ADVANCES IN PROPERTY PRICE MODELING

Taher Buyong
Spatial and Numerical Modeling Laboratory
Inastitut of Advanced Technology
Universiti Putra Malaysia
43400 UPM-Serdang
taher@putra.upm.edu.my

ABSTRACT

Current trends in mass appraisal of properties are the use of the MRA model. MRA has severe limitations in that it ignores spatial dependence and spatial heterogeneity; two spatial effects inherent in property data. As a result, residuals of MRA display spatial autocorrelation and heteroscedasticity that violated the presumption of OLS estimation used in MRA and may lead to biased estimated coefficients. Consequently, the accuracy of predicted property prices may be questioned. Several advanced models have emerged that try to overcome the limitations of MRA that are grouped as spatial, local and geostatistical models. These advanced models that incorporate spatial effects are proven to be superior to MRA in predictive performance in many studies and one of them should be suitable to be used in the mass appraisal of properties in the country. This paper demonstrates the problem of MRA and the selection of the most appropriate model, using MRA as the benchmark, to deal with a rent and transaction datasets. Problems in regressing residential property data are also highlighted.

Keywords: Mass appraisal, rating valuation, spatial models, local models, geographically weighted regression, spatial prediction.

1. Introduction

Determining the market price of properties in a region, termed mass appraisal, is an important professional task for valuation professionals. The current market prices of properties are required for purposes such as the basis of property taxes, understanding trends and dynamics of property prices, etc. It is normal that for any region, properties with known market prices make up for only a small percentage of properties in the region; there are very few properties that were recently transacted (sold) that their market prices are known. The market prices of the majority of the properties that have never been transacted (unsold) are unknown and must be specially determined.

There are several ways in which the prices of unsold properties can be determined in the context of mass appraisal. In the past, the sales comparison method that was designed for single property valuation was used. Considering that the number of properties to be valued may reach several hundred thousands, the method was inefficient. Currently, there is a move towards utilizing statistical-based method and the Multiple Regression Analysis (MRA) model (Adair and McGreal, 1988; Benjamin, Guttery, and Sirmans, 2004) is almost the automatic choice for all users. It should be aware that MRA has limitations in modeling property prices and other statistical-based methods that overcome the limitations in MRA have emerged.

The limitations of MRA are due to spatial effects in property data, primarily spatial dependence and spatial heterogeneity (Anselin, 1988). Advances in statistical-based method have produced models that are grouped as spatial models (Anselin, 1988), local models (Fotheringham, Brunsdon, and Chalton, 2002) and geostatistical kriging models (Dubin, 1998) that try to overcome the limitations of MRA. Spatial models focus on spatial dependence, local models focus on spatial heterogeneity while geostatistical models exploit spatial dependence. Studies proved that they generally performed better than the MRA based on a dataset that each model deals with (Hernandez, Yeates, and Lea, 2003; Paez, Long, and Farber, 2008; Bitter, Mulligan, and Dall'erba, 2007; McCluskey et al., 2000; Chica-Olmo, 2007). There is no study that has empirically analyzed all potential models and selected the most appropriate model for a dataset; Case et al. (2004) and Farber (2004) are probably the closest studies. Since data generating processes are not unique and property data behaves differently in different regions, potential models should be investigated of their suitability in dealing with the dataset at hand.

This paper aims to demonstrate that MRA is not suitable for the modeling of property prices when spatial effects are present and the most appropriate model from several advanced models to deal with the dataset must be

selected based on a series of assessments including prediction accuracy. The composition of the paper is as follows. Section two discusses major issues in property price modeling arising from spatial effects. Section three describes the advances in regression analysis and the issues they addressed. Section four presents the procedure for selecting the most appropriate model for the dataset at hand while Section five provides examples of model selection based on estimation diagnostics and prediction accuracies for a sample rent and transaction datasets. Section six presents related discussions and Section seven concludes the paper by highlighting the important points.

2. Major Issues

There are several issues related to statistical-based property price modeling but our focus is on issues arising from spatial effects of property data (Long, Paez and Farber, 2007). Two spatial effects are recognized for property data, that is, spatial dependence and spatial heterogeneity (Can and Megbolugbe, 1997; Kestens, Theriault, and Rosiers, 2006; Zhang, Ma, and Guo, 2009) and, the consequences due to them are spatially correlated residuals and heteroscedasticity. The first two sub-sections discuss spatial effects and the third sub-section discusses the effect of spatially correlated residuals and heteroscedasticity on predicted prices.

2.1 Spatial dependence

Spatial dependence, also known as spatial autocorrelation, is the existence of a functional relationship between a property and other properties (Anselin, 1988; LeSage and Pace, 2009; Pace, LeSage and Zhu, 2009). There are several ways in which spatial dependence can exist in property price modeling.

Spatial dependence follow from the theory that space is an important element in structuring objects and human behavior that give rise to a variety of interdependencies in space. As a result, what happen at a location is determined in part by what happen at other nearby locations. We say that an object such as a property is related to every other properties and the strength of the relationship decays with increasing distances between the properties such that beyond certain distance, the relationship is negligible. This form of spatial dependence occurs among property prices. Houses priced below RM40,000 are clustered together and distant away from clusters of houses priced above RM500,000, and a seller (and a buyer) takes the prices of recently transacted neighboring houses as clues when transacting a house. This form of spatial dependence also occurs among some property characteristics. Houses with two bedrooms and less than 50 sq m of floor areas are clustered together and distant away from clusters of houses with six bedrooms and more than 400 sq m of floor areas.

Spatial dependence can also follows from the theory of errors in that it arises due to spatial interactions between residuals because of omitted/unobserved, mis-specified or incorrectly measured property characteristics that have spatial pattern. This form of spatial dependence exists when distance to city center is omitted from a model when it is a determinant of house price; distance to city center can create pattern of different house prices across the study region. Another example is using data of average household income from national census as a neighborhood independent variable; boundaries of census tract often do not coincide with boundaries of independent variable neighborhood.

Spatial models account spatial dependences by parametrizing the effects in their functional models.

2.2 Spatial Heterogeneity

Spatial heterogeneity, also called spatial non-stationarity or spatial variability, is the lack of stability in the relationship under consideration in a region of interest (Fotheringham, Brunsdon, and Charlton 2002; Bitter et al., 2007). For example, house prices are known to be directly proportional to distance to city centers; house prices tend to decline as distances to city centers increase. However, houses located at suburbs with excellent commercial centers may be able to fetch better prices; apart from obtaining the same level of services from commercial centers, suburb houses enjoy cleaner environment. Thus, the relationship between house price and distance to city center do not necessarily show a constant trend but may change across a region. The relationship may start with a negative relationship near to the city center (decline in price as distance to city center increases) and change to positive at the suburbs. The relationship may again change to negative as it reaches country sides. There are many other situations where relationships among property characteristics show spatial variability; the relationships are not homogeneous over a region.

Two categories of spatial heterogeneity are recognized (Fotheringham, Brunsdon, and Charlton 2002); (i) discrete spatial heterogeneity, and (ii) continuous spatial heterogeneity. Discrete spatial heterogeneity occurs when the different relationships follow a spatial structure such as between different sub-regions or sub-markets (urban, sub-urban and rural). Discrete spatial heterogeneity is taken into account in the form of separate models for the different sub-regions. Continuous spatial heterogeneity occurs when different relationships vary smoothly across a region; the relationships do not necessarily follow a spatial structure. This kind of spatial heterogeneity is a bit difficult to account for and is usually modeled by randomly breaking up the region concerned into smaller sub-areas.

2.3 Neglecting Spatial Effects

The MRA functional model relating property prices and characteristics, in matrix form, is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{y}$ is the vector of property prices; $\mathbf{X}$ is the matrix of property characteristics with ones in the first column; $\mathbf{b}$ is the vector of relationships (coefficients); and $\boldsymbol{\varepsilon}$ is the vector of errors (residuals). MRA is a global model in that one model is fitted to a region of interest (Figure 1).

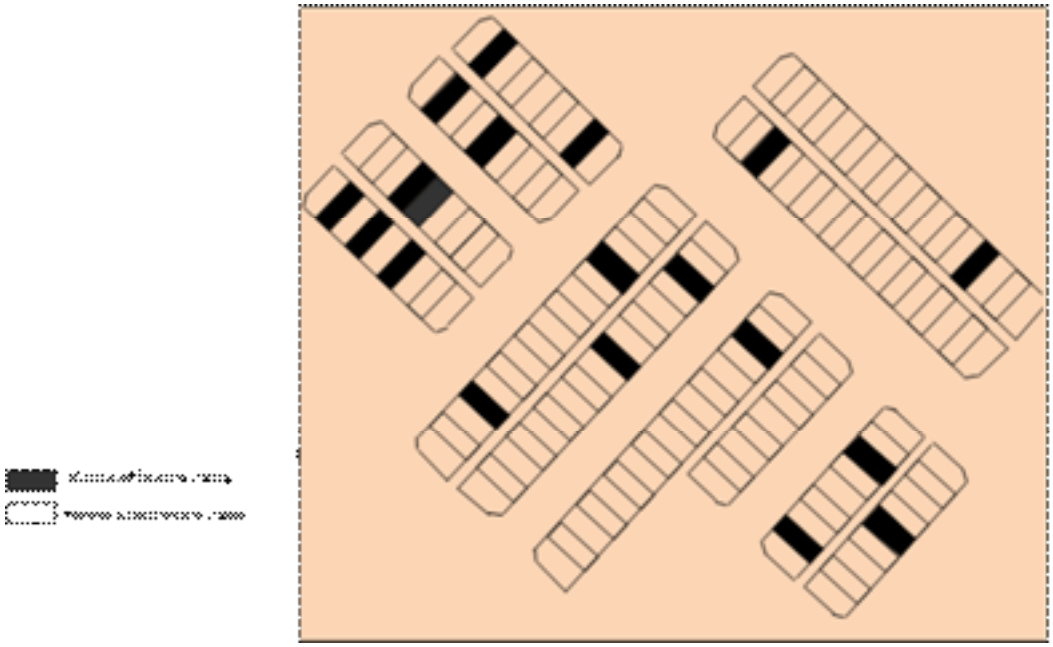

Figure 1 Global model fits one model to a region

The model neglects spatial effects totally; we have discussed in Sections 2.1 and 2.2 that property data are subjected to spatial dependence and spatial heterogeneity. Disregarding these effects lead to model misspecification and disturbs the residuals; the distribution and the structure of the variance-covariance matrix (Fotheringham, Brunsdon, and Charlton 2002; LeSage and Pace 2009). The expectation of the residuals is no longer zero $(E[\boldsymbol{\varepsilon}] \neq 0)$, the variance-covariance matrix of the residuals is a non-diagonal matrix signifying spatially correlated residuals (spatial autocorrelation) with non-constant diagonal elements (heteroscedasticity). Spatial autocorrelation and heteroscedasticity violate the presumption of the least squares estimation that the residuals must be uncorrelated and normally distributed with zero mean and constant variance, i.e., $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$. This makes the ordinary least squares estimated parameters employed by MRA biased and unsuitable for inference. The ending effect is that the predicted property prices are inaccurate and unreliable.

3. Current Trends/Advances

Several models that improve on the MRA have emerged. They can be categorized as spatial, local and geostatistical models.

3.1 Spatial Models

Spatial models are also known as spatial autoregressive models. The models improve on the MRA model by including spatial dependence parameters in the functional model; spatial dependence is explicitly modeled and estimated along with the coefficients. Four types of spatial models are recognized depending on where the modeling of spatial dependence occurs (Anselin 1988; LeSage and Pace, 2009): (i) Spatial Lag Model (SLM), (ii) Spatial Error Model (SEM), (iii) General Spatial Model (GSM), and (iv) Spatial Durbin Model (SDM).

The functional model of the SLM that models spatial dependence in the price is expressed as,

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\mathbf{b} + \mathbf{e} \tag{2}$$

The functional model of the SEM that models spatial dependence in the residuals is expressed as,

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{u}; \quad \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \mathbf{e} \tag{3}$$

The functional model of the GSM that models spatial dependence in both the prices and residuals is expressed as,

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}\mathbf{b} + \mathbf{u}; \quad \mathbf{u} = \lambda\mathbf{W}_2\mathbf{u} + \mathbf{e} \tag{4}$$

The functional model of the SDM that models spatial dependence in the characteristics is expressed as,

$$\mathbf{y} = \rho\mathbf{W}_1\mathbf{y} + \mathbf{X}\mathbf{b}_1 + \mathbf{W}\mathbf{X}\mathbf{b}_2 + \mathbf{e} \tag{5}$$

The notation **y** is the vector of property prices; ρ and λ are parameters of spatial dependence; **W** is the spatial weight matrix specifying the inter dependence of properties; **X** is the matrix of property characteristics with ones in the first column; **b** is the vector of coefficients; and **e** is the vector of residuals. From the perspective of model specifications, spatial models are more complete with the addition of spatial dependence parameters in the functional model.

The spatial weight matrix, **W**, defines the properties that the prices, characteristics, residuals, or any combination of these are interdependent. The interdependence is based on the concept of spatial neighbors. The spatially dependent properties receive a value of one while other properties receive the value of zero in the weight matrix. The property in question which

is the diagonal element receives the value of zero to prevent the property from influencing itself. Thus, the weight matrix **W** for a network of property centroids in Figure 2 is shown in Figure 3.

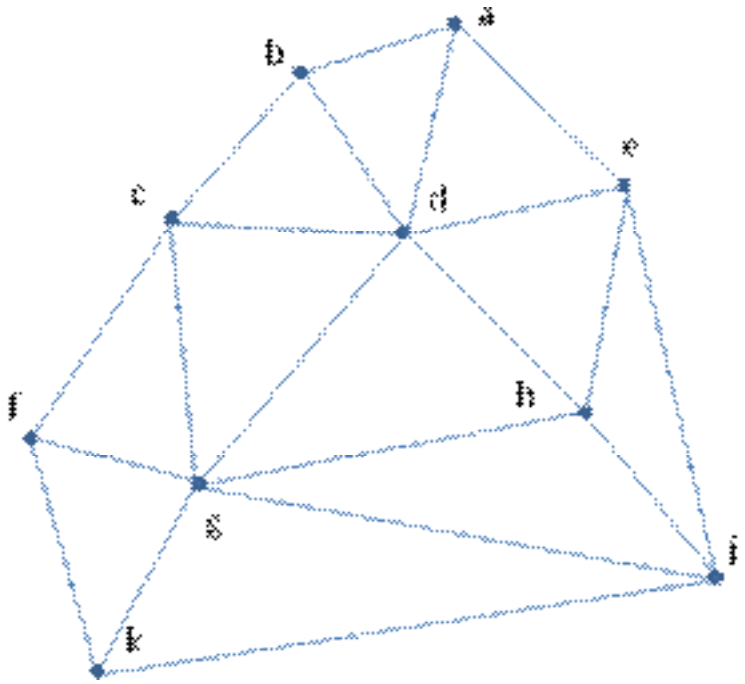

Figure 2 Connected property centroids are spatial neighbors

$$W = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ j \\ k \end{array} \begin{array}{c} \begin{array}{cccccccccc} a & b & c & d & e & f & g & h & j & k \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

Figure 3 Weight matrix **W** following spatial neighbors of Figure 2

Spatial models are global models like MRA. The only difference is that spatial dependence of properties is specified in the models (Figure 4). Estimating spatial models are carried out using the method of maximum likelihood where the spatial dependent parameters ρ and λ are estimated along with the coefficients (Anselin, 1988; LeSage and Pace, 2009). Accounting spatial dependence reduces, if not eliminates, spatial

autocorrelation and heteroscedasticity. Spatial models, however, ignore spatial heterogeneity which is the other component of spatial effects.

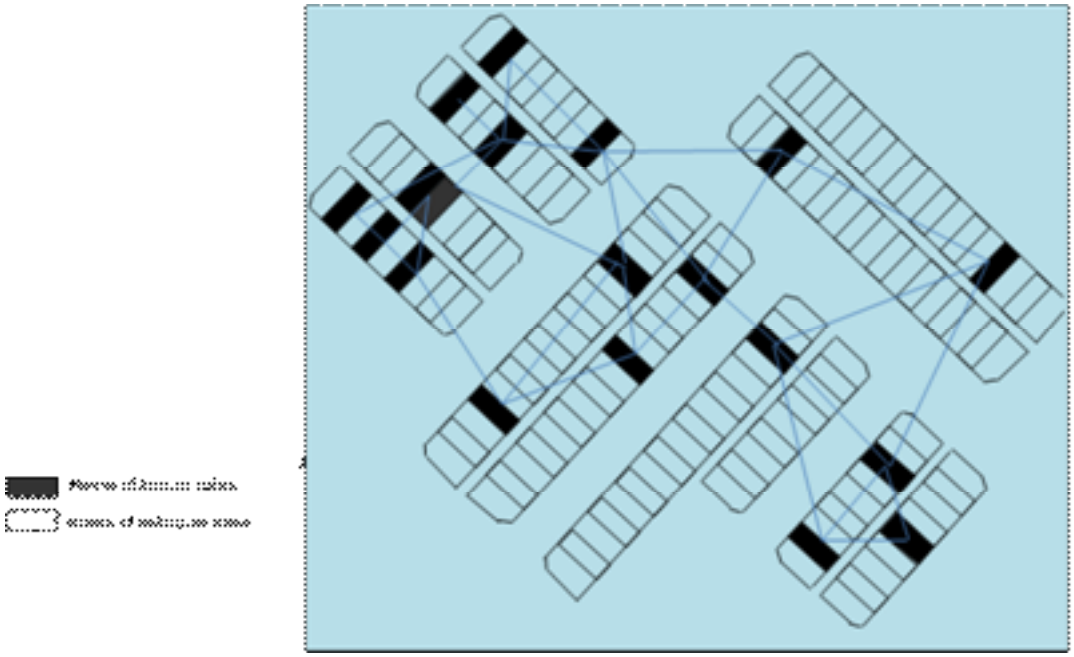

Figure 4 Spatial models are global models but with specified spatial dependence

3.2 Local Models

Two types of local models exist; Geographically Weighted Regression (GWR) and Moving Window Regression (MWR). Both local models consider several local areas (windows) for a region of interest and fit the MRA model to each of the windows. The regression points (centers of the windows) are at properties with known prices and the windows are allowed to overlapped. At each regression window, only a subset of observations nearest to the regression point enters the regression and all other observations are ignored. Figure 5 shows the concept of local modeling.

The functional model of local models in matrix form is given as,

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \mathbf{e}_i; \quad i = 1, 2, 3, \ldots, k \tag{6}$$

where **y** is the vector of prices; **X** is the matrix of property characteristics with ones in the first column; **b** is the vector of coefficients; and **e** is the vector of residuals. The notation i is the i^{th} regression window; and k is the total number of windows.

Property prices in a regression window are weighted differently in GWR and MWR. In MWR, the prices are weighted equally resulting in an identity weight matrix. In GWR, several weighting strategies are available and they utilize some distance decay functions resulting in a non-identity weight

matrix. For the same regression window, the weight matrix of GWR and MWR differ only in the values of the diagonal elements. Due to the structure of weight matrix, MWR estimation is done via ordinary least squares while GWR estimation is done via weighted least squares.

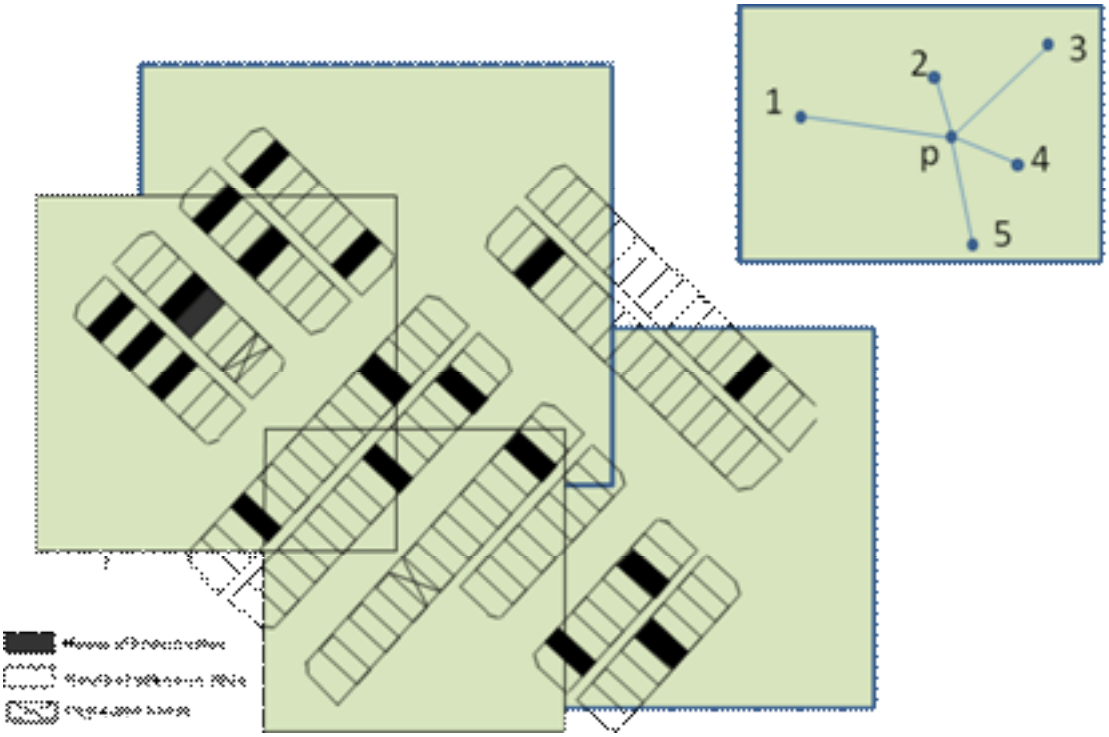

Figure 5 Local models fit a model to several local areas in a region

Local models acknowledge spatial heterogeneity and regressing in smaller windows across a region of interest permit this effect to be captured. Each regression window produces a set of estimated coefficients and variation in their values between different windows, which signifies spatial heterogeneity in the region, can be evaluated (Leung, Mei, and Zhang, 2000). By considering a subset of observations nearest to regression points, local models also acknowledge spatial dependence. However, GWR and MWR take spatial dependence in two different ways. Spatial dependence of an observation in GWR is a function of the distance from the observation to a regression point; it is at its maximum value when the observation is at the regression point and diminishes to a smaller value when the observation is distant away from the regression point. Spatial dependence of an observation in MWR is taken to be a constant of maximum value irrespective of its distance to the regression point. Spatial dependence of local models is reflected in their weight matrices.

3.3 Geostatistical Models

Geostatistical models, also called kriging models, use spatial dependence of prices of sold properties to predict the prices of unsold properties in a region of interest. This is possible by assuming a stochastic view of space and property price is a continuous random variable over the region. It is random because the factors that determine the price of property are numerous, largely unknown in detail and interact with complexity. However, the random data generating process of property price is spatially dependence meaning that the price of a property is dependent on the prices of nearby properties. This is very valid because random processes at two nearby locations are alike.

The methods begin with analysis of spatial dependence of property prices and constructing an experimental semivariogram describing the overall spatial dependence of the prices over a region of interest. Even though the semivariogram may show spatial dependence over the maximum separation distance of properties in the region, meaningful spatial dependence is limited to very small separation distances, called range (Figure 6). A theoretical semivariogram model is fitted to the determined experimental semivariogram. The fitted semivariogram model that represents the model of spatial dependence of property prices in the region is then used in the prediction process.

Figure 6 Meaningful spatial dependence of houses can only be obtained for range up to 750m

Several kriging models are available such as simple, ordinary, universal and so on, that can be used in the prediction and the general kriging model is given as,

$$\hat{z}(p_0) = \sum_{i=1}^{k} \lambda_i z(p_i) \quad (7)$$

where $\hat{z}(p_0)$ is the predicted price of property p_0; $z(p_i)$ is the price of property p_i; λ_i is the weight of the price of property p_i; and k is the number of properties in the neighborhood used in the prediction. The weights are

obtained after solving the related kriging systems of equations where spatial dependence of prices represented by the semivariogram model plays major roles.

Geostatistical kriging models prescribe a different operational procedure in relation to MRA, spatial models and local models discussed before due to different underlying theory. The models exploit spatial dependence in the effort to predict property prices instead of modelling it as in other advanced models. When only a limited number of neighbourhood properties are considered in the prediction, the models implicitly account spatial heterogeneity.

4. Model Selection

No single model is good for all datasets because data generating processes are different at different regions. Some datasets are dominated with spatial dependence that spatial models may be relevant. Some datasets are dominated with spatial heterogeneity that local models are relevant. Some datasets have no spatial effects (although very rare) that MRA model is relevant. The most appropriate model to deal with a dataset must be empirically determined.

In the absence of prediction, model selection has to be based on regression diagnostics such as the Moran's *I* spatial autocorrelation coefficient, Jarque-Bera statistic, Breusch-Pagan statistic, significance of spatial dependence parameters (ρ and λ), and significance of the variation of estimated coefficients. There is no hard-and-fast rule of how these parameters should be assessed but there are guidelines that can be followed (LeSage, 1999; Anselin, 2005). In property price modeling, it is always possible to predict the prices of sold properties such that the accuracy of prediction generated by the different models (predicted prices vs observed market prices) can be assessed. In such cases, the accuracy of prediction measures must also be considered in the model selection process and more weight should be given to it. A model that has accounted all the factors that could have led to inferior results should be able to predict more accurately.

5. Examples - Rents and Transactions

Two datasets are used to illustrate model selection, i.e., a rent and transaction datasets, to reflect the use of the two types of data in mass appraisal for property tax in the country.

Rent dataset

The first dataset comprise 1399 rent records of houses sampled in 2008 in two adjacent townships of Bandar Baru Bangi and Bandar Bukit Mahkota within Majlis Perbandaran Kajang. The specified functional model is

$$Price = b_0 + b_1 LA + b_2 MFA + b_3 AFA + b_4 TYP + b_5 POS + b_6 CEI + b_7 AGE + b_8 CBD + b_9 NQ$$

The dataset was randomly divided into two parts called calibration and prediction datasets comprising 90% and 10% of rent records, respectively. The calibration dataset was for calibration or estimation while the prediction dataset was for testing out-sample prediction. All models were estimated using the calibration dataset and the diagnostics analyzed to ascertain treating the dataset with spatial and local models (Table 1). All spatial and local models fit the dataset extremely well outperforming the MRA model with the exception of the SLM judging from the adjusted R^2 values. The values of Moran's *I* indicate that the residuals are still spatially correlated; the spatial regressions have not completely eliminated the spatial autocorrelation. The SLM, SEM and GSM have managed to model spatial dependences, as indicated by the values of ρ and λ parameters. Non-normality of residuals of GWR is less than that of MRA judging from the Jarque-Bera value although spatial models are better than GWR. Heteroscedasticity of GWR is less compared to MRA judging from the value of Breusch-Pagen. All variables show significant spatial heterogeneity across the study region (Table 2).

Table 1 Results of calibrating all models using KJ-Rent dataset

Diagnostics	SLM	SEM	GSM	GWR	MWR	MRA
Adj R^2	0.8057	0.9381	0.9434	0.9708	0.9730	0.8311
Moran's *I* (Prob)	0.1937 (0.0000)	0.2602 (0.0000)	0.0486 (0.0000)	0.1487 (0.0000)	0.1854 (0.0000)	0.1655 (0.0000)
ρ (Prob)	0.5800 (0.0000)	-	0.2690 (0.0000)	-	-	-
λ (Prob)	-	0.8590 (0.0000)	0.5920 (0.0000)	-	-	-
Jarque-Bera (Prob)	938 (0.0000)	499 (0.0000)	4419 (0.0000)	1013 (0.0000)	1283 (0.0000)	1022 (0.0000)
Breusch-Pagen (Prob)	1601 (0.0000)	1562 (0.0000)	1494 (0.0000)	1601 (0.0000)	1561 (0.0000)	1829 (0.0000)

Table 2 Spatial heterogeneity test of KJ-Rent data

Variable	GWR Probability	MWR Probability
LA	0.0000	0.0000
MFA	0.0000	0.0000
AFA	0.0000	0.0000
POS	0.0000	0.0000
TYP	0.0000	0.0000
CEI	0.0000	0.0000
AGE	0.0000	0.0000
CBD	0.0000	0.0000
NQ	0.0000	0.0000

Geostatistical analysis of the KJ-Rent dataset was also carried out where Figure 7 shows the experimental semivariogram of the rent price variable. It can be seen spatial autocorrelation exists in the dataset where the range is about 800 m. This range value seems reasonable for house price variable.

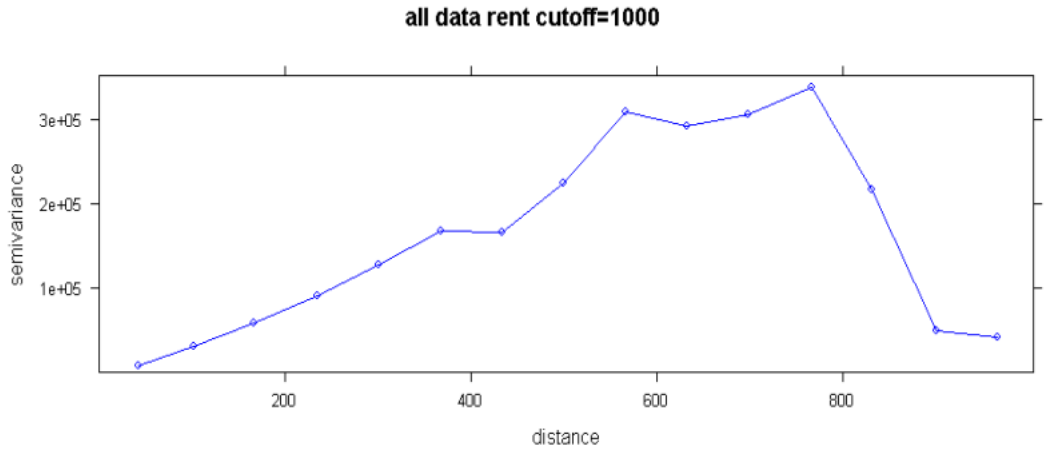

Figure 7 Experimental semivariogram of rent price of KJ-Rent dataset

The estimated coefficients of the respective models were then used to predict the rent prices of houses in the prediction dataset and the predicted prices were compared with their observed prices using some prediction accuracy assessment measures (Table 3). For median ratio measure, all investigated models meet the standard and perform better than MRA, except SLM, with GSM and MWR perform the best. For coefficient of dispersion measure, only GWR meet the standard and perform better than MRA. For price related differential measure, only SEM and GWR meet the standard with GWR performs the best. For percentage of prediction less than 10 percent error measure, all models meet the standard, except GSM, with GWR performs the best. However, SLM does not outperform MRA. It is concluded that GWR performs the best with the KJ-Rent dataset for meeting the standards and outperforming the MRA in all measures, and performs the best in three measures.

Table 3 Prediction accuracy assessment of KJ-Rent dataset

	R_{med}	**COD**	**PRD**	**P<10%E**	**RMSE**
Standard	0.9-1.1	<10	0.98-1.03	>50%	-
MRA	*0.97*	13.47	*1.03*	*53.57*	157743
SLM	1.60	30.83	1.21	*52.85*	152736
SEM	*0.97*	13.47	*1.03*	*53.57*	159064
GSM	***1.00***	14.59	1.07	45.00	150630
GWR	*0.97*	***7.20***	***1.01***	***75.71***	**71809**
MWR	***1.00***	18.06	0.92	*72.85*	73949
Kriging	1.03	12.26	*1.03*	*50.34*	93458

Note: Underlined indicates outperforming the MRA model; Italic indicates meeting the standard; Bold indicates the best performance in the measure

Transaction Dataset

The second dataset comprises 463 houses transacted in 2008 within Majlis Bandaraya Johor Bahru. The specified functional model is

$$Price = b_0 + b_1 LA + b_2 MFA + b_3 AFA + b_4 STO + b_5 BED + b_6 TIT + b_7 AGE$$

The dataset was also randomly divided into calibration and prediction datasets as in the rent dataset. All models were estimated using the calibration dataset and the diagnostics analyzed to ascertain treating the dataset with spatial and local models (Table 4). All models fit the dataset very well outperforming the MRA model judging from the adjusted R^2 values. The Moran's I values indicate that small spatial autocorrelation still present in the residuals. The Moran's I for MWR shows no value but it is statistically insignificant. The SLM, SEM and GSM managed to model spatial dependences that range from 0.27 to 0.71 and all are significant. The Jarque-Bera tests show significant non-normal distribution of the residuals like the MRA model. The Breusch-Pagan tests show significant heteroscedasticity like the MRA with GWR being the least severe. This is further reinforced by the non-significant variation in the relationships as shown by the spatial heterogeneity test (Table 5). The results of the spatial and local models diagnostics are almost in agreement with the MRA model that there are no spatial effects in the dataset.

Table 4 Results of the calibration of models using JB08-All dataset

Diagnostics	SLM	SEM	GSM	GWR	MWR	MRA
Adj R^2	0.73	0.82	0.83	0.99	0.94	0.72
Moran's *I* (Prob)	0.13 (0.00)	0.18 (0.00)	0.03 (0.01)	0.03 (0.01)	0.00 (0.73)	0.12 (0.00)
ρ (Prob)	0.42 (0.00)	-	0.27 (0.00)	-	-	-
λ (Prob)	-	0.71 (0.00)	0.36 (0.00)	-	-	-
Jarque-Bera (Prob)	740.80 (0.00)	2070.84 (0.00)	2151.81 (0.00)	59284.92 (0.00)	11637.80 (0.00)	2310.78 (0.00)
Breusch-Pagen (Prob)	601.52 (0.00)	992.31 (0.00)	1008.05 (0.00)	521.89 (0.00)	679.29 (0.00)	681.54 (0.00)

Table 5 Spatial heterogeneity test of JB08-All dataset

Variable	GWR Probability	MWR Probability
LA	0.1800	0.2000
MFA	0.3300	0.4500
AFA	0.7100	07800
TIT	0.6400	0.7400
BED	0.0000	0.0000

STO	0.0000	0.0000
AGE	0.8400	0.8400

Geostatistical analysis of JB08-All dataset was carried out where Figure 8 shows the experimental semivariogram of the transacted price variable being fitted with a semivariogram model. It can be seen that the range of spatial autocorrelation is about 900 meters.

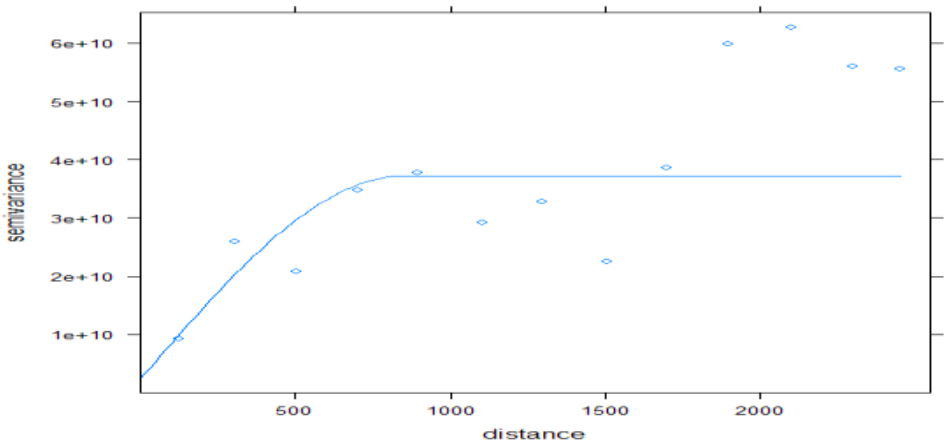

Figure 8 Fitted semivariogram model of JB08-All dataset

Table 6 shows the results of the prediction accuracy assessment of JB08-All dataset. For median ratio measure, all investigated models meet the standard and all models perform better than MRA except SEM with GWR, MWR and kriging being equally the most accurate. For coefficient of dispersion measure, only GWR meet the standard and predicts the most accurate even though GSM, GWR, MWR and kriging perform better than MRA. For price related differential measure, only GWR, MWR and kriging meet the standard, perform better than MRA with GWR predicts the most accurate. For percentage of prediction less than 10 percent error measure, only GWR, MWR and kriging meet the standard even though GSM, GWR, MWR and kriging perform better than MRA and, GWR predicts the most accurate. For root means square error measure, all models except SEM perform better than MRA with GWR predicts the most accurate. It is clear that GWR is the model that can predict most accurate with JB08-All dataset with MWR and kriging models as the alternatives.

Table 6 Prediction accuracy assessment of JB08-All dataset

	R_{med}	**COD**	**PRD**	**P<10%E**	**RMSE**
Standard	0.9-1.1	<10	0.98-1.03	>50%	-
MRA	1.04	29.59	1.13	30.02	133350.55
SLM	*1.01*	34.32	1.07	22.03	132555.69

SEM	1.06	33.80	1.18	21.38	142041.65
GSM	1.01	27.38	1.08	31.53	105623.50
GWR	***1.0***	***4.79***	***1.01***	***85.53***	**20442.51**
MWR	***1.0***	10.08	*1.02*	*65.23*	60487.14
Kriging	***1.0***	11.65	1.08	*83.87*	83807.04

Note: Underlined indicates outperforming the MRA model; Italic indicates meeting the standard; Bold indicates the best performance in the measure

6. Analyses and Discussions

The procedure for selecting the most appropriate model was concisely described in Section four and numerical examples of rent and transaction datasets were presented in Section five.

From the examples, MWR loses out to GWR in both datasets due to the structure of its weight matrix in regression windows. The weights of prices (rent and transaction) in MWR windows are non-varying which do not really represent the data generating process. The prices of nearby houses have more influence than prices of distant houses for any subject house. The distance decay weight structure of GWR windows is much closer to the data generating process. Spatial models lose out to GWR because spatial heterogeneity is more dominant compared to spatial dependence in the dataset. This fact is true because of significant difference in many factors between the two townships that the rent samples were taken: Bandar Baru Bangi is much closer to Kuala Lumpur and Putra Jaya, the main employment and business centers. Bandar Baru Bangi is more developed with varied facilities and services compared to Bandar Bukit Mahkota. Spatial dependence which measure 0.3097 on Moran's I scale is probably dominated by spatial heterogeneity that is highly significant for all house characteristics (Table 5). The same arguments are applicable to transaction samples that represent neighborhoods (*taman*) of varying development, environments and characteristic within Johor Baru. Spatial heterogeneity test of both datasets proves the arguments.

The analysis of rent dataset for model selection is very consistent throughout the entire process that there is high probability of the existence of spatial effects in the dataset and alternative models may be required. The assessment of prediction accuracy confirmed this statement and GWR is selected. The model selection process of transaction dataset is not as straightforward as the rent dataset. Model estimation diagnostics suggest non-existence of spatial effects meaning that MRA is appropriate to deal with the dataset. However, the prediction accuracy assessment reveals that MRA is not the appropriate model and GWR is selected.

These results, however, does not mean that GWR is the model to be used for every region in the country. Spatial and local models are probable only when spatial dependence and heterogeneity are significant in a dataset at

hand as usually indicated by a high Moran's I measure and reinforce with prediction accuracy assessment. Otherwise, MRA model can do the job; MRA which is less complicated and more familiar to users should be used when the dataset displays no spatial effects. We have seen that low Moran's I measure of the transaction dataset may also result in low prediction performance of MRA (transaction dataset) because Moran's I coefficient cannot exhaustively account for spatial effects The exact spatial or local model to be used must be found by performing the estimation diagnostics relating to spatial dependence and spatial heterogeneity as well as prediction accuracy assessment. However, more weight should be given to prediction accuracy assessment because it represents the goal of the property price modeling. The fact that there is no guarantee that the diagnostics exhaustively account for two spatial effects and the possibility of existence of other type of spatial effects warrant more weight be given to the prediction accuracy assessment.

It is proper to mention here that three major problems currently exist in using transaction records in property price modeling: (i) geocoding of transacted properties, (ii) incorrect model specification, and (iii) actual Data Generating Process (DGP) that does not support hypothesized DGP. The first issue is distinct while the second and third issues are interrelated.

It has been a standard practice in Malaysia that geocoding is facilitated by cadastral maps or topographic maps produced by the Department of Survey and Mapping, Malaysia. For geocoding of transacted properties, cadastral maps are more appropriate. However, cadastral maps produced prior to 2010 do not contain properties with qualified titles. Only cadastral maps produced in 2010 and beyond, after the implementation of the e-cadastre, contain properties with qualified titles. With the amendment of the National Land Code that permitted qualified titled properties to be transacted, many qualified titled properties are being transacted and geocoding of these properties using cadastral maps break down. For example, about 6,000 landed residential properties were transacted in Johor Bahru in 2008 but only 463 transactions managed to be geocoded; about 10,000 landed residential properties were transacted in Hulu Langat in 2009 but only 735 transactions managed to be geocoded. Based on the experience in dealing with transactions for the last 10 years in Hulu Langat and Johor Bahru, about 90% of transacted properties cannot be geocoded. This means the locations of these properties cannot be analytically quantified. Advanced property price modeling cannot be used and GIS that is supposed to facilitate many property valuation and management functions cannot be utilized. This problem will continue forever because these qualified titled properties will continue to exist and are transacted, unless appropriate actions are taken.

It has been hypothesized that basic physical characteristics of properties such as land area, main floor area, ancillary floor area, number of storey and number of bedroom have positive relationships and age has negative

relationship to price, linear or otherwise, following the theory of spatial process. This has been proven in many studies (McCluskey et al., 2000; Hernández, Yeates and Lea, 2003; Bitter, Mulligan, and Dall'erba, 2007; Paez, Long and Farber, 2008). The specified model which may include other variables is applied to any region of interest. Unfortunately, the DGP of property data in Malaysia, transaction records in particular, does not support the hypothesized spatial processes. This is evidenced from the regression analysis of many sets of transaction records of landed residential properties in Johor Bahru and Kajang.

Tables 7 and 8 show the independent variables that are significant with p-value indicated in parenthesis for JB08 and KJ09 dataset. The blank cells indicate non-significant independent variables. In Table 7, among land area, main floor area and ancillary floor area, two storey bungalows have only land area with positive significant relationship at RM785 per square meter (last row). One storey bungalows have land area and ancillary floor area with positive significant relationship at RM160 per square meter and RM1387 per square meter, respectively. Judging from the p-values, these two values are not very reliable; they have large variances. Comparing the land area relationship of RM267 per square meter for one story linked and RM362 per square meter for one storey low cost link one wonders what makes land of low cost houses more expansive. There is large difference between land price of single and double storey semi-detached houses. There exist inconsistency of significant of estimated independent variables and thus their large variances. All these show our transaction records do not support the hypothesized DGP and with the adopted model specification. The same conclusions are drawn when analyzing Table 8 for transactions in Kajang in 2009.

Table 7 Independent variables for JB08 dataset that are of expected signs and more than 95% significance level

Dataset	LA	MFA	AFA	STO	TYP	BED	POS	TIT	AGE
All house types	472 (0.0000)	907 (0.0000)	643 (0.0068)	59345 (0.0003)		31793 (0.0014)		235971 (0.0000)	1281 (0.0245)
1 S Linked	267 (0.0000)	725 (0.0000)	927 (0.0003)						1255 (0.0000)
2 S Linked	275 (0.0055)	1297 (0.0001)							1920 (0.0119)
1S Liked LC	362 (0.0000)		802 (0.0102)			18491 (0.0241)			
2 S Linked LC									
1 S Semi-D	312 (0.0023)	1059 (0.0221)						143718 (0.0038)	2162 (0.0398)
2 S Semi-D	1242 (0.0000)	890 (0.0094)	1114 (0.0014)			57740 (0.0099)			
1 S Detached	160 (0.0151)		1387 (0.0129)					347805 (0.0000)	
2 S Detached	785 (0.0000)							411009 (0.0002)	

Tables 7 and 8 represent examples of the experience gained. In reality, transaction records of landed residential properties for the past ten year of both towns were analyzed and the conclusion made was that the property price model needs re-specification for application in Malaysia. Among the factors that are worth looking at are (i) heterogeneous ethnics of Malaysian population, (ii) certain believes of some ethnic group, (iii) availability of houses, (iii) purchasing power of citizen, (iv) prioritized factors when buying houses of citizen, and (v) segmenting the housing market (Goodman and Thibodeau, 2003; Bourassa, Hoesli and Peng, 2003; Aminah Md Yusof, 2007).

It is fortunate that the JB08-All house types dataset has many significant independent variables that allows the analysis of the hedonic price models to be carried out.

Table 8 Independent variables for KJ09 dataset that are of expected signs and more than 95% significance level

Dataset	LA	MFA	AFA	STO	TYP	BED	POS	TIT	AGE	DIS
All house types	115 (0.0087)	870 (0.0000)	2312 (0.0002)	57328 (0.0007)	78470 (0.0000)	36127 (0.0003)	26274 (0.0096)		-1526 (0.0258)	3010 (0.0026)
1 S Linked							14017 (0.0040)			2267 (0.0001)
2 S Linked	107 (0.0000)	1189 (0.0000)	953 (0.0025)				33327 (0.0000)	-29634 (0.0001)	-1164 (0.0020)	2537 (0.0000)
1 S Linked LC									-999 (0.0277)	
2 S Linked LC										
1 S Semi-D										
2 S Semi-D		2983 (0.0000)						-202509 (0.0013)	-9938 (0.0104)	
1 S Detached						483461 (0.0013)				
2 S Detached								766971 (0.0483)		

7. Conclusions

Spatial effects inherent in property data comprise spatial dependence and spatial heterogeneity. Unaccounting spatial effects in property price modeling results in poor model specification which then leads to spatial autocorrelation and heteroscedasticity. These violated the presumptions of OLS estimation used in MRA and produced biased estimated coefficients. The ending consequence is the reduced accuracy of property price prediction. These are the characteristics of the MRA model which represent the drawback of the model.

Accurate modeling of property prices must account for spatial effects. Spatial models do it by explicit modeling of spatial dependence in the functional models. Spatial models, however, disregard spatial heterogeneity. Local models explicitly account spatial heterogeneity by fitting the MRA model to

several smaller windows until a region of interest in covered. The models implicitly account spatial dependence by only considering property prices in the regression windows. GWR models spatial dependence much more closer to reality than MWR; spatial dependence decreases with distance. Geostatistical kriging models exploit spatial dependence when using semivariograms, instead of accounting it, in predicting property prices and account spatial heterogeneity when only prices near predicted properties are considered. Geostatistical models are applicable only if spatial dependence is significant.

Theories dictate that local models, GWR in particular, that accounts both spatial effects is a superior model, MRA that disregards both spatial effects is an inferior model and the various spatial. However, data generating process is not unique and no model is good for every dataset. The most appropriate model to deal with a dataset must be selected based on a series of assessments. Significant weight should be given on the prediction accuracy assessment because that is the goal of the property price modeling exercise and there is no guarantee that the diagnostics employed exhaustively account for both spaial effects.

Property price model needs re-specification for application in Malaysia. Current model adopted from the west proved inappropriate. Transacted properties must be also able to be geocoded in order to take advantage of advances in property price modeling and other technological tools to enhance property valuation and management.

Acknowledgements

The author gratefully acknowledges the funding of this research from INSPEN under Agreement NAPREC (R&D) No 1/07 and MOSTI under Project No 01-01-04-SF0972.

References

Adair, A., and McGreal, S. (1988). The Application of Multiple Regression Analysis in Property Valuation, *Journal of Property Valuation and Investment*, 6(1): 57-67.

Aminah Md Yusof (2007). Malaysian Housing Investment Information Price Modeling, Final Report NAPREC Research, Institute of National Valuation, Kajang.

Anselin, L. (1988). Spatial Econometrics: Methods and Models, Dordrecht: Kluwer Academic Publisher.

Anselin, L. (2005). *Exploring Spatial Data with GeoDa: A Workbook*, Spatial Analysis Laboratory, Department of Geography, University of Illinois, Urbana-Champaign, USA.

Bitter, C., Mulligan, G. F. and Dall'erba, S. (2007), Incorporating Spatial Variation in Housing Attribute Prices: A Comparison of Geographically Weighted Regression and the Spatial Expansion Method. *Journal Geographical System,* 9: 7-27

Benjamin, J. D., Guttery, R. S., and Sirmans, C. F. (2004). Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation, *Journal of Real Estate Practice and Education*, 7: 65-77.

Bourassa, S. C., Hoesli, M. and Peng V. S. (2003). Do Housing Submarkets Really Matter?, *Journal of Housing Economics*, 12(1):12-28.

Can, A. and Megbolugbe, I. (1997), Spatial Dependence and House Price Index Construction, *Journal of Real Estate Finance and Economics*, 14: 203-222.

Chica-Olmo, J. (2007). Prediction of Housing Location Price by a Multivariate Statistical Method: Cokriging, *Journal of Real Estate Research*, 29(1): 99-114.

Case, B., Clapp, J., Dubin, R., Rodriguez, M. (2004), Modeling Spatial and Temporal House Price Pattern: A Comparison of Four Models, *Journal of Real Estate Finance and Economics*, 29(2): 167-191.

Dubin, R. A. (1998). Spatial Autocorrelation: A Primer, *Journal of Housing Economics*, 7: 304-327.

Farber, S. (2004). A Comparison of Localized Regression Models in an Hedonic House Price Context. Centre for the Study of Commercial Activity, Ryerson University, Toronto, Canada.

Goodman, A. C. and Thibodeau, T. G. (2003). Housing Market Segmentation and Hedonic Prediction Accuracy, *Journal of Housing Economics*, 12(2):181-201.

Hernández, T., Yeates, M., and Lea, T. (2003), *Residential property valuation : an application of geographically weighted regression (GWR)*. Centre for the Study of Commercial Activity, Ryerson University, Toronto, Canada.

Kestens, Y., Theriault, M., Rosiers, F. D. (2006), Heterogeneity in Hedonic Modelling of House Prices: Looking at Buyer' Household profiles, *Journal Geograph System*, 8: 61-96.

LeSage, J. (1999). The Theory and Practice of Spatial Econometrics, Department of Economics, University of Toledo, Ohio, USA. Available at http://www4.fe.uc.pt/spatial/doc/spatialeconometricsII.pdf (accessed on 12th February, 2008).

LeSage, J. and R. K. Pace (2009). Introduction to Spatial Econometrics, Boca Raton: CRC Press

Long, F., Paez, A., Farber, S. (2007), Spatial Effects in Hedonic Price Estimation:A Case Study in the City of Toronto, Centre for Spatial Analysis, McMaster University, Canada.

McCluskey, W. J., Deddis, W. G., Lamont, I. G., and Borst, R. A. (2000). The Application of Surface Generated Interpolation Models for the Prediction of Residential Property Values, *Journal of Property Investment and Finance*, 18(2): 162-176.

Pace, R. K., LeSage, J. and Zhu, S. (2009). Impact of Cliff and Ord on the Housing and Real Estate Literature, *Geographical Analysis*, 41:418-424/

Paez, A., Long, F., and Farber, S. (2008). Moving Window Approaches for Hedonic Price Estimation: An Empirical Comparison of Modelling Techniques, *Urban Studies*, 45(8): 1565-1581.

Zhang, L., Ma, Z., and Guo, L. (2009). An Evaluation of Spatial Autocorrelation and Heterogeneity in the Residuals of Six Regression Models, *Forest Science*, 55(6): 533-548.