Comparing Methods: Using Multilevel Modelling and Artificial Neural Networks in the Prediction of House Prices based on property, location and neighbourhood characteristics

Yingyu Feng[1] and Kelvyn Jones[2]

School of Geographical Sciences,

University of Bristol

Abstract

Two advanced modelling approaches, Multi-Level Models and Artificial Neural Networks are employed to model house prices. These approaches and the standard Hedonic Price Model are compared in terms of predictive accuracy, capability to capture location information, and their explanatory power. These models are applied to house prices in the Greater Bristol area, 2001-2013 using secondary data from the Land Registry, the Population Census and Neighbourhood Statistics so that these models could be applied nationally. The results indicate that MLM offers good predictive accuracy with high explanatory power, especially if neighbourhood effects are explored at multiple spatial scales.

KEYWORDS: House Prices, Multilevel Modelling, Artificial Neural Networks, Predictive Accuracy

1 Introduction

The principal objective of this paper is to present two advanced quantitative approaches, Multi-Level Models (MLM) and Artificial Neural Network (ANN). They are also compared with the baseline, widely-deployed, the standard Hedonic Price Model (HPM) in terms of predictive accuracy, capability to capture location information, and their explanatory power. There is no published work to date comparing ANN with MLM, both conceptually and in terms of practical adequacy. The use of a much larger dataset than previous publications is another important contribution of this study.

The rest of the paper is organized as follows. Section 2 provides a schematic specification of each modelling approach. Section 3 considers how the models are operationalised, in terms of study area and data, scenarios to capture locational information, and performance measures for competing models. The results are presented in section 4 with the conclusions in section 5. A review of the previous studies utilising ANN and MLM are included in the full paper.

2 Specification of HPM, MLM and ANN

2.1 The Hedonic Price Model

The Hedonic Price Model (HPM) was first introduced by Lancaster (1964) and formalised by Rosen (1974). Traditional HPM regresses the house prices on a range of its constituent attributes (e.g. number of bedrooms, type of house, or distance to the city centre and so on) and is usually calibrated using the ordinary least squares (OLS) method.

2.2 Multilevel Model

MLM (Goldstein, 1999) combines the micro-level equation, describing the within-neighbourhood between-house relationship for individual properties and the macro-level equation, the between-neighbourhood relationship, into one model, denote as

$$\boldsymbol{y_{ij} = \beta_0 + \beta_1 x_{ij} + (u_{0j} + u_{1j} x_{ij} + \mathbf{e}_{ij})} \quad (1)$$

1 yf13468@Bristol.ac.uk
2 kelvyn.Jones@Bristol.ac.uk

Where y_{ij} and x_{1ij} represent the house price and house-type for level-1 house i in level-2 neighbourhood j, respectively. If x_{1ij} is a binary variable with 1 being detached house and 0 being non-detached property, $\beta_{0j}$is the mean price of a non-detached property in a neighbourhood and β_{1j} is the mean differential for detached properties in a neighbourhood. The first part of the combined equation is called the fixed part representing the means and the second part (within the bracket) is called the random part. The terms u_{0j} and u_{1j} represent the unexplained price differentials of neighbourhood j in term of intercept and slope after taking into account house types. The distributional assumptions are denoted as

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix})$$

The random residuals e_{ij} represent the between-house price deviation within-neighbourhood j, and follow a normal distribution $e_{ij} \sim N(0, \sigma_e^2)$. These random effects are unmeasured or latent variables representing the differential 'desirability' of a neighbourhood for non-detached and detached property – the premium or discount people are willing to pay to live in that location. This allows the unobserved neighbourhood characteristics to play a role 'behind the scenes' in the effects of the observed variables on house prices (Snijders & Bosker, 1999).

2.3 Artificial Neural Networks

ANN is a form of artificial intelligence that consists of a number of interconnected processing elements, or neurons, that mimic the functions of biological neurons to process information in parallel (Caudill, 1988). One of the most popular ANNs is Multi-Layer Perceptron (MLP) as it is capable of mapping the non-linear relationships within the data. MLP typically contains an input layer, an output as the last layer, and one or more hidden layer(s) between the input and output layer. Each layer consists of a number of 'neurons', which are interconnected to the neurons in the immediate adjacent layers by a set of weights, representing the strength of connection. Back-propagation (BP) training algorithm (Rumelhart et al., 1986) is commonly used to train the ANNs. Once the relationship is established, the networks can be used to predict house prices. Figure 1 shows the structure of MLP with a single hidden layer.

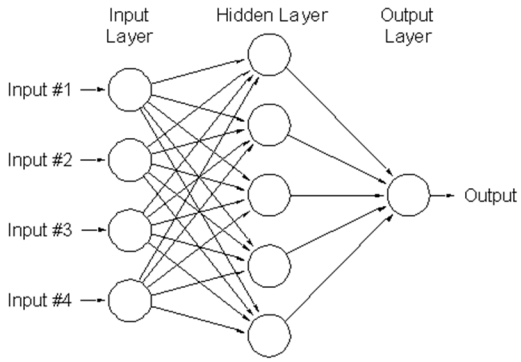

Figure 1 Multi-Layer Perceptron with a single hidden layer

3 Data and scenario comparison

3.1 Setting and Data

This study selects Greater Bristol as the study area. The house prices and property attributes (displayed in Table 1) are from the Land Registry of England and Wales (http://www.landregistry.gov.uk/) and neighbourhood characteristics (displayed in Table 2) are abstracted from the 2001 census data and the Neighbourhood Statistics website (https://neighbourhood.statistics.gov.uk/). The location of each sale is geocoded based on the unit postcode of the property to obtain its Ordnance Survey National Grid references. Output Areas (OAs) are selected as the lowest level of neighbourhoods, nested in lower layer super output areas (LSOAs) and then middle layer super output areas (MSOAs). The full dataset is 65,302 house sales, out of which 61,161 sales in 2001-2012 are used for model calibration and the rest 4141 sales in 2013 are used for prediction.

Table 1 Definition and explanation of house price data and attributes

Variables	Definition and explanation
Price	Sale price stated on the Transfer deed in thousands (£'000)
Yrmth	The year and the month when the sale was completed as stated on the Transfer deed, expressed in numerical form
Type	"Det" for Detached house
	"Semi" for Semi-Detached house
	"Terr" for Terraced houses
	"Flat" for Flats/Maisonettes
New	"New" for a newly built property
	"Old" for an established residential building
Duration	Types of legal interests in land:
	"Free" for freehold, where the legal interest in land is held by the owner of the land
	"Lease" for leasehold, where the interest in land or property is held by the tenant who lets the property from the landlord
East	The Ordnance Survey postcode grid reference: Easting
Nth	The Ordnance Survey postcode grid reference: Northing
Dist	Euclidean distance to the city centre, Cabot Circus in Bristol

3.2 Scenarios for comparison

Three scenarios are designed with different representations of space and place. Scenario 1 use property characteristics only, and three models are specified for each modelling approach, named HPM1, MLM1 and ANN1. In scenario 2, the Grid references of the property location (Easting and Northing) are included as additional explanatory variables. Six models, HPM2 (a)/(b), MLM2 (a)/(b) and ANN2 (a)/(b) are specified under this scenario where situation (a) uses the absolute locations while (b) additionally includes the distance to the city centre. In scenario 3, the Grid references and distance to city centre are replaced by measured neighbourhood characteristics. Another three models are calibrated, HPM3, MLM3 and ANN3. As the MLM requires the specification of the higher-level units, the nested multiple neighbourhood structure (OAs, LSOAs, and MSOAs) is used in all the scenarios.

3.3 Performance measures

In order to reach a balanced view of a model's performance, we have used R^2 as the goodness-of-fit measure, Mean Absolute Error (MAE) as the accuracy measure and Mean Absolute Percentage Error (MAPE) as a relative measure of accuracy. We have also examined the face validity of the models, the extent to which the explanatory power of property and neighbourhood variables in accounting for house price. Here we are particularly concerned with which variables are the most influential in determining the predicted prices.

Table 2 Definition of neighbourhood variables

Variables	Definition and Explanation	Level
IMD	2004 Index of Multiple Deprivation score (IMD)	LSOA
IMDbar	2004 deprivation score on "Barriers to Housing and Services", measuring barriers to housing such as affordability and geographical barriers to key local services.	LSOA
IMDenv	2004 Deprivation score in the living environment, comprising the 'indoors' living environment which measures the quality of housing and the 'outdoors' living environment for air quality and road traffic accidents.	LSOA
IMDcrime	2004 Deprivation Crime Domain Score, which measures the rate of recorded crime for four key dimensions of crime: burglary, theft, criminal damage and violence.	LSOA
Green	Green space area percentage of total land use area	OA
Det_area	Proportion of detached house	OA
Terr_area	Proportion of terrace house	OA
Flat_area	Proportion of flats	OA
Room	Average number of rooms per household, used as proxy of average size of properties	OA
Noheat	Proportion of houses that have no central heating	OA
Unemploy	Proportion of people aged 16-74 who are not in employment, including retired, students aged over 16 years old and other people	OA
Lnincome	Natural log of Experian income at MSOA level in 2004	MSOA
SocRent	Proportion of social rented from council or others	OA
PriRent	Proportion of private rented from council or others	OA
Occupancy	The Occupancy Rating provides a measure of under-occupancy and over-crowding. It relates the actual number of rooms to the number of rooms 'required' by the members of the household	OA
Young	Proportion of people aged 17 or under	OA
Old	Proportion of people aged 65 or older	OA
Black	Proportion of black ethnic	OA
Noedu	Proportion of people have no academic or professional qualifications	OA
Degree	Proportion of people have at least First degree or Higher degree	OA

4 Empirical results and discussion

Due to limited space, detailed model results are not included here but available on request. The comparison of goodness-of-fit are presented in Table 3 and the accuracy comparisons are based on the 4141 hold-out samples in 2013 and summarised in Table 4. All performance measures show that MLM is superior to ANN and HPM in each scenario, indicating that the specification of neighbourhood is helpful in house price predictions, even when the locations or neighbourhood characteristics have been included in the model. Once the appropriate hierarchical structure of housing market has been defined in MLM, location and neighbourhood characteristics will only further explain the price variation between neighbourhoods, but will not further improve the predictive accuracy.

Table 3 Comparisons of Goodness-of-fit

	Scenario 1			Scenario 2						Scenario 3		
	HPM1	MLM1	ANN1	HPM2a	MLM2a	ANN2a	HPM2b	MLM2b	ANN2b	HPM3	MLM3	ANN3
R^2 (in-sample)	0.39	**0.75**	0.39	0.43	**0.75**	0.41	0.43	**0.75**	0.47	0.68	**0.75**	0.69
R^2 (hold-out)	0.23	**0.75**	0.23	0.30	**0.75**	0.26	0.31	**0.75**	0.38	0.65	**0.74**	0.67

Table 4 Comparisons of prediction accuracies for hold-out samples

Hold-out sample:	**Scenario 1**			**Scenario 2**						**Scenario 3**		
4141 cases	HPM1	MLM1	ANN1	HPM2a	MLM2a	ANN2a	HPM2b	MLM2b	ANN2b	HPM3	MLM3	ANN3
MAE(lnP)	0.319	**0.178**	0.318	0.304	**0.178**	0.313	0.303	**0.178**	0.286	0.210	**0.178**	0.216
MAPE(lnP)	5.89%	**3.29%**	5.85%	5.61%	**3.29%**	5.76%	5.59%	**3.29%**	5.26%	3.89%	**3.30%**	4.00%
MAE(expo lnP)	80.4	**48.6**	80.1	77.0	**48.6**	79.0	76.9	**48.6**	73.7	56.3	**48.8**	56.6
MAPE(expo lnP)	30.9%	**17.5%**	30.0%	29.4%	**17.5%**	29.8%	29.4%	**17.5%**	27.2%	20.7%	**17.6%**	20.7%

In terms of model explanatory power, we have presented the relative importance of each predictor graphically. Figure 2 shows the predicted house price (after exponentiation) on a common scale. The predictions are for all sixteen types of property, holding everything else at their mean value (Figure 2a). The time effects (Figure 2b) are for the typical property in a typical neighbourhood. The effects of the neighbourhood characteristics (Figure 2 c-f) show the relationship between house price and each variable, holding all other variables constant at their typical value. In each figure, 95% confidence intervals of the in-sample model are also plotted. It can be seen that the average size of the property in OA, property characteristics and the time when it was sold are very important predictors.

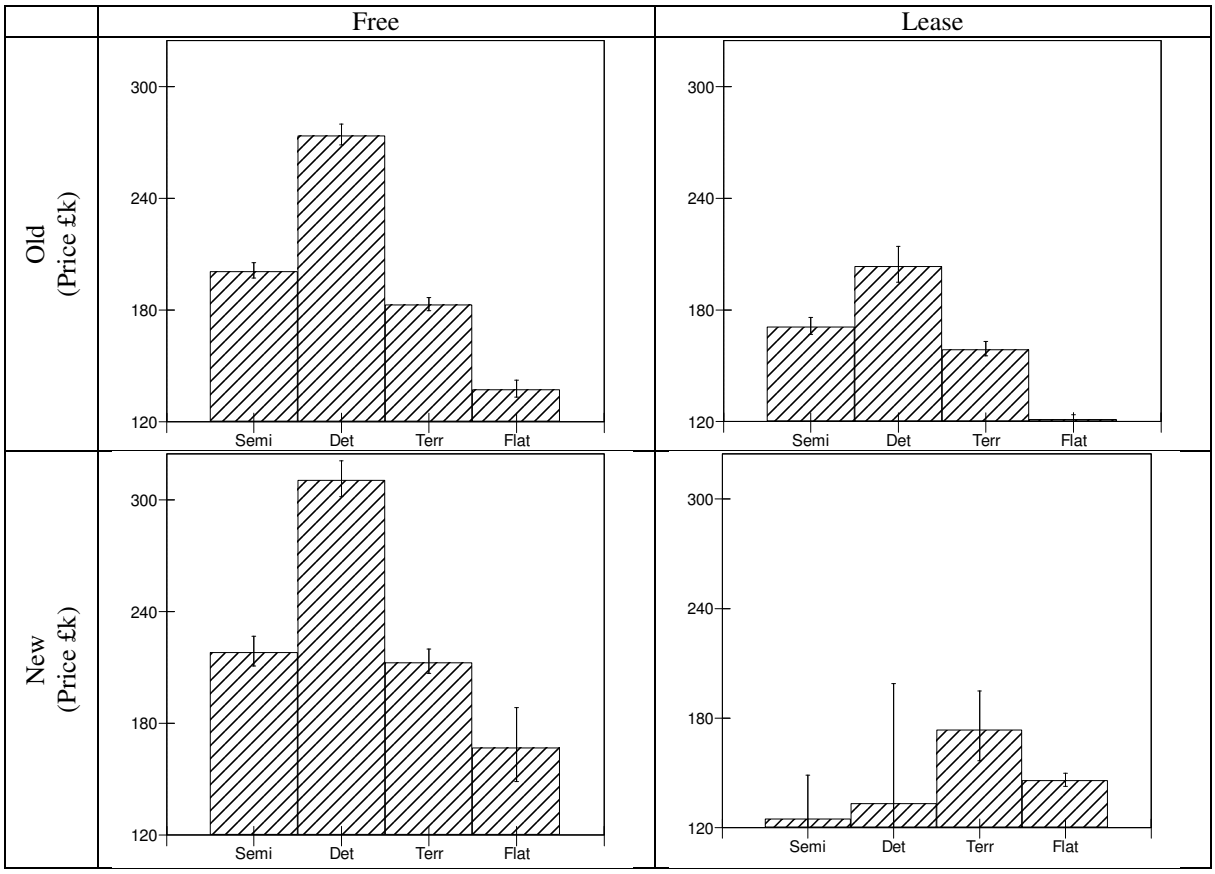

Figure 2 (a) Effect size of predictors

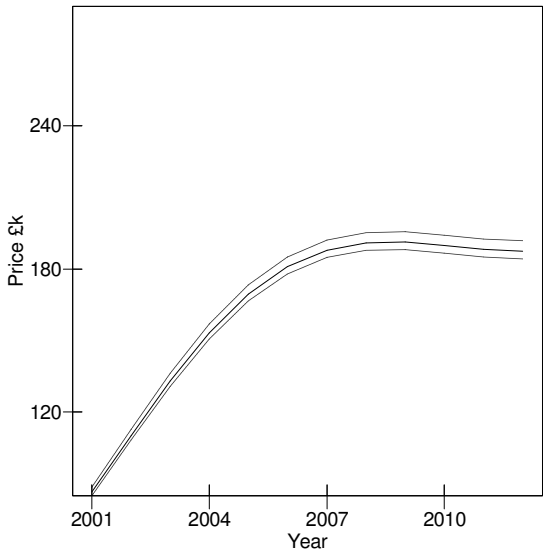

Figure 2 (b)

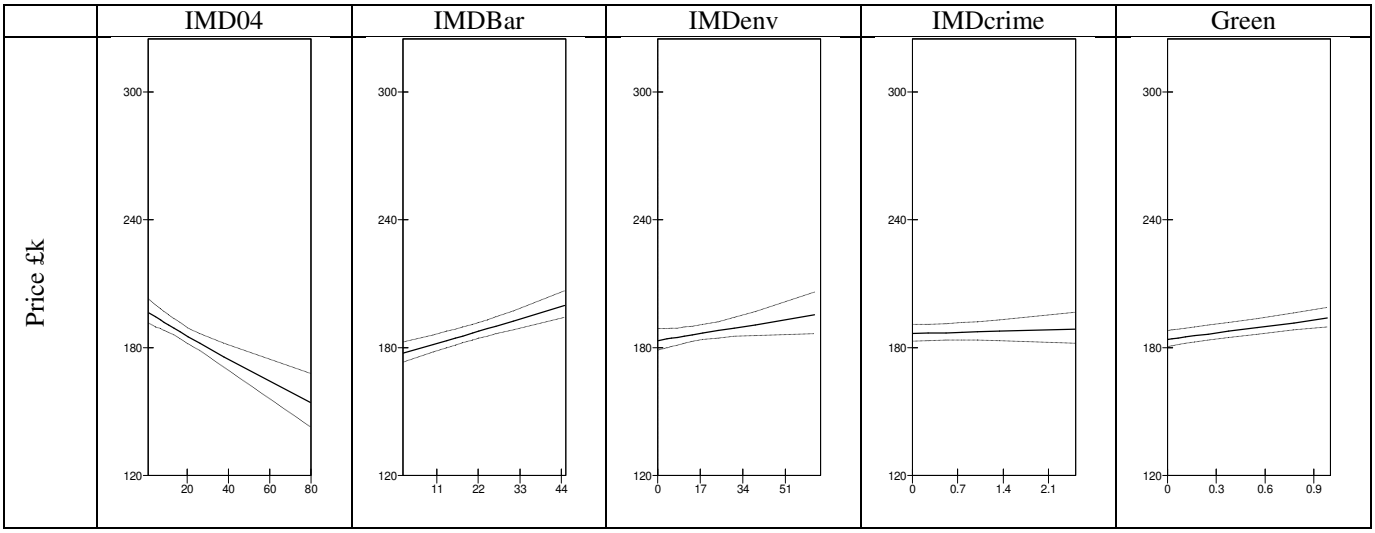

Figure 2 (c)

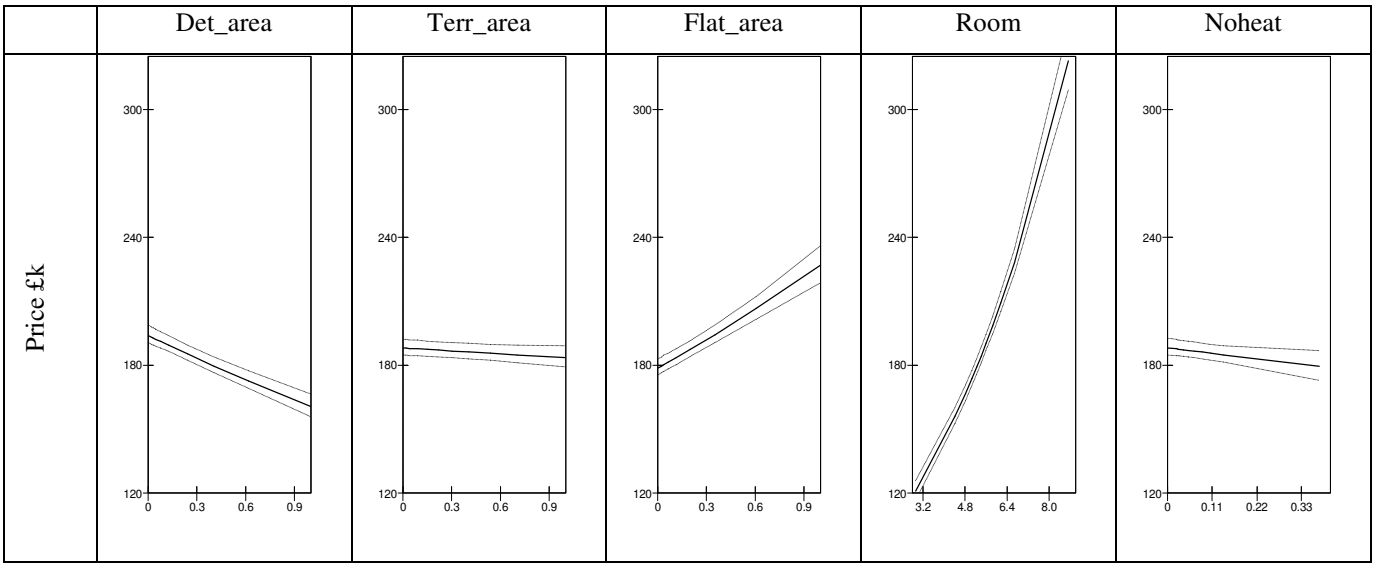

Figure 2 (d)

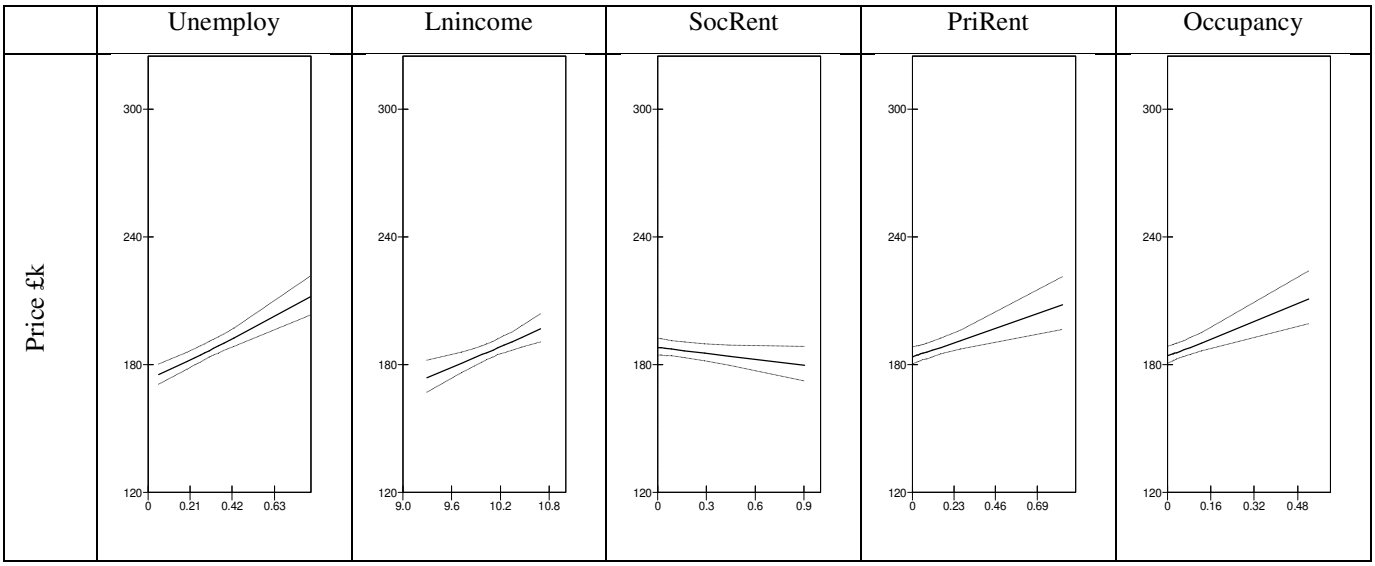

Figure 2 (e)

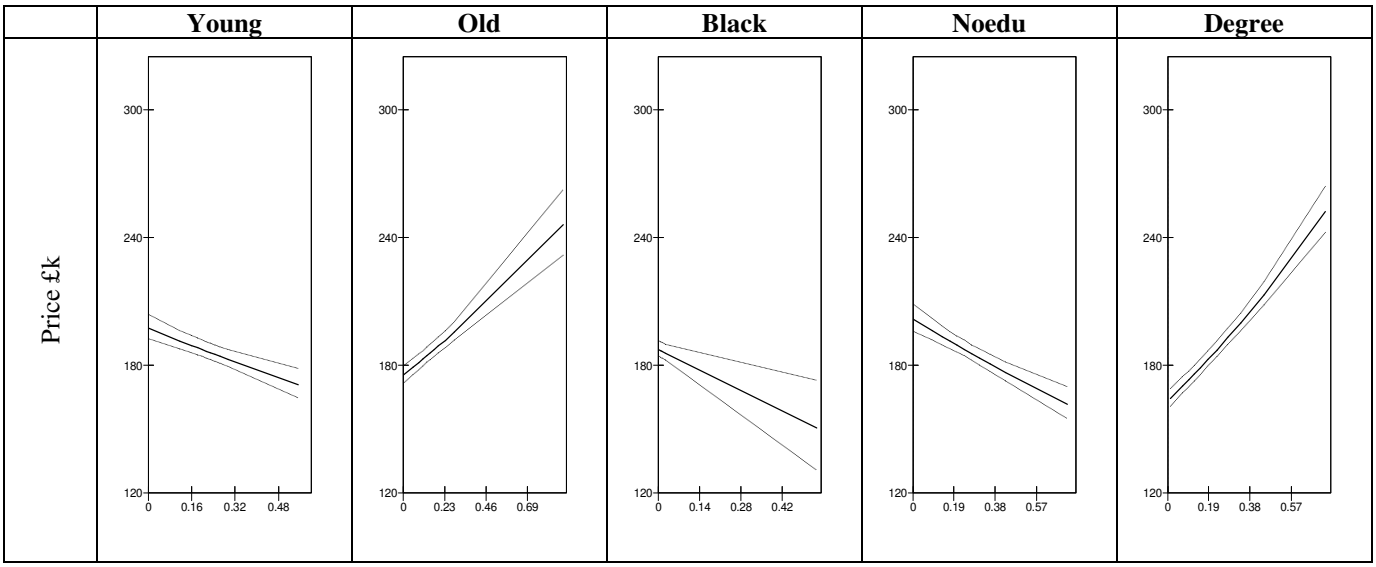

Figure 2 (f)

Table 5 summarised the size of the effect for the range from the minimum to the maximum for each neighbourhood variable and the 16 house types in the order of their importance in MLM3. There are some similarities between MLMs and ANN3, but also some major differences for some predictors such as the percentage of old people and the 2004 IMD crime score.

Table 5 Size of the effect in MLM3 versus ANN3

Variable	Min (£k)	Max (£k)	Range (£k)	Level	MLM3 Rank	ANN3 rank
Room	121	323	202	OA	1	1
16 Types	121	311	190	Property	2	4,12,14
Yrmth	87	191	105	Property	3	2
Degree	164	252	88	OA	4	3
Old	175	246	71	OA	5	19
Flat	178	227	48	OA	6	6
IMD04	154	196	42	LSOA	7	5
Noedu	162	202	40	OA	8	9
Black	150	187	37	OA	9	15
Unemploy	175	212	37	OA	10	7
Det_area	160	194	34	OA	11	11
Occupancy	184	211	27	OA	12	16
Young	171	197	26	OA	13	8
PriRent	184	208	24	OA	14	18
LnIncome	174	197	23	MSOA	15	13
IMDbar	177	200	22	LSOA	16	22
IMDenv	183	195	12	LSOA	17	23
Green	184	194	10	OA	18	20
Noheat	179	188	9	OA	19	21
SocRent	179	188	9	OA	20	17
Terr_area	183	188	4	OA	21	24
IMDcrime	184	189	4	LSOA	22	10

5 Conclusions

This paper illustrates the use of MLM and ANN approach to modelling housing prices and compares them with the widely accepted HPM approach in terms of goodness-of-fit, predictive accuracy and explanatory power. Neither ANN nor HPM is capable of including neighbourhood in the model due to the large number of categorical variables required for practical specification, while MLM is able to specify by simply defining them as macro-level units. The results indicate that MLM offers good predictive accuracy with high explanatory power, especially if neighbourhood effects are explored at multiple spatial scales.

References

Caudill, M. (1988). Neural Network Primer: Part III, AI Expert, pp53-59.

Goldstein, H.(1999). *Multilevel Statistical Model*. Arnold.

Lancaster, K.J. (1966), A New Approach to Consumer Theory, *Journal of Political Economy,* 74, pp. 132-157.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economy*, Vol. 82, pp. 34-55.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning Representations By Back-Propagating Errors. *Nature*, 323(6088), pp.533–536.

Snijders, Tom A B and Bosker, Roel J. (1999). *Multi- level Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Thousand Oaks, CA: Sage

Acknowledgements

This work was supported by the Economic and Social Research Council [grant number EDUC.SC3325]. The house price paid data covers the transactions received at Land Registry in the period 1st Jan 2001 to 31st Dec 2013. © Crown copyright 2013

Biography

Yingyu Feng is a PhD candidate at the University of Bristol. Her research interests include spatial modelling, multilevel analysis, GIS technology, neural networks and their applications in spatial analysis.

Kelvyn Jones is Professor of Quantitative Human Geography at the University of Bristol. He is an Academician of the Social Sciences and featured in the top 20 most cited human geographers of the last half century as of 2009. In 2013 he was awarded the Murchison Award of the Royal Geographical Society for 'publications on quantitative geography'.