Journal of Mathematics and Statistics 2012, 8 (4), 419-434
ISSN 1549-3644

doi:10.3844/jmssp.2012.419.434 Published Online 8 (4) 2012 (http://www.thescipub.com/jmss.toc)

Performance of Multiple Linear Regression and Nonlinear Neural Networks and Fuzzy Logic Techniques in Modelling House Prices

Siti Amri and Gurudeo Anand Tularam

Centre for Environmental Futures,
Faculty of Science, Environment, Engineering and Technology (ENV), Griffith University, Australia

Received 2012-06-14, Revised 2012-09-08; Accepted 2012-12-18

ABSTRACT

House price prediction continues to be important for government agencies insurance companies and real estate industry. This study investigates the performance of house sales price models based on linear and non-linear approaches to study the effects of selected variables. Linear stepwise Multiple Regression (MR) and nonlinear models of Neural Network (NN) and Adaptive Neuro-Fuzzy (ANFIS) are developed and compared. The GIS methods are used to integrate the data for the study area (Bathurst, Australia). While it was expected that the nonlinear methods would be much better the analysis shows NN and ANFIS are only slightly better than MR suggesting questions about high R^2 often found in the literature. While structural data and macro-finance variables may contribute to higher R^2 performance comparison was the goal of this study and besides the Australian data lacked structural elements. The results show that MR model could be improved. Also, the land value and location explained at best about 45% of the sale price variation. The analysis of price forecasts (within the 10% range of the actual prediction) on average revealed that the non-linear models performed slightly better (29%) than the linear (26%). The inclusion of social data improves the MR prediction in most of the suburbs. The suburbs analysis shows the importance of socially based locations and also variance due to types of housing dominant. In general terms of R^2, the NN model (0.45) performed only slightly better than ANFIS 0.39) and better than MR (0.37); but the linear MRsoc performed better (0.42). In suburb level, the NN model (7/15) performed better than ANFIS (3/15) but the linear MR (5/15) was better than ANFIS. The improved linear MR (6/15) performed nearly as well as the non-linear NN. Linear methods appear to just as precise as the more time consuming non linear methods in most cases for accounting for the differences and variation. However, when a much more in depth analysis is required non linear methods may prove to be more valuable. More research is needed in the area of house price modelling including more structural elements, modern buyer beliefs and the nature and type of risks noted in modern times.

Keywords: Neural Network (NN), Adaptive Neuro-Fuzzy (ANFIS), Multiple Regression (MR), Global Financial Crisis (GFC), International Monetray Fund (IMF), Statistics

1. INTRODUCTION

Housing investment is an important investment in Australia and worldwide and it is the means through which many have increased their wealth in the past. It is probably not possible for house prices to increase without bounds and as such the investment choice has been questioned causing a number of commentators (Stapledon, 2009) predicting a price decreases in not too distance future if only to stabilize the market. Whatever the market place does, it is nonetheless important to study the growth of house prices to predict prices for the use of real estate

Corresponding Author: Gurudeo Anand Tularam, Senior Lecturer in Mathematics and Statistics, Science Environment Engineering and Technology (ENV) Griffith University Brisbane, Australia

agents, insurance companies and government departments. Government agencies and councils can use the predicted prices to determine future rates payable amounts and use them to conduct forward planning such as budgeting for the corridors of growth in cities and so on.

In 2008, the International Monetray Fund (IMF) reported that the Australian house price was overpriced by around 20% (IMF, 2008). Given that house prices have softened in many overseas countries it is not surprising then that Australians are debating whether similar corrections may occur here. After the Global Financial Crisis (GFC) in 2007-8 the Australian government provided a number of stimulus measures and one of them to the housing industry; namely, the First Home Owners Grant that has allowed the Australian housing market to remain relatively strong. This has caused some heating of the market and the government acted swiftly increasing cash rates in order to slow the market and this has indeed slowed the growth rate and economy significantly in Australia. Since there has been over 70% owners occupiers in Australia, there is a vested interest in a strong housing market (ABS, 2009). In addition, the market has been favourably aided by government policy of negative gearing for investors to be active. The capital gains made in the past and the above mentioned tax policy have led to a significant increase in private investors. Evidence of this is noted in the proportion of investors taking out new housing loans increasing from 11-27% between 1990-2008 (ABS, 41020) but this appears to be slowing caused by many factors such as the global financial crisis and increasing interest rates. Nonetheless, significant increase in rental prices continue to lure new investors into the market.

As noted there is a need to estimated house prices but this is a challenge due to incomplete and "noisy" data of the market (Prasad and Richard, 2006). The data used in house price estimation is often based on historical sales records representing a small percentage of the total housing stock in a given area. Newly constructed homes also dominate the sales market and often there is lack of details regarding improvements or renovations as well as those pertaining to structural data. In this regard, the price reporting from previous sales records may not represent the true value of housing market appreciation. This suggests that the price patterns for many areas may be inaccurate due to a lack of appropriate sample information.

The housing system is a complex system. Housing data is complex and is kept in multiple data bases that is subject to ever changing financial and urban environment. However, there are some important factors impacting on prices such as accessibility, neighbourhood-quality, environmental, structural and speculation to name a few (Yates, 2002). Buyer's decision to pay above average prices for property in a premium location adds to the complexity in the development of accurate models. The local level factors include population demand, migration and other social variables, and the global level factors such as government policies and mortgage interest rates all have an impact on house price estimation (**Fig. 1**).

1.1. Modelling of House Prices

While there has been work done on house price modelling much is yet to be understood about house price fluctuations, patterns and growth (Farlow, 2004). The models can be grouped as those that use urban economics perspective focusing mainly on spatial; and those that use macroeconomic or financial economics. The latter focusing on housing as an asset using optimal methods in a consumption type model for example. Authors have used models based on heterogeneous assets labelled hedonic regression and also time series related models (Cannaday *et al.*, 2005; DEWHA, 2008; Nagaraja *et al.*, 2011; Shiller, 1993) while others use spatiotemporal models (Holly *et al.*, 2010; Yu *et al.*, 2007); Some have used the Torbin's Q model when others use genetic algorithmic type models (Wilson *et al.*, 2004). Amri and Bossomaier (2003) have attempted an agent based model but much work is yet undone in this area.

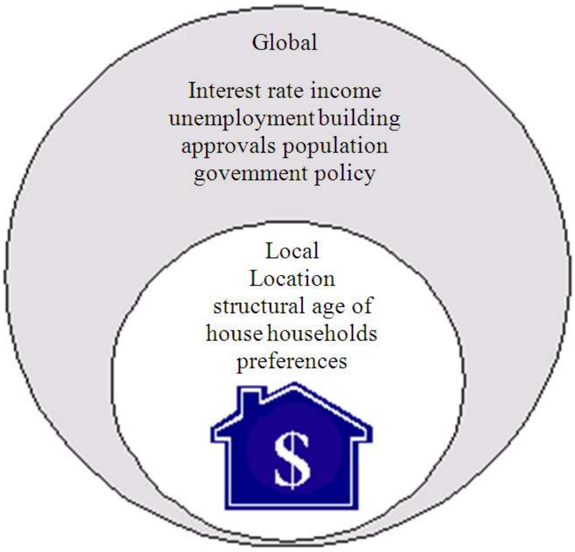

Global
Interest rate income unemployment building approvals population government policy
Local
Location structural age of house households preferences
$

Fig. 1. Complex housing system

A common feature of the housing market is that houses with similar structural attributes appear to differ in price based on their spatial attributes (neighborhood and location). The existences of several submarkets within the urban landscape have been subjected to several studies (Kauko, 2002; 2003). The importance of location in the housing submarket formation is also noted; that is, houses in different locations are not a perfect substitute. Goodman and Thibodeau (2003) used over 40 variables in hedonic modelling of house price estimation in Dallas (Texas, US) achieving 34% price range estimate within 10% of the observed prices. Their model included detailed structural information but rather sparse location attributes. In Puerto Allegre (Brazil), Gonzalez (2006) achieved high price estimation accuracy ($R^2 = 97$) using 14 variables (7 structural, 6 location and 1 time) based on a Fuzzy Rule Model (Fuzzy Rule is based on knowledge and simulation of expertise). The defining of the housing markets raises many questions: How does location affects house price? How does the local housing market evolve? How do consumers' motivational factors such as values and goals (housing preferences) affect property valuation?

Although much has been done in the past, much is still unknown as there seem to be not many answers forthcoming to the above questions in the literature. Farlow (2004) followed the economic and financial framework to its logical conclusion but failed to explain the patterns in house prices in UK. Siti-Nabiha *et al*., (2005) used an agent based model but little evidence agent models exists or those that include neural networks, fuzzy logic. Also, there is little evidence of detailed inclusion of structural element in the Australian modelling scene. It seems that there is a need to conduct more comparative studies that may help investigate house price modelling accuracy, explanation and prediction.

The primary aim of this study is to develop an appropriate database of house sales data so that a GIS can be used to develop a linear and two nonlinear models of house pricing in order to compare performance of each.

Data from Bathurst City (NSW) is used and models are compared in terms of accuracy of prediction using error analysis. Because of the ease of use and economy of the linear models, another aim is to attempt to increase accuracy of the linear model. The findings of this study may form a basis for the development of large scale dynamic house price models for other Australian cities.

1.2. Difficulties with Data for Modelling

As noted earlier, house price prediction has many challenges that include the information gap in sale information used that often constitute a small fraction of the actual housing stock for a given area. The housing stock is comprised of new constructed houses and existing residential properties. Houses that are sold in a specific period may be biased towards particular housing categories representing random availability of existing stock and new stocks. Also matching spatial and aspatial data adds another difficulty where data is lost due to currency problems as illustrated in **Fig. 2**.

The effect of data cleaning for linking with the GIS leads to a reduction of sales data size (**Fig. 2**). The sales records are more up-to-date than the land parcel data and this can be attributed to the difference in data currency and also the fact that the cadastre data are continuously evolving. In Bathurst local government area, there are about 200 new houses being added to the existing housing stock (BHS, 2001). The parcel data used for this study were obtained from the council in early 2003 and did not contain the new parcels development for 2003 and 2004.

Another disadvantage of existing data sets is their static nature. The housing system consists of processes and drivers including institutions and individuals behaviors. Financial and urban environment, continuing structural renovations and speculation all contribute to a dynamic price fluctuations market. The housing system processes and drivers interact to create spatial patterns and distributions of prices (i.e. housing markets). Data on the housing markets are often compiled as a snapshot of an event. In the case of spatial data, the focus has been on the land attributes. GIS serves as an expert system to analyze these spatial data sets but fails to represent the individuals and societies that influence the distribution and pattern of the housing system. The behavior of these individuals and their interactions on the housing consumption, create a dynamic system not captured by the traditional GIS. Temporal data modeling of the housing system, using data that are constantly changing, has not been attempted to date in the housing market.

1.3. Data Preparation

There are a number of sophisticated ways of developing and analysing complex data sets such as house sales price dataset. The Geographic Information System (GIS) data processing allows integration of data at different scales and can be applied to such datasets. In this system detailed housing characteristics can be combined with population statistics as well as policies or financial data. Studies have shown that modelling price estimation at a larger scale is feasible with the GIS technology (Lorenz *et al*., 2007). The varying mix of variables adopted by these studies reflects the data constraints and availability encountered in house price analysis.

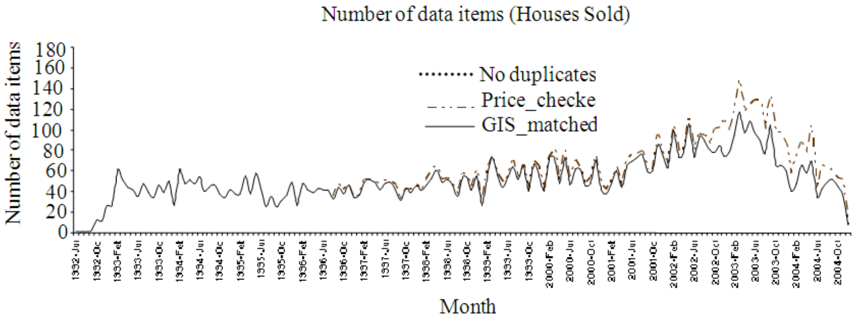


Fig. 2. Data Items (Houses Sold) variations

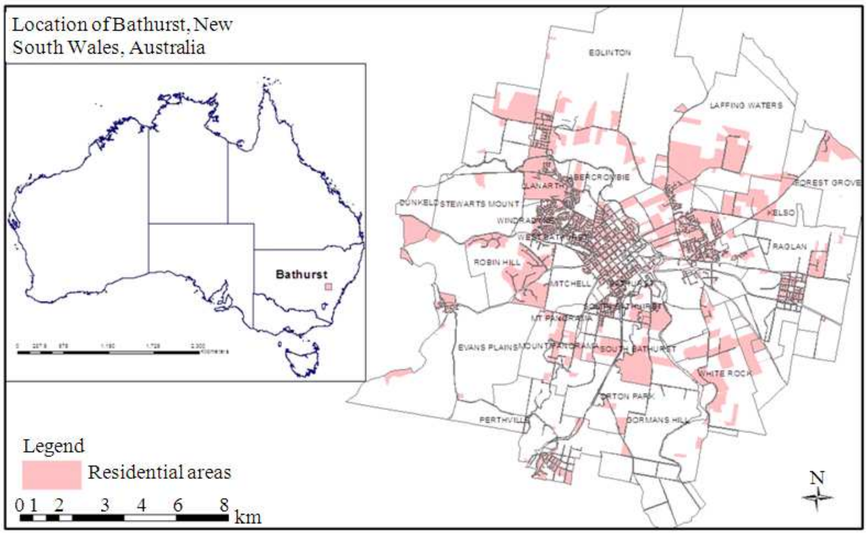


Fig. 3. Study area-Bathurst (NSW), Australia

Essentially, transaction sales records have varying standards in different countries and it seems that in the past for example structural data was not standard in Australian data bases.

Three data processing stages were employed on the sales data for preparation for matching to the cadastre database. These pre-processing included removing duplicates records, price-checking and GIS-matching. Three data lists were derived at each of the processing stages. After removing duplicate records, the number of sales record was reduced to 9265. This new list included repeat sales i.e. houses that have been sold more than once during the 13 year period. The sales data are further checked for price anomalies that don't reflect market price. The pruning process elliminated data that are under $20000 and over $1000000 with land area that are below 1000 square metres. In addition, sales records with no area information are also deleted from the subsequent list. The total price-checked sales record was reduced to 8841 unique sales. The GIS matching process reduced the data to 7849 unique sale records.

The effect of data cleaning for linking with the GIS data is the reduction of data size. The reasons can be attributed to the difference in data currency. The cadastre data are dynamic and were obtained from the council in early 2003 and did not contain the complete parcels record for 2003 and 2004. Data lost is consistent over the months 3-6 per cent for price-checked and 9-13% for GIS matched.

It is to be noted that some previous studies in large scale price estimation use a minimum of 1 or 2 variables to represent location characteristics with surprising level of accuracy (Lorenz *et al*., 2007). The notable difference between the data model used in this study and the others are the absence of structural data for the houses.

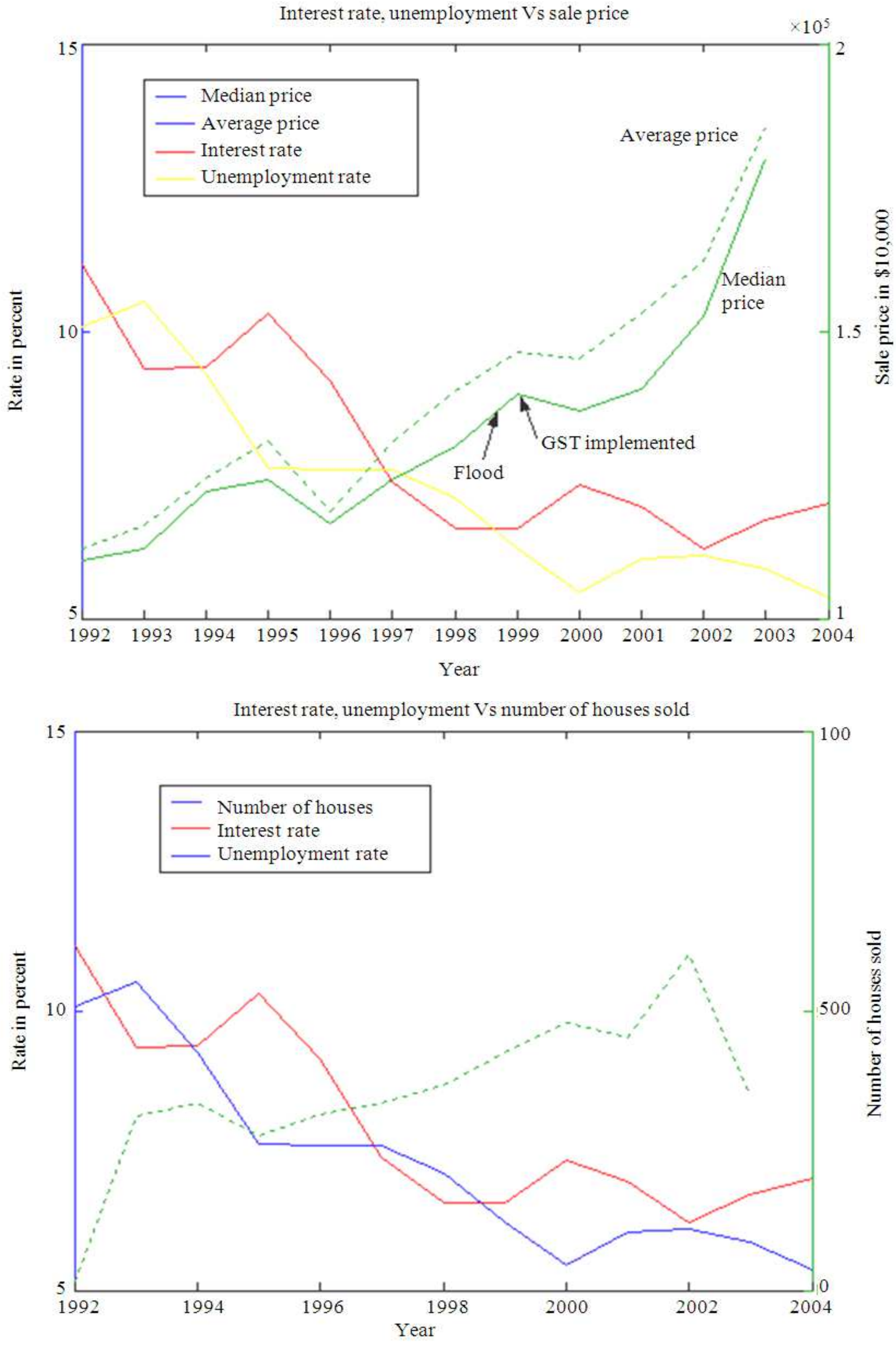


Fig. 4. Interest rate, unemployment Vs sale price and transaction volume (Bathurst, 1992-2004)

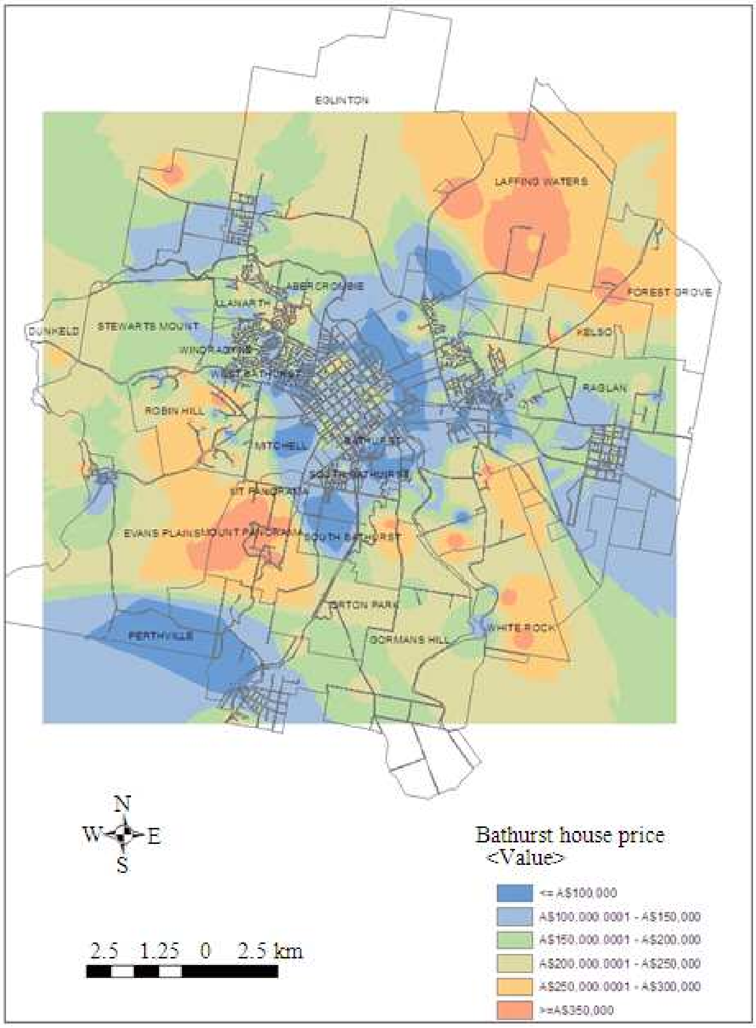


Fig. 5. Sale Price distribution in corresponding suburbs in Bathurst

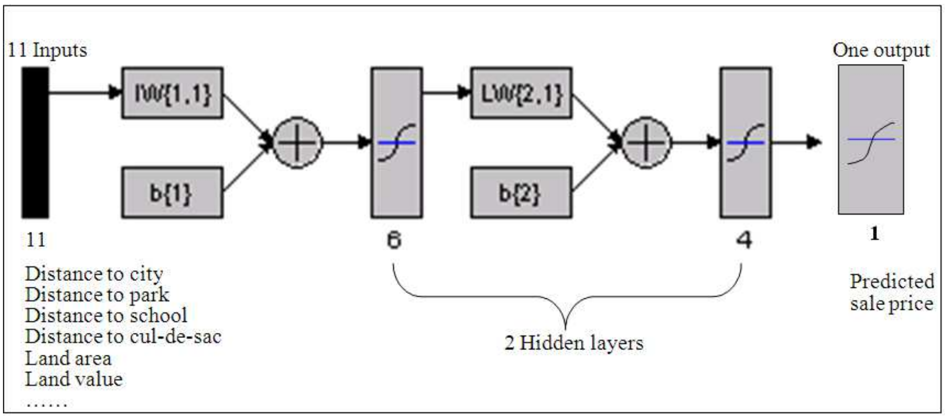


Fig. 6. Neural Networks (NN) architecture

1.6. Study Area

Bathurst city is built on the banks of the Macquarie River and was established in 1815 declaring a municipality in 1913. In 2011, it had a population of about 40 000, with the population increasing steadily from less than 32 000 in 1991 to over 36 000 in 2006. The urban rolling hills provide view to the distant ranges with Great Western Highway connecting city to Sydney (200 km). The city has a population annual growth of 2.2% and the median house price is $250 000 for 2010.

The initial settlement in the city is located along the Macquarie River with the older houses susceptible to flooding. Housing development is restricted in flood prone areas. However, flood mitigation has reduced flood risk in selected parts where medium density development has been initiated (Bathurst Plan, 2001). The spatial distribution of residential space has several geographic constraints (**Fig. 3**). The major development restrictions are related to flooding, the air corridor and special conservation areas that include the Mt Panorama racing circuit. There is a 50DBA Noise contour imposed around this area and houses are restricted within this zone. There are 2 major educational facilities; namely, Charles Sturt University and TAFE located in the vicinity. There are over 10 000 land parcels designated as residential. The inner city area was originally subdivided in 1833 into rectangular sections (200×200m) in a grid pattern. Some of the lots have been developed into medium-density housing or units and about 76% of dwellings are detached houses. 20.7 per cent consists of semi-detached units and flats and the remainder is commercial or government buildings (Housing Strategy, 2001). The parcel data is supplied by the Bathurst City Council consisting of detailed spatial characteristics of the house; the housing type is also included. The parcel data are dynamic, with about 200 new houses being added annually. The land values are based on the valuer estimates calculated every 3 years that form the basis for rate calculations; the land value of the house does not necessarily reflect current market conditions. The sales data from 1992-2004 are supplied by the Valuer General (NSW) linked to the parcel data. In a given year, houses for sale constitute a small percentage of the entire residential stock (less than 1%). The Great Western Highway, which links Bathurst and Sydney, has recently been upgraded improving commute time between the cities. With this change, the spill-over of residential demand from Sydney is expected to influence house price in the area. Approximately 50 km further west of Bathurst is another regional mining centre: Orange. Both Bathurst and Orange have almost similar population size. In summary, this small city provides a test bed for experimenting and conducting comparative work on modelling house prices.

1.7. Bathurst-the Study Area Statistics

Statistical and spatial trends of Bathurst housing from previous sales record covering the years between 1992-2004 are presented. The city has experienced positive growth averaging 5% each year during this time period. Renovation of existing homes is a significant housing activity but currently there is no record to link it to the houses sold. The median and the average house price move opposite to the interest rate and unemployment rate. **Figure 4** shows the implementation of GST in 1999 decreased average and median price; with -1 and -2% change respectively. The median and average prices were low in 1996 while in contrast the volume of sales for that year increased indicating that lower price houses were sought after. The flood event in 1999 (August) did not appear to have much effect to the general price trend. This is probably due to the fact that the flood is localised to the flood prone areas in the suburb of Kelso and CBD adjacent to the Macquarie River.

Existing houses sold annually averaged about 400 and new houses averaged about 200. New dwellings constituted about 50% of the total volume sold over the years. The number of houses sold varied inversely to the interest rate and unemployment as expected. **Figure 4** shows that lower interest and unemployment rate resulted in higher sales. **Table 1** shows that negative changes occurred in 1995, 2001 and 2003. The First Home Owners Grant (FHOG) was introduced in 1999 (when GST was introduced) supporting a buoyant housing market. The FHOG appeared to affect the average and median house price changes but not the volume of houses sold, suggesting an active housing market. **Figure 5** shows that in 2003 more expensive houses were transacted. The Bathurst housing market is dominated by newly built houses that are bigger in living spaces and in line with overall Australian trend.

Figure 5 shows house price spatial pattern with cheaper houses along the river and the great western highway. More expensive homes with larger land are on the periphery. Mt. Panorama, Robyn Hill and White Rock suburbs are predominantly with large homes on acreage. The cheaper housing can be found in the suburbs of Mitchell, South Bathurst, West Bathurst and in the villages of Eglinton, Raglan and Perthville. Lower prices are also found clustered in some parts of the city centre with higher prices for the acreage properties near the city perimeter. The expensive homes along Mitre and Howick Streets in the centre are period homes priced for their architecture.

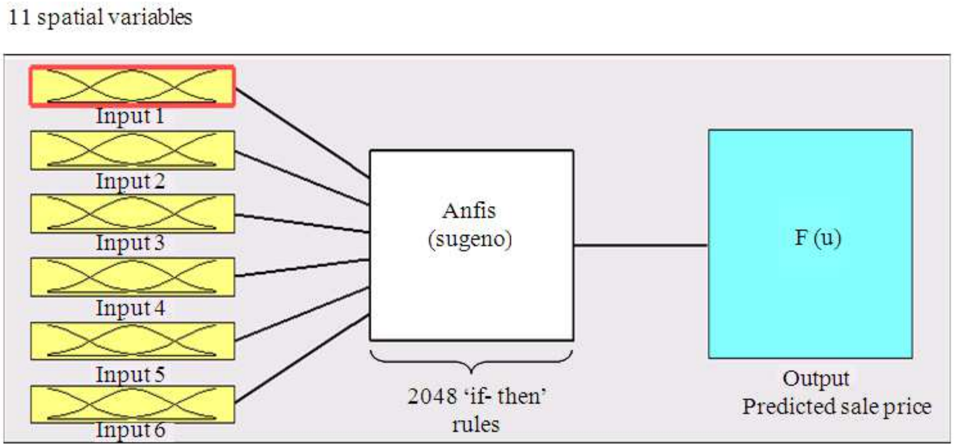


Fig. 7. The ANFIS architecture

Table 1. Average price and transaction volume in Bathurst (1992-2004)

Year	Avg. price	Percentage of Change	Med. price	Percentage of change	Houses sold	Percentage of change
1992	112230		110000		15.0	
1993	116140	3	112000	2	311	
1994	124460	7	122000	9	334	7
1995	130960	5	124000	2	276	-17
1996	118540	-9	116500	-6	314	14
1997	130510	10	124250	7	334	6
1998	139610	7	130000	5	367	10
1999	146260	5	139000	7	425	16
2000	145270	-1	136250	-2	478	12
2001	153070	5	140000	3	451	-6
2002	162170	6	152500	9	602	33
2003	185490	14	180000	18	348	-42
2004						
Mean	138726	5	132208	5	389	8

1.8. Models Used

Three models were used namely, Multiple Regression (MR), Neural Network (NN) and Neuro-Fuzzy Methods (ANFIS) to develop housing price models. All models were tested with the ten fold cross validation technique to test performance. The dataset was listed randomly and then divided into ten parts of equal proportions. Each part was further grouped into training and testing sets. The former constitutes 90% of the divided data and the later comprises the remaining 10%. According to Witten and Frank (1999), the ten-fold cross-validation technique, is a standard method for assessing generalization. In the following, each model is briefly explained in turn.

1.9. Multiple Regression (MR)

In MR model, the dependent variable-observed house price, is regressed against a set of predictors (or independent variables). The selected variables assumed to contribute to the house prices include land value, the size of the land (in square metres), distance to the main central business area, distance to the closest school, distance to park, distance to cul-de-sac, the topographical attributes (elevation, slope and aspect) and the map coordinates of the house.

The MR model is a function of the variables specified in Equation 1:

$$\begin{aligned}&\text{SalePrice} = f(\text{Land value, Area, Dcity,}\\&\text{Dschool, Dpark, Dculd,}\\&\text{Height, Slope\_deg, Aspect, Yearsold,}\\&\text{ X coordinates, Y-coordinates,}\mu)\end{aligned} \quad (1)$$

In abbreviated form the MR variables are:

SalePrice = Price of house in Australian dollars
Land_value = Estimated land value in Australian dollars

Area = The size of the land parcel in square metres
Dcity = Distance to the central business district
Dschool = Distance to the closest school
Dpark = Distance to the closest park
Dculd = Distance to the cul-de-sac
Height = The elevation of the parcel centroid
Slope_deg = The slope of the parcel centroid
Aspect = The aspect of the parcel centroid
Yearsold = The year when the house was sold
X and Y = Map coordinates of the parcel centroid
μ = Error term

A step-wise regression approach initially included all the independent variables and in turn discarded those collinear varaibles. The assumptions underlying such as the relationship between the dependent and independent variable is linear; it is assumed that no severe multicollinearity exists, outliers are not influencing analysis, errors are homoscedastic and errors are not autocorrelated.

1.10. Networks (NN)

Unlike MR no prior knowledge about the relationships between input and output variables is required since NN is data driven (Zhang *et al*., 2009). NN allows nonlinearity with complex mix of variables (Rossini, 1997). A major weakness for comparison of the model is that it performs like a "black box" with complex training procedure. The influence of a particular variable on the predicted price is not evident. A multilayer feed forward network is used with Levenberg-Marquardt (trainlm) method. The networks are then trained until a low mean-square error is achieved. The architecture that provides the best fit for the data is the networks with two hidden layers and an output layer with a (Coolen *et al*., 2002 ABS, 2009) configuration (**Fig. 6**). Over fitting occurs at a (Rossini, 1997; ABS, 2009) configuration when the testing error is considerably larger than the training error. Each of the layers uses the hyperbolic tangent sigmoid (tansig) transfer function to simulate the networks. The choice of function is based on the characteristic of the data (Maier and Dandy, 2000; Kalman and Kwasny, 1992).

1.11. Adaptive Neuro-Fuzzy System (ANFIS)

To account for model and parameter uncertainties a third approach neurofuzzy technique is also used. The spatial attributes are represented in a qualitative terms to describe their locational accessibility or proximity to amenities. For example, nearness to near to a school will depend on the buyer perceptions of accessibility. A buyer will have to decide on an arbitrary cut-off point, maybe 500 m or less to fix the limit considered accessible. That is, houses exceeding cut-off point have non accessible characteristics. In the fuzzy logic approach not accessible can be represented by a range of possible values. A function describing nearness is called the Membership Function (MF). The MF is the central idea that separates fuzzy sets and crisp sets. The MF defines how each point in the input space is mapped to an output MF with the value between 1 and 0 (Fuzzy Logic Toolbox, 2001). The shape and gradient of the MF dictates the weight of the input for mapping to the output function. Arbitrary selection of the MF may or may not truly reflect the input/output data for analysis.

In ANFIS an Adaptive Neuro-Fuzzy Inference System (ANFIS) is used to model house prices. The advantage of this technique is that the MF of the input parameters are chosen automatically using neuro-adaptive learning techniques, incorporated in the Fuzzy Logic Toolbox. These neuro-adaptive techniques provide a method for the fuzzy modeling procedure to learn information about a data set, in order to compute the membership function parameters that best allow the associated fuzzy inference system to track the given input/output data (Fuzzy Logic Toolbox, 2001). The price estimation will use selected spatial attributes as fuzzy inputs (11 inputs) and combined with Neural Networks to predict price (**Fig. 7**) with two MF chosen for input. The defuzzification process has 211 'if-then' rules. The 'dsigmf' (difference between 2 sigmoid functions) function is selected to describe the fuzzy inputs and "genfis2" is used to reduce the number of rules to less than 10.

The MR and NN models require minimum variables manipulation for model processing. In case of the Neuro-Fuzzy ANFIS model, subjective definitions of membership functions for independent variables may have comprised model accuracy. In all the three models, the same data are used to perform the analysis. For each technique, there can be further refinement and improvement; in this study, ease of use or simplicity and economy are consideration for refinement.

2. DISCUSSION

In terms of MR, NN and ANFIS models, the overall performance and performance by suburbs are investigated. The methods are evaluated based on accuracy of performance by analyzing adjusted R^2 values, mean square errors and the nature of spatially predicted values. **Table 4** summarizes performance of each model. The adjusted R squared and the prediction error percent within 10 and 20% of the actual sale price is also shown.

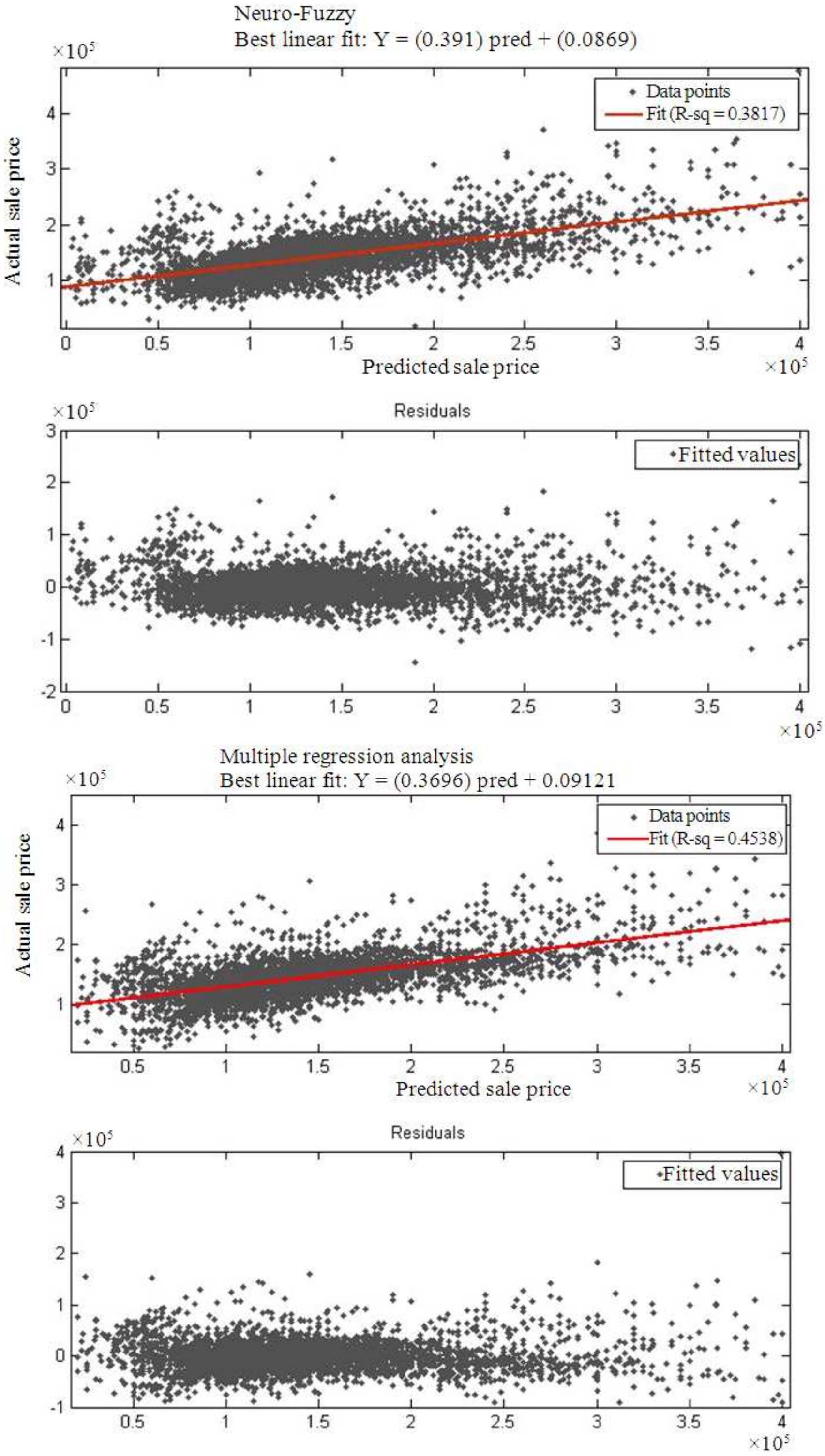


Fig. 8. NN, ANFIS and MR fit and residual plots

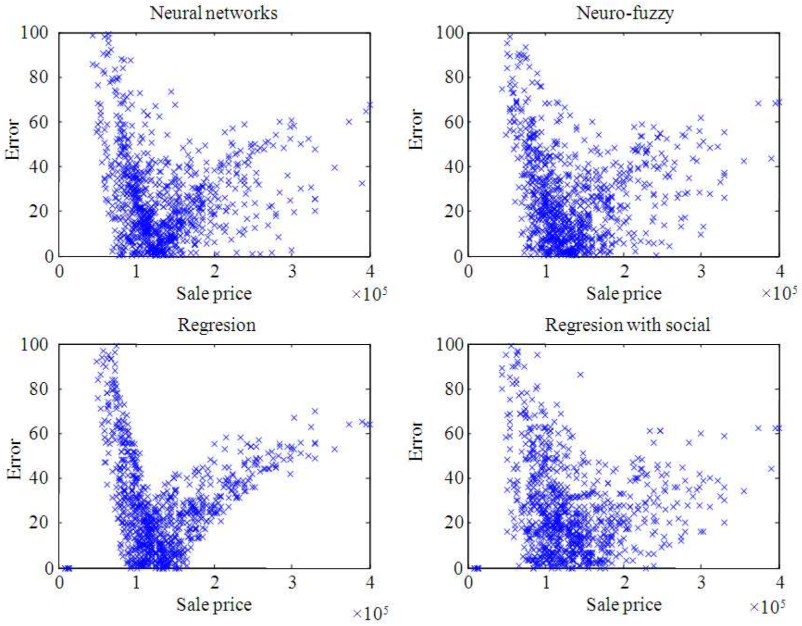


Fig. 9. Residual plots of the linear and non-linear models for Bathurst City suburb

Table 2. Comparison of MR, NN and NF models

	Neural networks	Neuro-fuzzy	Multiple regression
Adjusted R-square:	0.4536	0.3815	0.3694
Error <=10%	1310(31%)	1199(28%)	1176(28%)
Error <=20%	2372(56%)	2201(52%)	2159(51%)

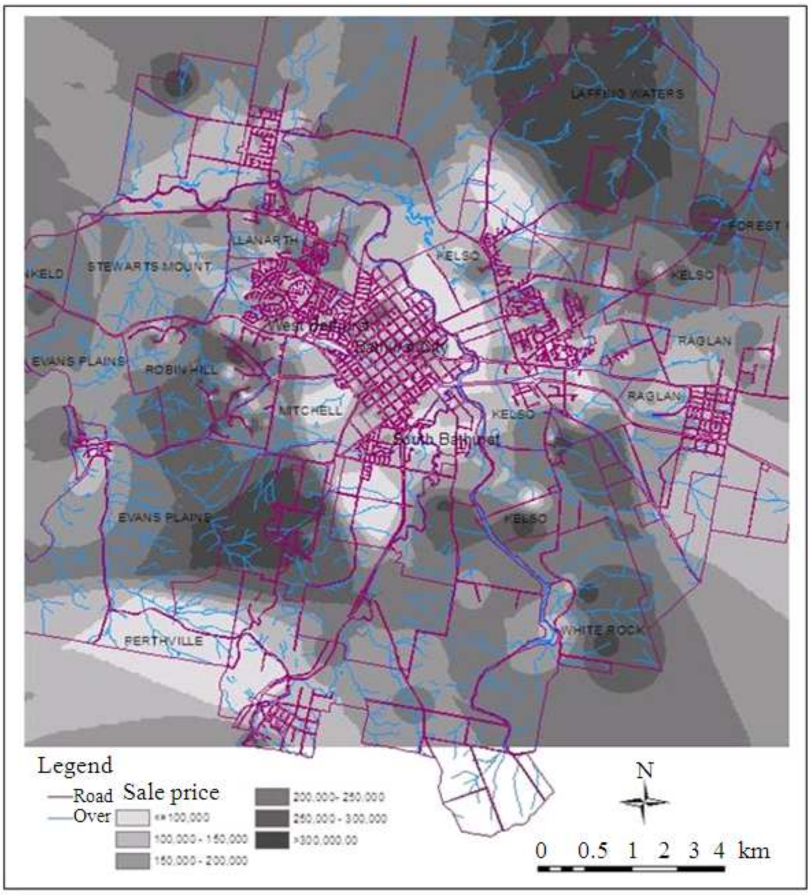
Table 2 shows that Neural Networks is the best performer followed closely by Neuro-Fuzzy and finally, the Multiple Regression. The **Table 2** also shows that error comparison revealed that all three models provide approximately 30% of the estimation that are within the 10% price range of the observed prices. Interestingly, all models provide more than 50% of price estimates that are within the 20% price range.

Each of the model performance is further analyzed using linear regression between the predictions and actual sale prices. **Figure 8** below shows the fit and residuals. All models show a good fit for houses in the range of $50000-$200000 but not for houses values outside this range. Based on adjusted R squared, the Neural Networks prediction showed better results than the fuzzy and MR. The results suggest systematic errors for houses with large differences to the local median. It is expected that the variance in selling prices will differ between the low and the high end of the housing market. There is greater potential range of prices for expensive houses than for low-priced houses. The variation unexplained in this study maybe due to a number of factors. The lack of structural, policy and macroeconomic factors such as movements in interest rate or change in tax incentives for example.

The MR model with parameters is shown in Equation 2:

$$\begin{aligned}&\text{Saleprice} = \\ &-1.720\text{E}7 + 1.4\text{Landvalue} + \\ &5491\text{Yearsold} \\ &+7491\text{Dcity} - 6305\text{Dculd} + \\ &1613\text{Slope} + 3\text{X} - 23052\text{Dpark}\end{aligned} \qquad (2)$$

The adjusted R-squared of the MR model is 37% indicating that the location and land value variables explained less than half of the variation. Land value explains about 20% of the sale variance. The coefficient of the independent variables in Equation 2 implies the direction and magnitude of influence to housing sale prices; housing location within the cul-de-sac or its proximity to a park has negative influence on price. The proximity to the central business district or the city centre, location (centroid coordinates x and y), land value and transaction date all have positive influence on the sale price. Clearly, there are other variables that could reduce variation. Coolen *et al.* (2002) recommended the inclusion of social factors such as household composition, age, income and current housing situation.



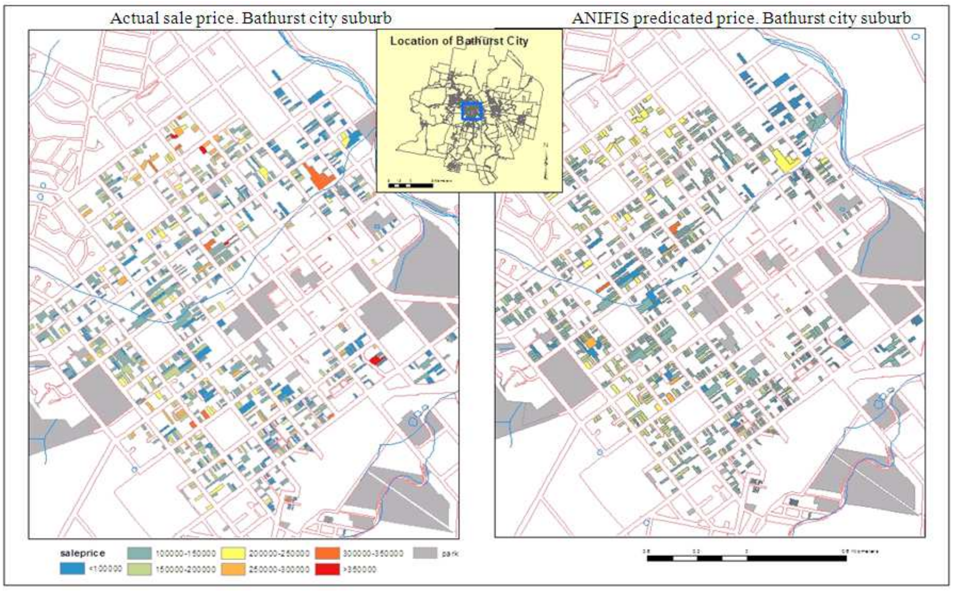

Actual sale price. Bathurst city suburb
Location of Bathurst City
ANIFIS predicated price. Bathurst city suburb
saleprice
<100000
100000-150000
150000-200000
200000-250000
250000-300000
300000-350000
>350000
park

Fig. 10. Actual Vs ANFIS price prediction map- the suburb of Bathurst City

For the purposes of an extended analysis, the MR model was used with the following census data added:

- Percent of the population that are owner occupiers
- Percent of the population that is living in public housing
- Percent of the population that is lone parent
- Percent of the population that is unemployed for male
- Percent of the population that is unemployed for female
- Percent of the population that have no cars
- Percent of the population that have 2 or more cars
- Percent of the population that is over 65 years of age
- Percent of the population that is under 34 and single
- Percent of the population that is of family households
- Percent of the population that is non-white households

The stepwise regression was applied on this additional data model and the MR equation determined is shown in Equation 3:

$$\begin{aligned}&\text{Saleprice} = -949465 + 1.2\text{landvalue}\\&+5537\text{yearsold} - 4053\text{dculd}\\&+1681\text{slope} + 1.4\text{y} - 17478\text{dpark} -\\&136656\text{owner} - 65225\text{public} - 75787\text{nocars}\\&+158193\text{morecars} - 52393\text{under}\\&34\sin gle + 61121\text{family}\end{aligned} \qquad (3)$$

The inclusion of social indicators improved the adjusted R-squared to 42% improving sale price prediction. Interestingly, this model excludes the distance to city centre and distance to cul-de-sac factors. The exclusion suggests that the model improves its variance with fewer location variables. The significant social variables noted are related to housing tenure, the ownership of cars and household composition at the local level.

Table 3 describes the performance of the models through detailed suburb analysis. In suburbs, the NN model (7/15) performed better than ANFIS (3/15) but linear MR (5/15) was not that low. The improved linear (6/15) almost performed as well as NN.

The variety of low, medium and high pricing houses in the different suburbs provide a unique housing market for Bathurst. The suburbs on the outskirts of the city such as Abercrombie, Llanarth, Kelso, Eglinton and Perthille are predominantly single detached housing. Bathurst City, South Bathurst, Gormans Hill, White Rock and West Bathurst have a significant medium and high-density housing. The expensive suburbs with median prices of over $250000 are in Llanarth, Robin Hill, White Rock and Mount Panorama. The suburbs with sold properties below $100000 are found in the largest three suburbs of Bathurst City, Kelso and West Bathurst.

The NN model performed better in terms of prediction accuracy for 7 of the suburbs. These suburbs are a mixed of low (Kelso, Eglinton, West and South Bathurst), medium (Windradyne, Llanarth) and acreage properties (Robin Hill). The MR model performed second with 5 suburbs - relatively small areas; Perthville and Raglan are villages on the outskirts of the city. Mitchell is a university accommodation area with high percentage of rental properties and students. Abercrombie and Gormans Hill are suburbs at the outer areas of the city.

ANFIS performed best in 2 suburbs - largest and in the smallest suburbs; Bathurst city and Mount Panorama respectively. The housing in Bathurst City suburb is predominantly medium price properties and Mount Panorama is premium properties on large acreage.

The addition of social data in the MRregsoc model affected performance. The MRregsoc model was better in 6 suburbs with a mix of villages (Perthville, Raglan), low (Kelso, Windradyne) and acreage properties (White Rock and Abercrombie). This suggests that social data does influence model performance in selected suburbs.

Table 4 describes the suburbs and the price forecasts that are within the 10% range of the actual prediction. The non-linear models perform slightly better than the linear models at average of 29% compared to 26%, except in the areas of Raglan, Abercrombie, Gormans Hill, Mitchell and Perthville. In MR, the inclusion of social data improves prediction in most of the suburbs except for Bathurst City, South Bathurst and Mitchell.

Figure 9 shows residual plots that are under 100% for the suburb of Bathurst City. This largest suburb in the study area includes the central business district area. Its proximity to the business area and educational facilities (Charles Sturt University and TAFE) made this suburb popular for renting. All models performed poorly for low and top range properties. Below $100000 price range, all models produced extremely large errors but in $100000-$200000, the error distribution and pattern are closer to the actual values. The residual plots for the suburb of Bathurst City also revealed not normal trends and the presence of outliers at both the lower and upper ends of the price range.

Figure 10 shows the actual sale price distribution for the Bathurst City suburb compared to the best performing model non-linear model ANFIS. The actual sale illustrated diversity of price ranges within the area. The ANFIS model prediction tends to standardize the price estimates into the middle range values. There is fewer frequency of price prediction in the top end of the price range.

Table 3. Performance of models by suburb in terms of adjusted R^2

Suburbs	Popn	Dwellings	Med price	<100K	100K-250K	>250K	Sold	Best model	Social data
Bathurst City	6600	2904	119000	0.3	0.65	0.05	957	anfis	
Kelso	6300	2157	135000	0.25	0.69	0.06	833	nn	regsoc
West Bathurst	4200	1453	115000	0.35	0.65	0.02	624	nn	
Windradyne	2600	891	150000	0.1	0.88	0.02	575	nn	regsoc
Eglinton	2300	729	136000	0.1	0.88	0.03	273	nn	
South Bathurst	1600	620	116800	0.28	0.71	0.01	255	nn	
Raglan	1100	343	140000	0.09	0.88	0.04	140	reg	regsoc
Abercrombie	1700	494	175000	0.11	0.7	0.19	140	reg	
Llanarth	1000	315	230000	0.08	0.53	0.39	137	nn	
Gormans Hill	520	222	106500	0.46	0.95	0.05	81	reg	
Robin Hill	840	243	247500	0.12	0.41	0.47	68	nn	regsoc
Mitchell	630	111	81250	0.67	0.31	0.02	54	reg	
Perthville	1000	305	118000	0.37	0.56	0.07	41	reg	regsoc
White Rock	440	108	254000	0.13	0.35	0.52	23	anfis	
Mt Panorama	440	150	350000	0	0.11	0.89	9	anfis	regsoc

Table 4. Performance by suburbs and errors

	Linear models				Non-linear models		
	----------------------------		----------------------------------		Average	Average non	
Suburbs	Sold	MR	MRregsoc	ANFIS	NN	linear models	-linear models
Bathurst City	957	0.24	0.236	0.260	0.242	0.24	0.25
Kelso	833	0.242	0.314	0.268	0.298	0.28	0.28
West Bathurst	624	0.199	0.312	0.292	0.345	0.26	0.32
Windradyne	575	0.314	0.368	0.294	0.353	0.34	0.32
Eglinton	273	0.26	0.366	0.366	0.370	0.31	0.37
South Bathurst	255	0.335	0.319	0.337	0.365	0.33	0.35
Raglan	140	0.357	0.450	0.286	0.293	0.40	0.29
Abercrombie	140	0.279	0.279	0.243	0.257	0.28	0.25
Llanarth	137	0.088	0.197	0.292	0.328	0.14	0.31
Gormans Hill	81	0.222	0.309	0.185	0.333	0.27	0.26
Robyn Hill	68	0.044	0.250	0.176	0.206	0.15	0.19
Mitchell	54	0.185	0.148	0.074	0.056	0.17	0.07
Perthville	41	0.317	0.366	0.268	0.293	0.34	0.28
White Rock	23	0.043	0.174	0.263	0.211	0.11	0.24
Mt Panorama	9	0	0.556	0.667	0.333	0.28	0.50
					Total average	0.26	0.29

3. CONCLUSION

This study explored the performance of linear and nonlinear models in predicting house prices for an Australian inland city Bathurst (NSW), Australia. The comparative study found that the non-linear models of NN and ANFIS perform better as expected but linear methods can be improved as well. The selected land value and location variables explained at best about 45% of the sale price variation. The exclusion of structural data and other macro finance variables possibly contribute to the lower prediction accuracy. Interestingly, sophisticated non-linear and linear models are almost similar in performance for house price prediction in many cases. Since the overall aim was to compare linear and non-linear models and assess performance in predicting house prices the lower R squared for all 3 models can be accepted; but the lower values can be attributed to existence of outliers, possibility of recording errors (due to paper records conversion to electronic data) and lack of structural information in Australian datasets. Comparison of the models at the suburbs level revealed variance in performance due to the type of housing that is dominant in the given suburbs. The price forecasts that are within the 10% range of the actual prediction revealed that the non-linear models perform slightly better than the linear models (29% compared to 26%); except in the areas of Raglan, Abercrombie, Gormans Hill, Mitchell and Perthville. The inclusion of social data improves prediction of MR in most of the suburbs except for Bathurst City, South Bathurst and Mitchell. The suburbs performance variation indicates the relevance of social location difference in house price prediction. In summary, the neural network model performs slightly better than the other non-linear model (ANFIS) and the linear models but it seems that the linear models can be improved by using judicious choice of data. Further study is needed with more recent sales data integrated with social aspects into the nonlinear models and given that structure is also important the inclusion of structural data into Australian datasets should be done to improve overall model prediction similar to those claimed overseas.

4. REFERENCES

ABS, 2009. Australian Social Trends.

Amri, S. and T. Bossomaier, 2003. Agent-based modelling of house price evolution. Proceedings of the Australian and New Zealand Intelligent Information Systems, Dec. 10-12, Sydney, pp: 15-20.

BHS, 2001. Bathurst local government housing report.

Cannaday, R., H. Munneke and T. Yang, 2005. A multivariate repeat-sales model for estimating house price indices. J. Urban Econom., 57: 320-342. DOI: 10.1016/j.jue.2004.12.001

Coolen, H., P. Boelhouwer and van K. Driel, 2002. Values and goals as determinants of intended tenure choice. J. Housing Built Environ., 17: 215-236. DOI: 10.1023/A:1020212400551

DEWHA, 2008. Energy Efficiency Rating and House Price in the ACT. Department of the Environment, Water, Heritage and the Arts.

Farlow, M.R. , 2004. NMDA receptor antagonists-A new therapeutic approach for Alzheimer's disease. Geriatrics Minnesota, 59: 22-27.

Gonzalez, S., 2006. Mass appraisal with genetic fuzzy rule-based systems. Property Manag., 24: 20-30. DOI: 10.1108/02637470610643092

Goodman, A.C. and T.G. Thibodeau, 2003. Housing market segmentation and hedonic prediction accuracy. J. Housing Economics, 12: 181-201. 10.1016/S1051-1377(03)00031-7

Holly, S., M. Pesaran and T. Yamagata, 2010. A spatio-temporal model of house prices in the USA. J. Econom., 158: 160-173. DOI: 10.1016/j.jeconom.2010.03.040

IMF, 2008. World economic and financial surveys, world economic outlook.

Kalman, B.L. and S.C. Kwasny, 1992. Why tanh: choosing a sigmoidal function. Proceedings of the IEEE International joint Conference on Neural Networks, Jun. 7-11, IEEE Xplore Press, Baltimore, MD, pp: 578-581. DOI: 10.1109/IJCNN.1992.227257

Kauko, J.K., 2002. Modelling the locational determinants of house prices: neural networks and value tree approaches, Phd thesis, University Utrecht.

Kauko, T., 2003. On current neural network applications involving spatial modelling of property prices. J. Hous. Built Environ., 18: 159-181. DOI: 10.1023/A:1023977111302

Lorenz, P.D., S. Truck and T. Lutzkendorf, 2007. Exploring the relationship between the sustainability of construction and market value. Property Manage., 25: 119-149. DOI: 10.1108/02637470710741506

Maier, R.H. and C.G. Dandy, 2000, Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environ. Modell. Software, 15: 101-124. DOI: 10.1016/S1364-8152(99)00007-9

Nagaraja, C., L. Brown and L. Zhao, 2011. An autoregressive approach to house price modeling. Annals Appl. Statist., 5: 124-149. DOI: 10.1214/10-AOAS380

Prasad, N. and A. Richards, 2006. Measuring housing price growth-using stratification to improve median-based measures.

Rossini, P.A., 1997. Application of artificial neural networks to the valuation of residential Property. Prodeedings of the 3rd Pacific Rim Real Estate Society Conference, (PRRESF' 97), New Zealand.

Shiller, R., 1993. Macro markets: Creating institutions for managing society's largest economic risks. Oxford University Press.

Stapledon, N., 2009. Housing and the global financial crisis: US versus Australia. Economic and Labour Relations Review, 19: 1-16.

Siti-Nabiha, A.K., Scapens and W. Robert, 2005. Stability and change: An institutionalist study of management accounting change. Account. Audit. Accountab. J., 18: 44-73.

Wilson, I., A. Jones, D. Jenkins and J. Ware, 2004. Predicting housing value: Genetic algorithm attribute selection and dependence modelling utilising the gamma test. Applications Artificial Intelligence Finance Econom. Advances Economet., 19: 243-275. DOI: 10.1016/S0731-9053(04)19010-5

Witten, I.H. and E. Frank, 1999. Data mining: Practical machine learning tools and techniques with Java implementations.

Yates, J., 2002. A spatial analysis of trends in housing markets and changing patterns of household structure and income.

Yu, D., Y. Wei and C. Wu, 2007. Modeling spatial dimensions of housing prices in Milwaukee, WI. WI, Environment Planning B: Planning Design, 34: 1085-1102.

Zhang, G., E.B. Patuwo and Y.M. Hu, 2009. Forecasting with artificial Neural Networks: The state of the art. Int. J. Forecasting, 14: 35-62.