KRIGING IN MASS APPRAISAL FOR RATING

[1]Taher Buyong, [2]Suriatini Ismail, [2]Ibrahim Sipan, [2]Mohd Ghazali Hashim and [1]Mohd Nurul Fakhri bin Adan

[1]Institute of Advanced Technology
Universiti Putra Malaysia
43400 UPM Serdang, Malaysia
Email: rehat_55@yahoo.com

[2]Faculty of Geoinformation Science and Engineering
Universiti Teknologi Malaysia
81310 UTM Skudai, Malaysia
Email:
suriatini@fksg.utm.my, ibrahim@fksg.utm.my, fakhri.adan@gmail.com

ABSTRACT

The application of geostatistical method, kriging, to predict the unknown property values from observed data of known property values in the context of mass appraisal is described in this paper. This method uses semivariogram to define the weight of neighboring observations in the prediction procedure. Kriging has been used successfully in predicting underground surfaces such as ore bodies and soil nutrients but receive less attention from valuation and econometrics experts. Several variations of kriging are available but the focus of the paper is on ordinary kriging. The theory is described and computational method applicable to hedonic price equation is shown. Transaction data within Majlis Perbandaran Kulai from year 2004 to 2006 are used as a case study. The results are compared with the traditional method of Multiple Regression Analysis (MRA) in term of its prediction performance.

Keywords: Property tax, Multiple Regression Analysis (MRA), spatial regression, spatial autocorrelation.

INTRODUCTION

Mass valuation of residential properties for rating purposes in Malaysia was proposed by Azhari Husin (1990). This attempt is justified because over 80% of taxable properties in local governments can be classified as residential properties. The procedures were based on the traditional Multiple Regression Analysis (MRA). The main problem with the MRA method is the ignorance of spatial dependence in the model. The importance of spatial dependence in property price analysis has been established and inappropriate treatment of property data with spatial dependence can lead to estimated coefficients that are either biased and inconsistent, or unbiased and consistent but inefficient. Estimated coefficients with such characteristics affect the reliability of hypothesis testing and prediction accuracy.

Ignoring spatial dependence in the MRA model produces residuals (errors) that are spatially auto-correlated. Many regression techniques, including MRA, see spatial autocorrelation as nuisance and necessary steps are taken to eliminate or reduce it. Spatial autocorrelation, however, has been useful in explaining house price and thus invaluable in the prediction process. Even though there are variations in the approach of prediction, i.e., the spatial econometrics and kriging approaches (Dubin, 1998; Basu, 1998), both approaches involves estimating the structure of the spatial autocorrelation in a dataset, and this information is incorporated in the prediction.

The theory of Kriging with respect to property price prediction has been empirically tested in studies in the western countries (Dubin, 1998; Martinez et. al., 2000; Chica-Olmo, 2007). However, it is necessary to conduct an evaluation of this method if it is to be adapted for rating mass valuation in Malaysia. A research to assess the predictive performance of the method is in progress and this paper presents preliminary findings of the research. In the next section, models and methods of MRA and kriging are discussed and this is followed by discussions on the data used for the analysis. Results and related discussions are found in the subsequent section. The paper concludes by highlighting important points.

MODELS AND METHODS

Multiple Regression Analysis

MRA is one of the most widely used statistical techniques to analyze data in order to describe, control and predict the value of one variable (Kutner M.H. 2004). MRA explores and quantifies the relationship between independent variables (land area, floor area, number of bedrooms, etc.) and a dependent variable (price). The MRA equation in matrix term is given by

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$$

where Y is the vector of dependent variable, β is the vector of coefficients, X is the matrix of independent variables and ε is a vector of independent random variable with $\mathbf{E}\{\varepsilon\} = 0$. If the number of dependent variable equals to or greater than the number of coefficients, β, the coefficients can be estimated by

$$\hat{\beta} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'X})\boldsymbol{Y}$$

The estimated vector of coefficients contains information concerning the relationships between independent variables and dependent variable in the study region. Thus, it can be used to predict dependent variable (price) of any house in the region, provided that the values of its independent variables (land area, floor area, number of bedrooms, etc.) are known. The equation to perform the prediction, in matrix from, given by

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} + \boldsymbol{\varepsilon}$$

where X is the matrix containing values of the independent variables, the characteristics of the houses, whose prices are to be predicted.

The fitness of the MRA model (how well the variability in the dependent variable is explained by the independent variables) can be measured by the coefficient of determination, R^2, adjusted coefficient of determination, $\overline{R}^2$, and standard error of the estimate, s_e. The equations for these measures are

$$R^2 = 1 - \frac{SSE}{SSTO}$$

$$\overline{R}^2 = 1 - \frac{(n-1)}{(n-m-1)} - \left(1 - R^2\right)$$

$$s_e = \sqrt{MSE}$$

where

SSE = $\sum (y_i - \hat{y}_i)^2$

SSTO = $\sum \left(y_i - \overline{y}_i\right)^2$

MSE = $\sum (y_i - \hat{y}_i)^2 / (n - m - 1)$

y_i = observed price, $\hat{y}_i$ = predicted price, and $\overline{y}_i$ = mean price.

n = number of observations

m = number of coefficients

The F-test is used to determine the relationship between the dependent variable and the set of independent variables of the population using sample data. The null and alternative hypotheses are

H_0: $\beta_1 = \beta_2 = \ldots = \beta_m = 0$
H_a: at least one $\beta_i \neq 0$

The *F*-test statistic is given by

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{m}}{\frac{SSE}{n-m-1}}$$

The *t*-test is used to determine the relationship between the dependent variable and each independent variable in the population using sample data. The null and alternative hypotheses are

H_0: $\beta_i = 0$
H_a: $\beta_i \neq 0$ for i = 1, 2, 3, …, m

The *t*-test statistic is given by

$$t = \frac{b_i - \beta_i}{s_{b_i}}$$

Ordinary Kriging

Kriging is an interpolation method that can predict values of unknown points of a random function, random field, or random process. These predictions are best linear unbiased estimators and also weighted linear combinations of the observed values. There are several variations of kriging methods, and ordinary kriging is one of them.

The general kriging equation is given as

$$Z(p) = \mu(p) + \varepsilon(p)$$

where Z(p) is the value of a random variable at point p, μ(p) is the deterministic component at point p and ε(p) is the stochastic spatially correlated errors at point p.

Ordinary kriging requires random variables of constant mean, □, that does not vary with location. Variations in the mean (called trend) must be removed if ordinary kriging, given by

$$Z(p) = \mu + \varepsilon(p)$$

is to be used. Trend in the random variables can be estimated by fitting a polynomial function to the study region and estimating its value at every data point. Removing trend form each data point leaves residual which is the spatially correlated error, ε(p). That is, the general kriging equation can be re-written as,

$$\begin{aligned} \varepsilon(p) &= Z(p) - \mu(p) \\ \text{if} \quad \mu(p) &= \hat{Z}(p) \\ \varepsilon(p) &= Z(p) - \hat{Z}(p) \end{aligned}$$

Removing trend leaves spatially correlated errors, ε(p), to work with in determining the semivariogram and prediction process instead of the observed Z(p) value of the random variable. Trend must be added back to the predicted spatially correlated error to obtain the true predicted values. Thus,

$$\hat{\varepsilon}(p) = \sum_{i=1}^{k} \lambda_i \varepsilon(p)$$

where k is the number of neighbor and,

$$\hat{Z}(p)_{final} = \hat{Z}(p) + \hat{\varepsilon}(p)$$

where $\hat{Z}(p)$ is the polynomial estimation.

The most important step in ordinary kriging (and all other kriging methods) is determining a semivariogram model that describes the spatial autocorrelation (spatial autocorrelation is related to semivariance) of pints in a region. This semivariogram model is analogous to □ coefficients in the MRA method; the model that was calibrated using points of known values in a region can be used to predict values of other unknown points in the region. The three most common semivariogram models are the spherical, exponential and Gaussian. The spherical model is given as,

$$\begin{aligned} \gamma(h) &= c_0 + c_1 \left\{ \left[\frac{3h}{2a} \right] - \frac{1}{2} \left[\frac{h}{a} \right]^3 \right\} && \text{for } 0 < h < a \\ &= c_0 + c && \text{for } h \geq a \end{aligned}$$

The exponential model is given as,

$$\gamma(h) = c_0 + c_1 \left\{ 1 - e^{\left(-\frac{h}{a}\right)} \right\}$$

and the Gaussian model is given as,

$$\gamma(h) = c_0 + c_1\left\{1 - e^{\left(-\frac{h}{a}\right)^2}\right\}$$

where c_0 is the nugget, c_1 is the sill, a is the range, h is the distance, b is slope of the semivariogram model, and e = 2.71828.

Spatial autocorrelation is practically zero when the distance between two points goes beyond its range. Thus, the number of known neighborhood points, k, and the shape of the neighborhood search (called the neighborhood search parameters), are two important parameters to be considered when performing prediction. Optimal neighborhood search parameters to be used with a particular semivariogram model can be determined from cross validation. Cross validation results in terms of mean errors, root-mean-square errors, standard errors, mean standardized errors and root-mean-square standardized errors provide the mean to assess and choose the optimal semivariongram model and the associated neighborhood parameters for prediction.

DATA

For the purpose of this paper, the analysis used a test dataset comprises 217 single storey terrace houses transacted between 2002 and 2006 within Majlis Perbandaran Kulai (MPKu) area. The dependent variable is the transacted price. The seven independent variables are shows in Table 1. Spatial variables quantifying accessibility and neighborhood, and temporal aspects of the data are ignored.

Table 1 Independent Variables

Variables	Descriptions
LandArea	Land area in square meter
Main_FA	Main floor area in square meter
Anc_FA	Ancillary floor area in square meter
Add_FA	Additional floor area in square meter
Position	Intermediate, end, corner
Floor	Floor finishing (tiles, cement, tiles and cement)
Title	Freehold, leasehold, others.

RESULTS AND DISCUSSIONS

Twenty-two records were randomly selected from the test dataset forming a prediction dataset while the remaining 195 records form a calibration dataset. The calibration dataset was used to calibrate the MRA and ordinary kriging models. The calibration of the MRA model resulted in a set of estimated coefficients which was then used in the prediction of prices in the prediction dataset. The calibration of the ordinary kriging model resulted in fitted semivariogram model which was used in the prediction of prices of the prediction dataset. The predicted prices of each model were compared with the observed or actual prices to determine the accuracy of the prediction.

MRA model

The MRA model calibration and prediction were carried out using SPSS package. Table 2 shows the results of the MRA model calibration. The value of the F-test statistic indicates that a regression relationship exist between price and the independent variable as a group. The adjusted R^2 of 0.60 is quite reasonable. The relationship between the price and each of the independent variables is very strong except ancillary floor area, as indicated by the value of the t-test statistic. This relationship is true even at 0.01% significance level (99% confidence level) except for position and title type which register 0.05% significance level (95% confidence level). The regression function of the MRA was found to be

Predicted value = 88713.241 + 82.999 (LandArea) + 191.901 (Main_FA) + 104.556 (Anc_FA) + 531.318 (Add_FA) + 5021.427 (Position) + 3343.322 (Floor) - 1204.429 (Title)

Table 2 MRA model calibration

F test	41.73
R^2	0.61
Adj R^2	0.60
S_e	9639.14

	Coefficients		
	B	**Std. Error**	**t statistic**
(Constant)	88713.241	8104.140	10.947***
Land Area	82.999	31.814	2.609***
Main_FA	191.901	86.572	2.217**
Anc_FA	104.556	94.580	1.105
Add_FA	531.318	70.122	7.577***
Position	5021.427	3249.797	1.545**
Floor	3343.322	920.052	3.634***
Title	-1204.429	902.780	-1.334**

Dependent variable: Price
***significant at 0.01%, **significant at 0.05%, *significant at 0.1%

Ordinary kriging model

The ordinary kriging model calibration and prediction were carried out using ArcGis Geostatistical Analyst. The results of the cross validation process of model calibration, with neighborhood search parameters k = 5 and shape = four sector with 45 degree offset, are shown in Table 3. The mean standardized errors of all models were close to zero indicating that the predictions of all models were unbiased; any of the models could be chosen based on this criterion. However, further selection processes narrowed down the selection to the spherical model. The spherical semivariogram model was selected because it had the smallest root-mean-square error among all models; the smallest root-mean-square error indicated that the model's predictions were much closer to the observed values. On top of that, its average standard error was almost equal to its root-mean-square error and its root-mean-square standardized error was much closer to one, when compared to other models, indicating that mean model produced valid standard errors of predictions. Therefore, the spherical semivariogram model was used in the next step of the analysis; to predict values of properties in the prediction dataset.

Table 3. Cross-validation results for ordinary kriging

	Spherical	**Exponential**	**Gaussian**
Mean:	91.51	85.09	74.45
Root-Mean-Square:	8807	8815	8853
Average Standard Error:	7978	7974	7967
Mean Standardized:	0.01186	0.01153	0.009541
Root-Mean-Square Standardized:	1.093	1.095	1.1

Prediction Accuracy

Information generated by model calibrations was used to predict prices of houses in the prediction dataset. The predicted prices of each model were compared with the observed or actual prices already available in the prediction dataset to determine which model gives better prediction. A model that predicts prices closer to the observed prices is said to be a better model in terms of prediction accuracy.

The accuracy of prediction was assessed using Coefficient of Dispersion (COD) performance which is the average absolute percentage of dispersion around the median predicted ratio. The COD assesses the accuracy of the predictive model with measures how closely the individual ratios are arrayed around the median ratio. The more closely

the ratios are grouped around the median, the more equitable was the predicted model. The COD can be calculated using,

$$COD = \frac{100}{R_m}\left(\frac{\sum_{i=1}^{n}\left|R_i - R_m\right|}{n}\right)$$

where;

COD = Coefficient of dispersion;

R_i = Predicted ratio for each houses; $R_i = \frac{\hat{p}_i}{p_i}$

R_m = Median predicted ratio;

n = Number of houses predicted.

Other useful measures are the minimum and maximum deviation of the predicted price from its observed price. Table 4 summarizes the assessment of prediction accuracy of both models. The COD measure indicates that the MRA and kriging models are equal. The minimum deviation measure indicates that the ordinary kriging model is better while the maximum deviation measure indicates that the MRA model is better. The COD is a better measure and thus, it can be concluded that both models perform equally in terms of prediction accuracy.

Table 4 Summary of the assessment of prediction accuracy

	MRA	**Ordinary kriging**
Minimum deviation	RM120787.006	RM119102.41
Maximum deviation	RM152285.016	RM178032.09
Coefficient of Dispersion (COD)	4.0%	4.0%

CONCLUSIONS

The main objective of this paper is to preliminary asses the performance of the kriging model in mass appraisal of residential properties in Malaysia. The focus of the assessment is on the prediction accuracy in order for the model to be useful for mass appraisal, without compromising other aspects. This was done by comparing the performance of the ordinary kriging model and the conventional MRA global model. It was found that, despite limitations of the test dataset (small size, uses structural variables only and ignoring temporal effects) which make the MRA model inferior, the kriging model failed to outperform the MRA model. The expectation is that the kriging model will predict better than the MRA model.

This research will proceed with testing of datasets without the limitations indicated above. Larger and more representative test datasets with structural, accessibility, and neighborhood variables will be built and out-sample prediction will be carried out. The model will also incorporate temporal aspects of the data. A superior MRA model can thus be specified. The investigated kriging models will also focus on the use of residuals of spatial hedonic models as the observed values of house prices to account for spatial dependence of observations. Co-kriging will also be investigated as many of the independent variables have shown significant association with price. It is hoped that the suitability of the MWR model to be used for mass appraisal of residential properties for rating in Malaysia will be ascertained.

ACKNOWLEDGEMENT

Financial support of this research under Agreement No: NAPREC (R&D 1/07) is gratefully acknowledged.

REFERENCES

Chica-Olmo J. (2007). Prediction of Housing Location Price by a Multivariate Spatial Method: Cokriging. Universidad de Granada, Granada, Spain.

Basu, S. and Thibodeau, T.G., (1998). Analysis of Spatial Autocorrelation in House Prices. *Journal of Real Estate Finance and Economics*, 17(1): 61-85.

Dubin R.A. (1998). Predicting House Prices Using Multiple Listings Data. Journal of Real Estate Finance and Economics, Vol. 17:1, 35-59.

F. long, A. Paez, S. Farber (2007). Spatial Effects in Hedonic Price Estimation: A Case Study in the City of Toronto. Center for Spatial Analysis, Working paper series.

Hamid, Abdul, bin Hj. Mar Iman (2007). Modelling Locational Factors Using Geographic Information System Generated Value Response Surface Technique To Explain and Predict Residential Property Value. Centre For Real Estate, Universiti Teknologi Malaysia.

Kutner M.H., Nachtsheim C.J., Neter J. (2004). Applied Linear Regression Models Fourth Edition, New York: McGraw-Hill/Irwin

Martinez M.A, Lorenzo, Rubio. (2000). Kriging Methodology for Regional Economic Analysis: Estimating the Housing Price in Albacete. JEL O47; International Advances in Economic Research, 6(3): pp. 438-450,

McCluskey W.J., Deddis W.G., Lamont I.G., Borst R.A. (2000). The application of surface generated interpolation models for the prediction of residential property values. Journal of Property Investment & Finance, Vol. 18 No. 2, 2000, pp. 162-176.

Taher Buyong (2007). Spatial Data Analysis for Geographic Information Science, Skudai: UTM Publisher.

Van Beers WCM and Kleijnen JPC (2003). Kriging for interpolation in random simulation. Journal of the Operational Research Society (2003) 54, pp. 255–262.

Webster R., Oliver M.A. (2001). Geostatistics for Environmental Scientists. John Wiley & Sons Ltd, Chichester.