heterogeneous Location Feature and Homogeneous Property Feature

As in figure 5.1, the cross-city estate dataset can be divided into two categories: homogeneous features and heterogeneous features. The homogeneous features usually describe the characteristics of the property such as direction, floor level, area. These features are usually city invariant. However, the location information like district names and subject property community names are city-dependent.

### 5.2.2 HFT+HLF Framework

Most of the ANN-based property appraisal methods have to replace and retrain the regression factors for each new city due to the heterogeneous location features. This model adopted addresses this problem by separating the homogeneous and heterogeneous features during the training. The homogeneous features are used to create a cross-city transferable model and the city-dependant heterogeneous features are used for location-based fine-tuning. As a result, the final network consists of two joined ANNs: a homogeneous feature transferable ANN and heterogeneous location fine-tuning ANN, as shown in Figure 5.2.

The homogeneous features such as the subjective property features and building features are the inputs for the transferable part. The output feature maps from the transferable ANNs are then concatenated with the heterogeneous location features and become the new inputs to the fine-tuning section of the proposed network. Since the homogeneous features are commonly shared between properties of different cities, this part of our neural network can be considered as a generic subject property relationships (apartment is majority type of real estate transaction in the chinese real estate market) and building features learning network which could be *transferred* between different cities. Through transfering this part of the network to a new city, the new city model only need to optimise the weight of the fine-tuning network. For example, the weights of the transferable ANN learned from the source city Beijing can be transferred to the new model of the target city Baotou and then fine-tuned with less data from the target city Baotou. For layer design, the researchers have tested for 1 hidden layer ANN(Do and Grudnitski, 1992; Yeh et al., 2018) and a 2-hidden-layers ANN (Khalafallah, 2008).

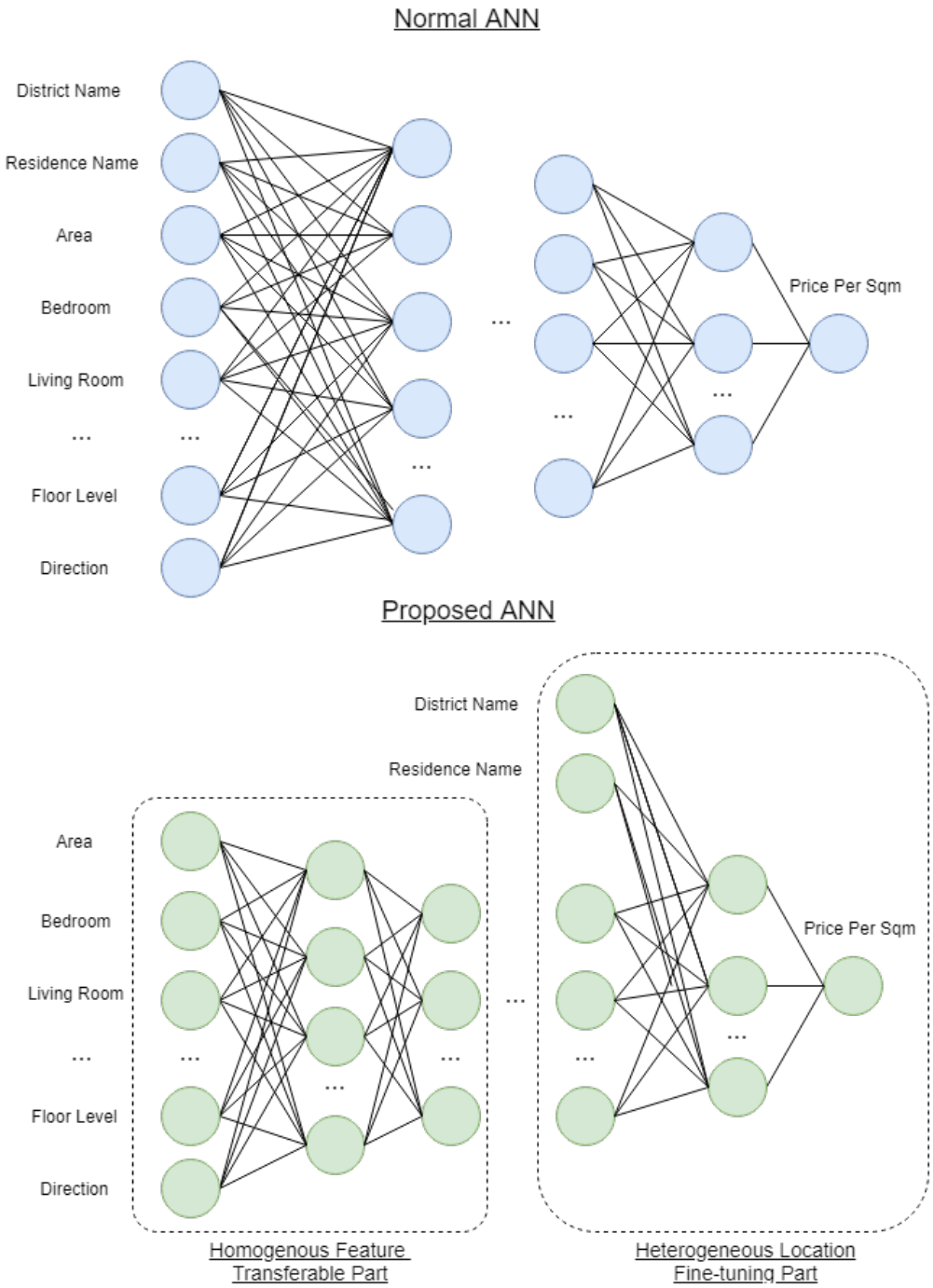


Figure 5.2: Normal ANN vs Proposed ANN

| City | Beijing | | | Shanghai | | |
|---|---|---|---|---|---|---|
| # Hidden Layers | RMSE | MAPE | R2 | RMSE | MAPE | R2 |
| 1 | 0.295 | 0.171 | 0.751 | 0.148 | 0.176 | 0.742 |
| 2 | 0.265 | 0.160 | 0.781 | 0.178 | 0.156 | 0.762 |
| 3 | 0.245 | 0.157 | 0.793 | 0.182 | 0.144 | 0.789 |
| 4 | **0.239** | **0.149** | **0.816** | **0.186** | **0.138** | 0.782 |
| 5 | 0.243 | 0.151 | 0.787 | 0.173 | 0.146 | **0.791** |

Table 5.1: Performance Score for Layer Selection

In this proposed method, extensive experiments have been conducted as shown in Table 5.1 to determine the number of layers needed for this proposed ANN structure. For convenient evaluation, the number of hidden layers of the transferable ANN and the fine-tuning ANN are set to be the same. From Table 5.1, a comparison between Beijing and Shanghai demonstrate that 4 hidden layers would given a better performance score for layer selection. Therefore, in these experiments, 4-hidden-layer structure yields the best performance for modeling the Beijing and Shanghai dataset in the fully supervised setting. The final proposed ANN consists of total 9 hidden layer: 4 hidden layers with $(100, 50, 20, 10)$ hidden nodes for the transferable ANN, 1 concatenation layer and 4 hidden layers with $(100, 50, 20, 10)$ hidden nodes for the fine-tuning ANN. After a 10 fold cross validation, all the hidden layers were equipped with 0.1 negative slope Leaky ReLU activation function, a 0.5 drop-out rate and Batch Normalization. This research use mean squared error (MSE) as the criterion loss for the price regression. The network can be optimised using the Adam algorithm with AMSGrad moving average variant. The overview of our architecture is shown in the lower section of Fig 5.2 with the standard traditional ANN structure added on top for comparison.

In the Figure 5.2, the top figure is the traditional ANN structure for property appraisal. The bottom figure is the proposed structure tailored for

cross-city transfer learning. The transferable part can be transferred to the model created for a different city. Only the parameters in the fine-tune part need to be learned based on the data from a new city

## 5.3 HTF+HLF Neural Network Design

The design of the neural network follows a modified version of the Kaastra and Boyd (1996)'s method to build the proposed ANN. The steps are discussed below and shown in Figure 5.3

### 5.3.1 Variable selection

There are 15 variables in the dataset have been selected for the Mass Appraisal System. This study divided the variables into three segments depending on their characteristics. (see Section 4.3.2)

### 5.3.2 Data Pre-processing

The data pre-processing mainly deals with missing data and outlier values from the sample dataset form this research. This process can help the neural network models to learn the pattern within the data more precisely by helping reduce incorrect trends and spurious fit before training.

### 5.3.3 Number of Inputs

In Neural Networks, the optimal number of input variable (input neurons) is significant and important for its learning ability and output performance. More vector intake from variables, the better the coverage of the model, How-

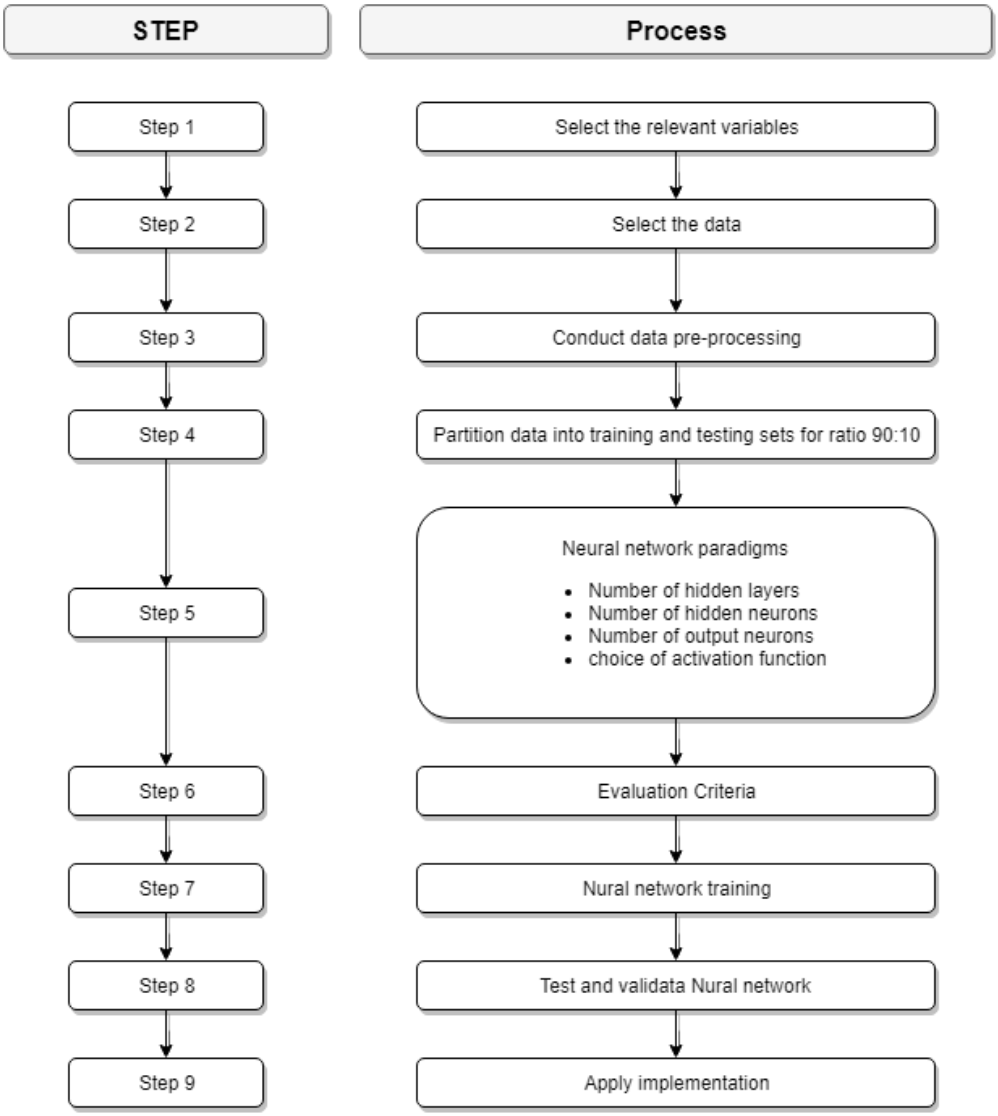


Figure 5.3: Design of Proposed Neural Network

ever, too many inputs adversely affect model capability and the efficiency of the network. In this research, all attributes supplied for the dataset which was initially provided by the national owned data company, Cityre, (claimed as one of the largest real estate data provider in China) were utilised. After pre-processing trials of the six cities' dataset , 14 input neurons and real estate price per square meter as the output neuron were selected.

### 5.3.4 Loss Function

In this study, the goal of the ANN was to produce the price per square meter as the output variable, after some trials the MSE loss function method of regression loss function was selected. The MSE is the sum of squared distances between the target variable and predicted values during the training process.

The loss can be described as:

$$\ell(x, y) = L = \{l_1, \ldots, l_N\}^\top, \quad l_n = (x_n - y_n)^2,$$

## 5.4 Confidence in HTF+HLF

A level of confidence in any method is needed, whether automated or manual.However, the data volume and data quality significantly affect the performance of ANNs.

| Measure | Formula | Minmum Value | Maxmum Value | Desirable Outcome |
|---|---|---|---|---|
| *RMSE* | $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$ | 0 | 1 | Min |
| *MAPE* | $\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert\hat{y}_i - y_i\rvert}{\lvert y_i\rvert}$ | 0 | 1 | Min |
| $R^2$ | $1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$ | 0 | 1 | close to 1 |

Table 5.2: Performance Measure

A number of performance measures were used to evaluate confidence in different ways. They were the root-mean-squared-error (RMSE), mean-absolute-percentage-error (MAPE) and R-squared Error ($R^2$). In order to compare the measures, the setting for training and validation were the same for all experiments. RMSE as an accuracy measures compares the sample standard deviation of the difference between observed and predicted values. MAE also gives the average of squared errors in the aggregate size of the errors in the model's predictions. $R^2$ indicates how well the models fit the training and validation data sets, which is the most common Key Performance Indicator (KPI) for model performance (Mas et al., 2016; Kok et al., 2017).

This Chapter has discussed the structure of a systematic process for generating reliable processes and results. The next chapter describes the detailed experiment design and the results produced.

# Chapter 6

# Experimental Design and Results

## 6.1 Introduction

This chapter presents the implementation of the automated aspect of the mass appraisal system. Specifically, it demonstrates the transferring of partial neural network learning from a source city and the fine-tuning of a smaller amount of location information for a target city, This new proposed semi-supervised models actual performance is then compared to that of a fully supervised Artificial Neural Network (ANN) method.

## 6.2 Data Labels

Properties in six cities covering first three Tiers of Chinese City Classification were selected for this research work. In addition to the proposed location transfer appraisal method, the work undertaken created a massive cross-city

dataset also achieved through this research. The labels for the data are shown in Table 6.1

| ***Label*** | ***Definition*** | ***Data Type*** | ***Measures*** |
|---|---|---|---|
| city | Name of the city | Categorical | [-] |
| districts | Distric of the property located in the subject city | Categorical | [-] |
| residence | The name of community area | Categorical | [-] |
| year | The year of building completion | Numerical | [Year] |
| building type | The building type from usage purpose | Categorical | [-] |
| bedroom | Number of bedroom | Numerical | [Units] |
| kitchen | Number of kitchen | Numerical | [Units] |
| floor | Level of floor | Numerical | [Level] |
| decoration | Degree of decoration level | Categorical | [-] |
| area | Usable floor area in square meter | Numerical | $[m^2]$ |
| livingroom | Number of livingroom | Numerical | [Units] |
| bathroom | Number of bathroom | Numerical | [Units] |
| drection | the facing of the property | Categorical | [-] |
| structure | internal structure of the property | Categorical | [-] |
| price | the price of the property | Numerical | [] |

Table 6.1: Labels for the data

## 6.3 Training and Settings

Cities of different Tiers normally have different (as shown in the core data discussed in Chapter 4) sizes of data available for the training samples. As shown in Table 4.4, the available data for first-tier cities (Beijing, Shanghai) is nearly 20 times more than for the third-tier cities (Hohhot, Baotou). As a result, the batch size and the learning rate have to be individually set for each city tier.

The pre-experiment for testing the effectiveness of different learning rate was run for six different data sets, the learning rate was set in a range 0.001 to 0.03 and batch size 64 to 256, then the best result of learning rate selected for proposed location transfer ANN model.

In the experiments undertaken, after a 10 fold cross validation process, the learning rate was set at 0.005 with batch size of 256 for Tier 1 cities, a 0.01 learning rate with batch size of 128 for Tier 2 cities and 0.02 learning rate with batch size of 64 for Tier 3 cities. The number of epochs was set 300 to ensure the network is fully converged during the training. The test network was implemented in PyTorch and the training times ranged from 2 hours to 30 minutes depending on the training sample size. The performance metrics used were Root Mean Square Error (RMSE), R-squared Error ($R^2$) and Mean Absolute Percentage Error (MAPE) as shown in Table5.2 in Section 5.4 .

## 6.4 Experiment 1: Traditional ANN vs Proposed ANN

Since the proposed network is a new ANN structure, the first experiment is to check whether the new structure affects the appraisal performance of the ANN model. In this experiment, this research compares the performance of the proposed transferable model with a normal non-transferable ANN model in a fully supervised single city setting. This experiment follows the common 10-fold cross-validation strategy which is the average of the experimental results based on 10 random splits of the dataset into 90% training samples and 10% testing samples. The detailed performance metrics are shown in Table 6.2

below.

| Model | Performance Metric | Tier 1 | | Tier 2 | | Tier 3 | |
|---|---|---|---|---|---|---|---|
| | | Beijing | Shanghai | Jinan | Qingdao | Hohhot | Baoto |
| Traditional ANN | RMSE (Lower Better) | 0.246 | 0.186 | **0.184** | **0.202** | **0.150** | 0.145 |
| | MAPE (Lower Better) | 0.151 | 0.138 | **0.142** | 0.157 | **0.121** | 0.108 |
| | $R^2$ (Higher Better) | 0.804 | **0.782** | **0.757** | **0.828** | 0.522 | **0.485** |
| Proposed ANN | RMSE (Lower Better) | **0.239** | **0.185** | 0.189 | 0.204 | 0.170 | **0.135** |
| | MAPE (Lower Better) | **0.149** | **0.136** | 0.147 | 0.159 | 0.129 | **0.102** |
| | $R^2$ (Higher Better) | **0.816** | 0.778 | 0.744 | **0.828** | **0.523** | 0.473 |

Table 6.2: Performance Comparison between Traditional ANN and Proposed Transferable ANN

The proposed model achieved similar and impressively low RMSE and MAPE scores for all six cities. This means that the model can converge well on all of the training datasets used. However, the low $R^2$ scores for the Tier 3 city indicate the poor regression performance. This is mainly due to the lack of good available data in the Tier 3 cities. This validates the authors claim that the cross-city transfer learning is necessary for property appraisal model, especially to those low volume Tier 3 cities. Comparing the proposed models with standard non-transferable ANN method, the proposed network yields a very similar overall performance. It proves that the new proposed architecture did not affect the overall appraisal performance.

## 6.5 Experiment 2: Semi-supervised Cross City Transfer Learning

Most other ANN appraisal methods have to train a new model from scratch for different cities or regions. This research proposed model from this research has the ability to transfer a partially pre-trained network learned from the source city to the target city. Experiment 2 was conducted to validate the performance of proposed model in a cross-city semi-supervised setting.

### 6.5.1 Transfer to Tier 1 Cities:

The first set of experiments is to test the proposed models ability to transfer the pre-trained network between the first tier cities: Beijing and Shanghai. The situation is that China, Tier 1 cities always face the first, new policy changes and the pressure for digitisation, also the real estate market is more active compared to the other Tiers of cities. Therefore, the Tier 1 cities usually have a larger amount of data available for training. The demand for transferring appraisal models conducted from lower Tier 2 or 3 cities to Tier 1 cities is highly unlikely in the real-world practice. As a result, this research conducted the Beijing to Shanghai and Shanghai to Beijing Transfer Learning in this experiment. For Shanghai to Beijing transfer, we first pre-train the model based on the training dataset of Shanghai and transfer the transferable part of the network to the new model for Beijing. Then, 10, 20, 30 records were selected from each location (each residential community) in Beijing to fine tune the Beijing model. The test dataset was randomly selected from the remaining dataset with the size of 10% of the overall Beijing dataset. The processes for

Beijing to Shanghai transfer are the same. As only a small amount of data from the Beijing training dataset was used, it can be considered as semi-supervised learning. The detailed performance metrics are shown in Table 6.3 and Table 6.4.

| Semi-Supervised | Training Size | **Shanghai ->Beijing** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 10 Records Per Residence | 12,020 | **0.186** | **0.139** | **0.829** |
| 20 Records Per Residence | 24,040 | **0.180** | **0.133** | **0.847** |
| 30 Records Per Residence | 36,060 | **0.174** | **0.129** | **0.845** |
| Fully Supervised (No Transfer Learning) | 127,441 | 0.239 | 0.149 | 0.816 |

Table 6.3: Performance Scores for Shanghai to Beijing Transfer.

| Semi-Supervised | Training Size | **Beijing ->Shanghai** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 10 Records Per Residence | 25,630 | 0.194 | 0.141 | 0.764 |
| 20 Records Per Residence | 50,860 | **0.185** | **0.135** | 0.776 |
| 30 Records Per Residence | 76,290 | **0.178** | **0.131** | **0.788** |
| Fully Supervised (No Transfer) | 243,222 | 0.185 | 0.136 | 0.778 |

Table 6.4: Performance Scores for Beijing to Shanghai Transfer.

By only using 20 records from each residential community, this research shows that a semi-supervised transferable model can quickly achieve the similar or even superior performance as compared with a fully supervised single city training set. If the training dataset is increased to 30 records per residential community, the overall performance even surpasses the single city supervised learning performance. If the training set is 20 records from each residential community, the overall size of the training data required only one-fifth of that for a fully supervised training dataset. As a consequence, the proposed transferable model can be significantly reduced data volume requirement by five times compared to traditional ANN modelling methods.

### 6.5.2 Transfer to Tier 2 Cities

Table 6.5, Table 6.6, Table 6.7 and Table 6.8 shows the experimental results for Tier 1 Cities to Tier2 Cities Transfer and Table 6.9 and 6.10 show those for within Tier 2 Cities Transfer. The Tier 1 to Tier 2 transfer models usually outperform the within Tier 2 transfer model. Because the Tier 1 cities have much more training data than Tier 2 and 3 cities, the transferable part of our network learned from Tier 1 cities usually create better predictive ability and robustness.

| Semi-Supervised | Training Size | **Beijing ->Jinan** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 10 Records Per Residence | 4,400 | 0.205 | 0.189 | 0.558 |
| 20 Records Per Residence | 8,800 | 0.193 | 0.165 | 0.642 |
| 30 Records Per Residence | 13,200 | **0.179** | **0.142** | **0.784** |
| Fully Supervised (No Transfer) | 47,124 | 0.189 | 0.147 | 0.744 |

Table 6.5: Performance Scores for Beijing to Jinan

| Semi-Supervised | Training Size | **Shanghai ->Jinan** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 10 Records Per Residence | 4,400 | 0.202 | 0.188 | 0.546 |
| 20 Records Per Residence | 8,800 | 0.198 | 0.173 | 0.612 |
| 30 Records Per Residence | 13,200 | **0.181** | **0.144** | **0.773** |
| Fully Supervised (No Transfer) | 47,124 | 0.189 | 0.147 | 0.744 |

Table 6.6: Performance Scores for Shanghai to Jinan

In table 6.5 and table 6.6, the experimental results for performance indicator of $RMSE$, $MAPE$ and $R^2$ for transfer from Beijing and Shanghai (Tier1 Cities) to Jinan (Tier2 city) are presented. Using 30 recods per residence, all three performance indicator are better compared with a fully supervised (No Transfer) result. This suggest that 30 record per residence will be enough

support for transfer from both Beijing and Shanghai to Jinan.

| Semi-Supervised | Training Size | **Beijing ->Qingdao** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 10 Records Per Residence | 5,790 | 0.302 | 0.204 | 0.624 |
| 20 Records Per Residence | 11,580 | 0.250 | 0.186 | 0.756 |
| 30 Records Per Residence | 17,370 | **0.203** | **0.155** | **0.836** |
| Fully Supervised (No Transfer) | 67,611 | 0.204 | 0.159 | 0.828 |

Table 6.7: Performance Scores for Beijing to Qingdao

| Semi-Supervised | Training Size | **Shanghai ->Qingdao** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 10 Records Per Residence | 5,790 | 0.296 | 0.218 | 0.636 |
| 20 Records Per Residence | 11,580 | 0.243 | 0.187 | 0.755 |
| 30 Records Per Residence | 17,370 | **0.201** | **0.153** | **0.838** |
| Fully Supervised (No Transfer) | 67,611 | 0.204 | 0.159 | 0.828 |

Table 6.8: Performance Scores for Shanghai to Qingdao

In Table 6.7 and Table 6.8, the result of transfer from Beijing and Shanghai to Qingdao, which shows similar performance results as the "transfer to Jinan". This means that transfer performance method will be better than the performance of a fully supervised ANN when 30 records per residence have been feed used. These also suggests that transfer from Tier 1 cities to Tier 2 cities, will also provide better performance models with less data required.

| Semi-Supervised | Training Size | **Qingdao ->Jinan** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 10 Records Per Residence | 4,400 | 0.290 | 0.212 | 0.495 |
| 20 Records Per Residence | 8,800 | 0.266 | 0.197 | 0.561 |
| 30 Records Per Residence | 13,200 | 0.241 | 0.175 | 0.614 |
| Fully Supervised (No Transfer) | 47,124 | 0.189 | 0.147 | **0.744** |

Table 6.9: Performance Scores for Qingdao to Jinan

| Semi-Supervised | Training Size | **Jinan ->Qingdao** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 10 Records Per Residence | 5,790 | 0.307 | 0.221 | 0.608 |
| 20 Records Per Residence | 11,580 | 0.281 | 0.217 | 0.659 |
| 30 Records Per Residence | 17,370 | 0.266 | 0.201 | 0.732 |
| Fully Supervised (No Transfer) | 67,611 | 0.204 | 0.159 | 0.828 |

Table 6.10: Performance Scores for Jinan to Qingdao

Table 6.9 and Table 6.10 shows that for internal transfer between Teir 2 cities, from the result, adding even 30 records does not increase the performance of transfer. which means inter transfer for teir2 cities will not produce a better performance model. This could because the lack of additional data voulume between Tier 2 cities. Therefore this study did not include evaluation for Tier 3 cities internal transfer, because of the consideration of data volume in Tier 3 cities is even lower than for Tier 2 cities.

### 6.5.3 Transfer to Tier 3 Cities

As the Tier 3 cities have a limited amount of training data and the supervised Tier 3 cities models have relatively poor performance (very low $R^2$ scores), also accoding to the reslut for internal transfer of Tier 2 city. The researcher did not test the Tier 3 to Tier 3 cities transfer evaluation. The focus was on evaluating the proposed network on Tier 1 or Tier 2 cities transfer to Tier 3. (shows in table 6.11, 6.12 , 6.13, 6.14, 6.15, 6.16, 6.17, 6.18,)

Similar performance was also achieved when transfer from Tier 1 Cities (Beijing, Shanghai) and Tier 2 Cities (Qingdao, Jinan) to target city Baotou (Tier 3) was compared. The result for this test are shown in Table 6.15, 6.16, 6.17 and 6.18.

| Semi-Supervised | Training Size | **Beijing ->Hohhot** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 5 Records Per Residence | 420 | 0.188 | 0.148 | 0.459 |
| 10 Records Per Residence | 840 | 0.172 | 0.139 | 0.486 |
| 15 Records Per Residence | 126,0 | **0.163** | **0.124** | **0.547** |
| Fully Supervised (No Transfer) | 5,048 | 0.170 | 0.129 | 0.523 |

Table 6.11: Performance Scores for Beijing to Hohhot

| Semi-Supervised | Training Size | **Shanghai ->Hohhot** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 5 Records Per Residence | 420 | 0.178 | 0.138 | 0.446 |
| 10 Records Per Residence | 840 | **0.166** | 0.133 | 0.512 |
| 15 Records Per Residence | 1260 | **0.161** | **0.126** | **0.588** |
| Fully Supervised (No Transfer) | 5048 | 0.170 | 0.129 | 0.523 |

Table 6.12: Performance Scores for Shanghai to Hohhot

Table 6.11 and Table 6.12, showed with 15 records of the data for the transfer from Tier 1 cities to Tier 3, Hohhot, results in a better performance model compared with fully supervised model on Hohhot.

| Semi-Supervised | Training Size | **Jinan->Hohhot** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 5 Records Per Residence | 420 | 0.188 | 0.144 | 0.509 |
| 10 Records Per Residence | 840 | 0.177 | 0.138 | 0.521 |
| 15 Records Per Residence | 1260 | **0.164** | **0.126** | **0.548** |
| Fully Supervised (No Transfer) | 5048 | 0.170 | 0.129 | 0.523 |

Table 6.13: Performance Scores for Jinan to Hohhot

| Semi-Supervised | Training Size | **Qingdao->Hohhot** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 5 Records Per Residence | 420 | 0.185 | 0.142 | 0.482 |
| 10 Records Per Residence | 840 | 0.174 | 0.132 | 0.555 |
| 15 Records Per Residence | 1260 | **0.155** | **0.123** | **0.585** |
| Fully Supervised (No Transfer) | 5048 | 0.170 | 0.129 | 0.523 |

Table 6.14: Performance Scores for Qingdao to Hohhot

As shown in Table 6.13 and Table 6.14 similar evaluation result occur Tier 2 cities on the transfer to Hohhot. This shows that Tier 2 city transfer to Tier 3 cities also delivers a better performing model. (Shown in Figure 6.13 and 6.14)

Similar performance is also achieved when transfer from Tier 1 Cities (Beijing, Shanghai) and Tier 2 Cities (Qingdao, Jinan) to target city Baotou (Tier 3). The result for this experiment are shown in Table 6.15, 6.16, 6.17 and 6.18

| Semi-Supervised | Training Size | **Beijing->Baotou** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 5 Records Per Residence | 550 | 0.140 | 0.105 | 0.424 |
| 10 Records Per Residence | 1100 | **0.135** | **0.099** | 0.456 |
| 15 Records Per Residence | 1650 | **0.121** | **0.090** | **0.513** |
| Fully Supervised (No Transfer) | 8652 | 0.135 | 0.102 | 0.473 |

Table 6.15: Performance Scores for Beijing to Baotou

| Semi-Supervised | Training Size | **Shanghai->Baotou** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 5 Records Per Residence | 550 | 0.141 | 0.105 | 0.436 |
| 10 Records Per Residence | 1100 | **0.133** | **0.100** | 0.455 |
| 15 Records Per Residence | 1650 | **0.127** | **0.099** | **0.520** |
| Fully Supervised (No Transfer) | 8652 | 0.135 | 0.102 | 0.473 |

Table 6.16: Performance Scores for Shanghai to Baotou

| Semi-Supervised | Training Size | **Jinan->Baotou** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 5 Records Per Residence | 550 | 0.142 | 0.107 | 0.347 |
| 10 Records Per Residence | 1100 | 0.137 | **0.100** | 0.448 |
| 15 Records Per Residence | 1650 | **0.127** | **0.096** | **0.516** |
| Fully Supervised (No Transfer) | 8652 | 0.135 | 0.102 | 0.473 |

Table 6.17: Performance Scores for Jinan to Baotou

| Semi-Supervised | Training Size | **Qingdao->Baotou** | | |
|---|---|---|---|---|
| | | RMSE | MAPE | $R^2$ |
| 5 Records Per Residence | 550 | 0.141 | **0.108** | 0.422 |
| 10 Records Per Residence | 1100 | 0.136 | **0.099** | 0.451 |
| 15 Records Per Residence | 1650 | **0.125** | **0.093** | **0.521** |
| Fully Supervised (No Transfer) | 8652 | 0.135 | 0.102 | 0.473 |

Table 6.18: Performance Scores for Qingdao to Baotou

Table 6.11, 6.12, 6.15 and 6.16 demonstrate the model transfer from Tier 1 cities (Beijing, Shanghai) to Tier 3 cities (Hohhot, Baotou). Table 6.13, 6.14, 6.17 and 6.18 demonstrate the model transfer from Tier 2 cities (Jinan, Qingdao) to Tier 3 cities (Hohhot, Baotou). Using only 15 records from each residential community, the proposed model already outperforms the fully supervised model. In other words, the performance of the proposed model yield significant improvement after transferring either Tier 1 to Tier 3 or Tier 2 to Tier 3 cities. This shows that by transferring the homogeneous property features learning network trained from a more substantial training data (Tier 1 or 2 cities) can help to boost the performance of low data volume Tier 3 cities.

# Chapter 7

# Discussion and Conclusion

## 7.1 Introduction

This chapter provides a summary of the findings from the results of our experimentation and analysis. Specifically, we discuss the results in relation to the objectives that were outlined in Chapter 1. Each of the following subsections detail the contribution from this research and they seek to address every one of those objectives identified. Following this discussion, this study proceeds to discuss about the managerial and policy implications from the findings. The former is largely focused more on the issues uncovered that will aid system implementers to looking to implement and utilise an AVM. The latter discusses the issues for the implementation in China and the broader national implications for such system to be used in a more intensive manner. Lastly, all research projects have their limits, and this study is not unique to further improvement. Hence, this study provides some suggestions for future researchers to pursue in this interested topic.

## 7.2 Discovery 1: The proposed cross-city AVM in reducing cognitive effort of human appraisers in the context of high-volume appraisal activities

There is certainly no comparison here between human appraisal effort and one that is supported by AVMs. In this study, AVMs that are powered by ANN could process thousands of appraisals within an hour, while human effort in practice requires 1 to 3 days to value a property. Where there is no recent transaction that accounts for the present value, human appraisers draw on similar properties from another city but make adjustments for economic-traits within the city to arrive at a final valuation price. The proposed ANN could achieve the same thing with a more rigorous, systematic and repeatable scientific approach on a large-scale basis. This certainly helps for in large-scale underwriting and refinancing purposes, where the proposed method can provide quick verification of property values, leading to significant savings in human resources. It allows investors and lenders to continuously update their risk profile as the debt service coverage ratio and loan-to-value information can be constantly updated with less effort.

The approach adopted can also provide different entities ranging from insurance companies to institutional investors with almost instant revaluation of their balance sheets, helping bypass the quarterly or annual activity that is tedious and time-consuming. This is even more useful when the market is highly volatile and requires almost instant reevaluation.

Since mass appraisals can be automated using an ANN network on

a large-scale basis, it allows for ease of integration into the development of financial trading strategies on a real-time basis. This will open the scope for new innovations in the underwriting process, such as valuing an real-estate asset as a new investment model. It also paves a way for possible future arbitrage trading backed by real estate investment trusts who can avoid a costly underwriting evaluation process.

## 7.3 Discovery 2: The proposed cross-city AVMs ability to consider unique and heterogenous traits

This study contributed to one of the pressing issues in mass appraisal models, that is, the lack of flexibility to account for cross-location in the evaluation exercise. Since data across different cities are composed of completely different location traits, traditional property appraisal models need to be adjusted from the beginning without assuming any prior information for each city. With large datasets this could have been a monumental task if not for the proposed novel semi-supervised homogeneous features transfer and heterogeneous location fine-tuning network that was devised in this study.

This studys proposed model has the ability to transfer the homogeneous feature learning component from a source city and fine-tune by a small amount for the target citys location features. This semi-supervised model can achieve a similar or even superior performance compared to that of a fully supervised model, with significantly less effort. A fully supervised model is akin has unconscious bias, consistency, clarity and repeatability benefits com-

pared to the best human effort because it is automated. However, this studys semi-supervised approach developed demonstrated that only one fifth of training data is required compared to a fully supervised ANN model for similar performance.

Hence, the developed method demonstrates that when real estate appraisal models are trained on data-rich cities, theinferred relationships (tacit knowledge) can be transferred to cities with insufficient real estate data without sacrificing accuracy. This approach reduces the period of data collection such that it lowers the models training cost and provide a better benchmarking and national and regional consistency when evaluating property values.

## 7.4 Discovery 3: The role of proposed cross-city AVM to reduce human bias in the smoothing of the appraised properties

The methods can implement the reduction in subsequent human bias (there may be bias in the original datasets uesd) and demonstrate the simplicity of using ANNs to execute reasonably good results with little effort when processing data. Bias can occur when a human valuer needs to classify an property into a collection of specific features or traits while judging across the many trade-offs required to arrive a linearising function for their decisions. This study avoids linear appraisal methods that produce significant errors and thus in mispricing. Specifically, this methods avoids the inability to capture appropriate value due to the nonlinearities of the characteristics impacting value as faced by hedonic valuation models. The designed cross-city model addressed the non-linearity

issue by demonstrating that it sidesteps the non-linear least squares approach in statistical models where there are not much guidance received from the functional form. In addition, the approach designed can reduce human bias as it does not rely on human intervention in addressing the property characteristics which uses simple counts for the number of baths or stories (floors) or location code. In such cases, mostly statistical models are applied, and this causes the matrix of explanatory variables to be filled with a lot of dummy variables, resulting in an inability to decide between the importance of those variables. This produces bad estimates for the pricing models coefficients. As such, this studys approach removes redundant information across the nodes and because there is no unnecessary information matrix to invert, there is no need for specification searches. Hence, this study's approach reduces equivocality, which often arise from the conflicting interests of stakeholders in the property valuation. This allows for an impartial and rational process at the core of a mass appraisal exercise.

## 7.5 Discovery 4: The proposed cross-city AVMs role in reduce uncertainties in present value estimate

This proposed approach helps in bridging the uncertainty involved in tax base valuation,where standardizing real estate tax valuatons and behavior will be able to enhance the tax calculation exercise. Since an automatic valuation model supported by technology is integral for proper mass appraisals, it can significantly help in the reform of fiscal and taxation systems and adapt quickly

to new policies. This is particularly important in the rapid and fast-paced development happening in the Chinese real estate market.

The challenge in reducing uncertainty also need to support the communication of results to let the public know clearly how the price valuation was reached. This is important for a widespread use of the ANN devised models as the accuracy and robustness of the appraisals need to meet the stakeholders confidence requirement. The reduction of uncertainties in the present value estimate will increase confidence from the international valuer's network and will allow investors to be more accepting of such techniques, based on a large-scale automated model.

## 7.6 Managerial Implications

The author anticipates that there will still be some resistance to adoption of AVMs within organisation active in the current system. This is because valuer's feel that their role will be undermined and their position within the industry weakened. However, AVMs powered by ANNs are not meant to reduce the significance of the valuer's but meant to assist them to arrive at a better decision and address more properties. Inevitably manual appraisal for standard properties will eventually be a thing of the past due to its potential for bias, corruption and time delays. Nonetheless, non standard and high value properties may still need significant human intervention but they can always be complemented using ANN methodology. The real-estate appraisal industry will have to take a significant role in educating government and institutional investors of the significance of AVMs role in aiding lenders and investors in managing their asset profile. However once government is aware of the con-

stancy and repeatability benefit resulting in improved trust, they may legislate for the use of AVM based methods. This does create an opportunity for governments to tweak the algorithms applied to better align property value with policy objectives.

## 7.7 Policy Implications

The concern of governments to accept automated large scale mass appraisals is in the fairness and reliability of the results. The concern is always on the black box nature of such methods, and it is difficult to communicate to government officials or the public that the appraisal values are self-learnt even though the prices may be closer to the true transaction value. In China, this is slightly easier to achieve because of the growing demand and high volume of property transactions. A large scale automated system is more likely to be adopted faster in such an enviroment. In fast moving economic circumstances the time than to wait for human appraisers to value the property can cause problem in itself. This is important taking into consideration the amount of investments and new property development that has been taking place in China in the last decade. New townships are formed and lenders are increasingly concerned that their risk portfolio may not be current. In addition, there is also imbalances in developments across regions in China. This imbalances create an opportunity for transferring knowledge of property valuations from one city to another. Using the proposed method developed, allows knowledge to be transferred from more developed regions to less developed regions. This means that less developed regions will attain significant insights into their new developments and lenders can use it as a form of predictive tool for their future planning.

## 7.8 Limitations and Suggestions for Future Research

The field of real estate appraisal is a promising avenue for the implementation of decision support system for appraisal assessment. Therefore, it is not only an important discipline for academia but it is also highly relevant for practitioners, commercial organisations and government bodies. There are still further challenges that need to be met from this research. Research could investigate the ethical issues in automated property price valuation. When a large central body with higher IT capability influence property prices, what are the measures that are in place that could protect institutional investors can is an area of research. The issues are similar to the research in ethics for AI are currently under much discussion. The work discussed could be extended by having applying an optimisation routing or the analogous to a regression method for feature extraction so that the value relationship each features contribute to the proportion of the valuation price can be determined.

The potential of the automated mass appraisal system as the tool of business innovation could be a destroyer to the current mass appraisal market, due the uniqueness of the method. This potentially could enable smaller firms to achieve similar or even better result as compared to existing large players in the appraisal market.

The present method may still be further explored especially on the the trial and error nature of the underlying ANN model that can create slightly different results each time the model is run. This is a may be an issue for predictive accuracy and consistency. Although this is not a problem statement for this study, it is worth appreciating that ANNs requires appropriate selection

of a number of variables. Future research may need to test their observations made here, with a number of variables with a diverse optimal parameters and verify how far the result deviates from the ones derived by the authors. There is no actual recommended settings for real-estate research but it was widely known that ANNs accuracy for classification problems is sensitive to what the analyst set for the parameters set by an analyst. These needs to be ascertained and tested by future research in mass-appraisal research with machine learning. Studies that wish to extend this research may also go into the domain of fuzzy features in the real-estate forefront. This would provide better assessment of the property value as fuzzy logic can better represent the uncertainty in the judging the many options and property features. This may also create scope for existing valuers in the future industry, as if they see no role, they will resist and perhaps even sabotage implementation of mass appraisal systems. The application of fuzzy logic should be evaluated with respect to the sample size, variable selection and the AI technique used in the AVM.

In conclusion, this research has investigated the previous and current approaches to this problem via an extensive literature review. Lack of data is a major problem with traditional approaches that cannot be readily applied for fast growing economies or economies where the data has not been digitized and connected. AVM of mass appraisal system using ANN with cross-city capability as devised and tested in this research helps address these issues and enables another step towards a mass appraisal system that is fit for a fast changing world, such as the one in China.

# References


Ahn, J. J., Byun, H. W., Oh, K. J. and Kim, T. Y. (2012), 'Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting', *Expert Systems with Applications* **39**(9), 8369–8379.

Arribas, I., García, F., Guijarro, F., Oliver, J. and Tamošiūnienė, R. (2016), 'Mass Appraisal of Residential Real Estate Using Multilevel Modelling', *International Journal of Strategic Property Management* .

Bagnoli, C. and Smith, H. (1998), 'The theory of fuzz logic and its application to real estate valuation', *Journal of Real Estate Research* **16**(2), 169–200.

Ball, M. J. (1973), 'Recent empirical work on the determinants of relative house prices', *Urban studies* **10**(2), 213–233.

Bin, O. (2004), 'A prediction comparison of housing sales prices by parametric versus semi-parametric regressions', *Journal of Housing Economics* **13**(1), 68–84.

Bohari, S. N., Saad, S. N. M., Marzukhi, F., Rahim, A. and Darim, A. D. A. (2015), Residential property valuation using gis, *in* 'Signal Processing & Its Applications (CSPA), 2015 IEEE 11th International Colloquium on', IEEE, pp. 131–134.

Borst, R. (1995), 'Artificial neural networks in mass appraisal', *Journal of Property Tax Assessment & Administration* **1**(2), 5–15.

Borst, R. and McCluskey, W. (2007), 'Comparative evaluation of the comparable sales method with geostatistical valuation models', *Pacific Rim Property Research Journal* **13**(1), 106–129.

Bourassa, S. C., Cantoni, E. and Hoesli, M. (2007), 'Spatial dependence, housing submarkets, and house price prediction', *The Journal of Real Estate Finance and Economics* **35**(2), 143–160.

Brown, G. R. and Matysiak, G. A. (2000), Real estate investment: A capital market approach, PhD thesis, Univerza v Mariboru, Ekonomsko-poslovna fakulteta.

Brown, K. H. and Uyar, B. (2004), 'A Hierarchical Linear Model Approach for Assessing the Effects of House and Neighborhood Characteristics on Housing Prices', *Journal of Real Estate and Education* .

Cao, J. A. (2015), *The Chinese real estate market: Development, regulation and investment*, Routledge.

Case, B., Clapp, J., Dubin, R. and Rodriguez, M. (2004), 'Modeling spatial and temporal house price patterns: A comparison of four models', *The Journal of Real Estate Finance and Economics* **29**(2), 167–191.

Cervelló, R., García, F. and Guijarro, F. (2011), 'Ranking Residential Properties by a Multicriteria Single Price Model', *Journal of the Operational Research Society* .

Coleman, J. W. and Larsen, J. E. (1991), 'Alternative estimation techniques for linear appraisal models', *The Appraisal Journal* **59**(4), 526.

Colwell, P. F. and Munneke, H. J. (2003), 'Estimating a price surface for vacant land in an urban area', *Land Economics* **79**(1), 15–28.

Crosby, N., Lavers, A. and Murdoch, J. (1998), 'Property valuation variation and the'margin of error'in the uk', *Journal of Property Research* **15**(4), 305–330.

D 'amato, M. (2007), 'Comparing Rough Set Theory with Multiple Regression Analysis as Automated Valuation Methodologies', *International Real Estate Review* .

De Mare, G., Lenza, T. L. L. and Conte, R. (2012), Economic evaluations using genetic algorithms to determine the territorial impact caused by high speed railways, *in* 'Proceedings of World Academy of Science, Engineering and Technology', number 71, World Academy of Science, Engineering and Technology (WASET), p. 1906.

Del Giudice, V., De Paola, P. and Forte, F. (2017), 'Using Genetic Algorithms for Real Estate Appraisals', *Buildings* .

Des Rosiers, F., Thériault, M. and Villeneuve, P.-Y. (2000), 'Sorting out access and neighbourhood factors in hedonic price modelling', *Journal of Property Investment & Finance* **18**(3), 291–315.

Diaz, J. and Hansz, J. A. (2002), Behavioral research into the real estate valuation process: Progress toward a descriptive model, *in* 'Real Estate Valuation Theory', Springer, pp. 3–29.

Do, A. Q. and Grudnitski, G. (1992), 'A neural network approach to residential property appraisal', *The Real Estate Appraiser* **58**(3), 38–45.

Downie, M.-L. and Robson, G. (2008), 'Automated valuation models: an international perspective'.

Fisher, I. (1906), *The nature of capital and income*, The Macmillan Company.

Fitch Ratings Structured Finance, K. (2012), 'Residential Mortgage Assumptions', (July), 2016–2018.

Francke, M. K. and Vos, G. A. (2004), 'The hierarchical trend model for property valuation and local price indices', *The Journal of Real Estate Finance and Economics* **28**(2-3), 179–208.

Fu, L.-M. (2003), *Neural networks in computer intelligence*, Tata McGraw-Hill Education.

Fu, Y., Liu, G., Papadimitriou, S., Xiong, H., Ge, Y., Zhu, H. and Zhu, C. (2015), Real estate ranking via mixed land-use latent models, *in* 'Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 299–308.

Galal, A. (2001), *Reforming land and real estate markets*, number June.
**URL:** *http://documents.worldbank.org/curated/en/2001/06/1346408/reforming-land-real-estate-markets%5Cnhttp://www-wds.worldbank.org/external/default/WDSConten*

Garrod, G. and Willis, K. (1992), 'The environmental economic impact of woodland: A two-stage hedonic price model of the amenity value of forestry in britain', *Applied economics* **24**(7), 715–728.

Geltner, D. (1989), 'Estimating real estate's systematic risk from aggregate level appraisal-based returns', *Real Estate Economics* **17**(4), 463–481.

Geltner, D., Miller, N. G., Clayton, J. and Eichholtz, P. (2001), *Commercial real estate analysis and investments*, Vol. 1, South-western Cincinnati, OH.

Geoghegan, J., Wainger, L. A. and Bockstael, N. E. (1997), 'Spatial landscape indices in a hedonic framework: an ecological economics analysis using gis', *Ecological economics* **23**(3), 251–264.

González, M. A. S., Soiberman, L. and Formoso, C. T. (2002), Explaining results in a neural-mass appraisal model, *in* '9th European Real Estate Society Conference (ERES)'.

Guan, J., Shi, D., Zurada, J. M. and Levitan, A. S. (2014), 'Analyzing massive data sets: an adaptive fuzzy neural approach for prediction, with a real estate illustration', *Journal of Organizational Computing and Electronic Commerce* **24**(1), 94–112.

Guntermann, K. L. and Norrbin, S. (1987), 'Explaining the variability of apartment rents', *Real Estate Economics* **15**(4), 321–340.

Harvey, J. (2016), *Urban land economics*, Macmillan International Higher Education.

Hassoun, M. H. (1995), *Fundamentals of artificial neural networks*, MIT press.

Herz, A., Sulzer, B., Kühn, R. and Van Hemmen, J. (1988), 'The hebb rule: Storing static and dynamic objects in an associative neural network', *EPL (Europhysics Letters)* **7**(7), 663.

Hill, R. C., Knight, J. R. and Sirmans, C. (1997), 'Estimating capital asset price indexes', *Review of Economics and Statistics* **79**(2), 226–233.

Hoesli, M., Giaccotto, C. and Favarger, P. (1997), 'Three new real estate price indices for geneva, switzerland', *The Journal of Real Estate Finance and Economics* **15**(1), 93–109.

Hoesli, M., Thion, B. and Watkins, C. (1997), 'A hedonic investigation of the rental value of apartments in central bordeaux', *Journal of Property Research* **14**(1), 15–26.

IAAO (2015), 'Standard on Mass Appraisal of Real Property', (December).

International Valuation Standards Council (2017), *International Valuation Standards (2017)*, International Valuation Standards Council, London, UK.

Jansen, S. J., Coolen, H. C. and Goetgeluk, R. W. (2011), *The measurement and analysis of housing preference and choice*, Springer.

Jeff, H. (2011), 'Programming neural networks with encog3 in java'.

Jenkins, D., Lewis, O., Almond, N., Gronow, S. and Ware, J. (1999), 'Towards an intelligent residential appraisal model', *Journal of Property Research* **16**(1), 67–90.

Ji, Y. (1999), 'Research on related issues of tax base appraisal', *State Assets Management* **6**, 26–28.

Ji, Y. and Fu, C. (2005), 'Mass appraisal: A tax base appraisal method for ad valorem tax', *Appraisal Journal of China* **11**, 5–9.

Ji, Y., Wang, C. and Fu, C. (2006), 'Application research of avm in mass appraisal systems of foreign countries', *Appraisal Journal of China* **3**, 13–17.

Kaastra, I. and Boyd, M. (1996), 'Designing a neural network for forecasting financial and economic time series', *Neurocomputing* **10**(3), 215–236.

Kang, H.-B. and Reichert, A. K. (1991), 'An empirical analysis of hedonic regression and grid-adjustment techniques in real estate appraisal', *Real Estate Economics* **19**(1), 70–91.

Kanojia, A., Khan, M. Y. and Jadhav, U. (2016), 'Valuation of Residential Properties by Hedonic Pricing Method- A State of Art', *International Journal of Recent Advances in Engineering & Technology (IJRAET)* .

Kauko, T. (2002), Modelling location in house prices–neural network and value tree approaches, PhD thesis, PhD Thesis. Utrecht.

Kauko, T. (2003), 'Residential Property Value and Locational Externalities: On the Complementarity and Substitutability of Approaches', *Journal of Property Investment & Finance* .

Kauko, T. and d'Amato, M. (2008), 'Introduction: Suitability issues in mass appraisal methodology', *Mass appraisal methods: An international perspective for property valuers* pp. 1–24.

Kauko, T. and dAmato, M. (2008), 'Mass appraisal methods', *An international perspective for property valuers (Chichester, UK* .

Khalafallah, A. (2008), 'Neural network based model for predicting housing market performance', *Tsinghua Science and Technology* **13**(S1), 325–328.

Kohonen, T. (1982), 'Analysis of a simple self-organizing process', *Biological cybernetics* **44**(2), 135–140.

Kok, N., Koponen, E.-L. and Martínez-Barbosa, C. A. (2017), 'Big data in real estate? from manual appraisal to automated valuation', *The Journal of Portfolio Management* **43**(6), 202–211.

Laakso, S. (1997), 'Urban housing prices and the demand for housing characteristics', *The Research Institute of the Finnish Economy (ETLA) A* **27**.

Lake, I. R., Lovett, A. A., Bateman, I. J. and Langford, I. H. (1998), 'Modelling environmental influences on property prices in an urban environment', *Computers, Environment and Urban Systems* **22**(2), 121–136.

LeCun, Y., Bengio, Y. and Hinton, G. (2015), 'Deep learning', *Nature* **521**(7553), 436–444.

Lee, Y.-L., Jung, C., Kuang, Y. Y. et al. (2003), Fair evaluation of real estate value in urban area via fuzzy theory, Technical report, European Real Estate Society (ERES).

Lentz, G. and Wang, K. (1998), 'Residential appraisal and the lending process: A survey of issues', *Journal of Real Estate Research* **15**(1), 11–39.

LIU, X.-s., Zhe, D. and WANG, T.-l. (2011), 'Real estate appraisal system based on gis and bp neural network', *Transactions of Nonferrous Metals Society of China* **21**, s626–s630.

Lu, j. (2014), 'Real estate appraisal model and empirical research based on genetic algorithm', *Computer Science* **11**, 75–78.

Lützkendorf, T. and Lorenz, D. (2005), 'Sustainable property investment: valuing sustainable buildings through property performance assessment', *Building Research & Information* **33**(3), 212–234.

Ma, H., Chen, M. and Zhang, J. (2015), 'The prediction of real estate price index based on improved neural network algorithm', *Advanced Science and Technology Letters* **81**, 10–15.

Manganelli, B., De Mare, G. and Nesticò, A. (2015), Using genetic algorithms in the housing market analysis, *in* 'International Conference on Computational Science and Its Applications', Springer, pp. 36–45.

Mansfield, J. R. and Lorenz, D. P. (2004), 'Shaping the Future: The Impacts of Evolving International Accounting Standards on Valuation Practice in the UK and Germany', *Property Management* .

Mas, v. H., Mauricio, A. and Crespo, F. (2016), 'Property Valuation using Machine Learning Algorithms : A Study in a Metropolitan-Area of Chile .', (May 2017), 97–105.

Mason, C. and Quigley, J. M. (1996), 'Non-parametric hedonic housing prices', *Housing studies* **11**(3), 373–385.

Mason, C. and Quigley, J. M. (2007), 'The curious institution of mobile home rent control', *Journal of Housing Economics* **16**(2), 189–208.

McCluskey, W. and Anand, S. (1999*a*), 'The application of intelligent hybrid techniques for the mass appraisal of residential properties', *Journal of Property Investment & Finance* **17**(3), 218–239.

Mccluskey, W. and Anand, S. (1999*b*), 'The Application of Intelligent Hybrid Techniques for the Mass Appraisal of Residential Properties', *Journal of Property Investment & Finance* .

McCluskey, W., Davis, P., Haran, M., McCord, M. and McIlhatton, D. (2012), 'The potential of artificial neural networks in mass appraisal: the case revisited', *Journal of Financial Management of Property and Construction* **17**(3), 274–292.

McCluskey, W., Deddis, W., Mannis, A., McBurney, D. and Borst, R. (1997), 'Interactive application of computer assisted mass appraisal and geographic information systems', *Journal of Property Valuation and Investment* **15**(5), 448–465.

Michaels, R. G., Smith, V. K. et al. (1990), 'Market segmentation and valuing amenities with hedonic models: the case of hazardous waste sites', *Journal of Urban Economics* **28**(2), 223–242.

Miller, N. G. (1982), 'Residential property hedonic pricing models: a review', *Research in Real Estate* **2**(1), 31–56.

Mo, L. (2014), 'Commercial real estate price evaluation in the circumstance of imposing real estate tax: Based on improved particle swarm optimization bpnn', *Journal of ZHONGNAN University of Economics and Law* **4**, 38–43.

Moodys Investors Service, K. (2012), 'Certification document 2012', pp. 1–108. **URL:** *http://www.moodys.com/sites/products/ProductAttachments/NRSRO 2013.pdf*

Moré, J. J. (1978), The levenberg-marquardt algorithm: implementation and theory, *in* 'Numerical analysis', Springer, pp. 105–116.

Needham, D., Franke, M. and Bosma, P. (1998), 'How the city of amsterdam is using econometric modelling to value real estate', *Journal of Property Tax Asessment and Administration* **3**, 25–46.

Nghiep, N. and Al, C. (2001), 'Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks', *Journal of real estate research* .

Orford, S. (2017), *Valuing the built environment: GIS and house price analysis*, Routledge.

O'Roarty, B., McGreal, S., Adair, A. and Patterson, D. (1997), 'Case-based reasoning and retail rent determination', *Journal of Property Research* **14**(4), 309–328.

Pace, R. K. (1995), 'Parametric, semiparametric, and nonparametric estimation of characteristic values within mass assessment and hedonic pricing models', *The Journal of Real Estate Finance and Economics* **11**(3), 195–217.

Pavlov, A. D. (2000), 'Space-varying regression coefficients: a semi-parametric approach applied to real estate markets', *Real Estate Economics* **28**(2), 249–283.

Quan, D. C. and Quigley, J. M. (1991), 'Price formation and the appraisal function in real estate markets', *The Journal of Real Estate Finance and Economics* **4**(2), 127–146.

Riddel, M. (2001), 'A dynamic approach to estimating hedonic prices for environmental goods: an application to open space purchase', *Land Economics* **77**(4), 494–512.

Ripley, B. D. (1993), 'Statistical aspects of neural networks', *Networks and chaosstatistical and probabilistic aspects* **50**, 40–123.

Robinson, R. (1979), *Housing economics and public policy*, Springer.

Rodriguez, M., Sirmans, C. and Marks, A. (1995), 'Using geographic information systems to improve real estate analysis', *Journal of Real Estate Research* **10**(2), 163–173.

Rosenblatt, F. (1958), 'The perceptron: a probabilistic model for information storage and organization in the brain.', *Psychological review* **65**(6), 386.

Rosenblatt, F. (1962), 'Principles of neurodynamics'.

Rossini, P. A., Kershaw, P. J. et al. (2008), 'Automated valuation model accuracy: some empirical testing'.

Rothenberg, J., Galster, G. C., Butler, R. V. and Pitkin, J. R. (1991), *The maze of urban housing markets: Theory, evidence, and policy*, University of Chicago Press.

Schulz, R., Wersing, M. and Werwatz, A. (2014), 'Automated valuation modelling: a specification exercise', *Journal of Property Research* **31**(2), 131–153.

Scott, I. and Gronow, S. (1990), 'Valuation expertise: its nature and application', *Journal of Valuation* **8**(4), 362–375.

Sheng, L. and Tang, A. (1998), 'A method of real estate market comparison approach appraisal based on neural networks', *Systems Engineering Theory and Practice* **3**, 52–56.

Simonotti, M. (2009), 'La segmentazione del mercato immobiliare per la stima degli immobili urbani', *Aestimum* (28).

Stevenson, S. (2004), 'Testing the statistical significance of real estate in an international mixed asset portfolio', *Journal of Property Investment & Finance* **22**(1), 11–24.

Sui, D. Z. (1992), 'A fuzzy gis modeling approach for urban land evaluation', *Computers, environment and urban systems* **16**(2), 101–115.

Sun, F. (2013), 'Application research of avm in mass appraisal systems of foreign countries', *Local financial research* **3**, 54–57.

Tan, K. H., Lim, C. P., Platts, K. and Koay, H. S. (2006), 'An intelligent decision support system for manufacturing technology investments', *International journal of production economics* **104**(1), 179–190.

Tay, D. P. and Ho, D. K. (1992), 'Artificial intelligence and the mass appraisal of residential apartments', *Journal of Property Valuation and Investment* **10**(2), 525–540.

Theebe, M. A. (2004), 'Planes, trains, and automobiles: the impact of traffic noise on house prices', *The Journal of Real Estate Finance and Economics* **28**(2-3), 209–234.

Tian, Y. and Yang, J. (2014), 'Application of geographic information system

on urban residential real estate mass appraisal.', *Applied Mechanics & Materials* .

Verkooijen, W. J. H. et al. (1996), Neural networks in economic modelling: An empirical study, Technical report, Tilburg University, School of Economics and Management.

Wang, J., Li, Z. and Tam, V. W. (2014), 'Critical factors in effective construction waste minimization at the design stage: a shenzhen case study, china', *Resources, Conservation and Recycling* **82**, 1–7.

Wang, Q. and Liu, Y. (2014), 'Improving tax base assessment system of the real estate tax in china by referring to international experiences', *International Taxation in China* **4**, 31–34.

Wang, Z., Chen, C., Guo, B., Yu, Z. and Zhou, X. (2016), 'Internet plus in china', *It Professional* **18**(3), 5–8.

Wei, q., Liu, S. and Jin, H. (2016), 'Applicability of the new standard of city-size classification in china', *Progress in Geography* **35**(1), 47–56.

White, H. (1989), 'Learning in artificial neural networks: A statistical perspective', *Neural computation* **1**(4), 425–464.

Widrow, B. and Lehr, M. A. (1995), 'Perceptrons, adalines, and backpropagation', *Arbib* **4**, 719–724.

Wilhelmsson, M. (2000), Traffic noise and property values, PhD thesis, Institutionen för fastigheter och byggande.

Wilhelmsson, M. (2002), 'Spatial models in real estate economics', *Housing, theory and society* **19**(2), 92–101.

Worzala, E., Lenk, M. and Silva, A. (1995), 'An exploration of neural networks and its application to real estate valuation', *Journal of Real Estate Research* **10**(2), 185–201.

Wyatt, P. (1996*a*), 'The development of a property information system for valuation using a geographical information system (gis)', *Journal of property Research* **13**(4), 317–336.

Wyatt, P. (1996*b*), 'Using a geographical information system for property valuation', *Journal of Property Valuation and Investment* **14**(1), 67–79.

Wyatt, P. (1997), 'The development of a gis-based property information system for real estate valuation', *International Journal of Geographical Information Science* **11**(5), 435–450.

Yang, L. (2002), 'Designing the system of real estate rent appraisal based on bp neural network', *Chinese Journal of Management Science* **4**, 23–27.

Yang, L. and Liu, K. (2004), 'Application of bp neural network in real estate appraisal', *Journal of Hebei Institute of Architectural Science and Technology* **21**, 105–109.

Yeh, I.-C. and Hsu, T.-K. (2018), 'Building real estate valuation models with comparative approach through case-based reasoning', *Applied Soft Computing* **65**, 260–271.

Yeh, I.-c., Hsu, T.-k. and Weight, C. (2018), 'Building real estate valuation models with comparative approach through case-based reasoning', *Applied Soft Computing Journal* **65**, 260–271.
**URL:** *https://doi.org/10.1016/j.asoc.2018.01.029*

Zhang, G., Patuwo, B. E. and Hu, M. Y. (1998), 'Forecasting with artificial neural networks:: The state of the art', *International journal of forecasting* **14**(1), 35–62.

Zhang, W. (2011), 'Application on real estate price prediction based on rs and pbnn', *Computer Simulation* **7**, 365–368.