Spatial stochastic regression modelling of urban land use

S H M Arshad[1], J Jaafar, M Z Z Abiden, Z A Latif and A R A Rasam

Centre of Studies for Surveying Science and Geomatics
Faculty of Architecture, Planning and Surveying
Universiti Teknologi MARA (UiTM), 40450, Shah Alam, Malaysia

E-mail: sitihasnizamuhdarshad@yahoo.com

Abstract. Urbanization is very closely linked to industrialization, commercialization or overall economic growth and development. This results in innumerable benefits of the quantity and quality of the urban environment and lifestyle but on the other hand contributes to unbounded development, urban sprawl, overcrowding and decreasing standard of living. Regulation and observation of urban development activities is crucial. The understanding of urban systems that promotes urban growth are also essential for the purpose of policy making, formulating development strategies as well as development plan preparation. This study aims to compare two different stochastic regression modeling techniques for spatial structure models of urban growth in the same specific study area. Both techniques will utilize the same datasets and their results will be analyzed. The work starts by producing an urban growth model by using stochastic regression modeling techniques namely the Ordinary Least Square (OLS) and Geographically Weighted Regression (GWR). The two techniques are compared to and it is found that, GWR seems to be a more significant stochastic regression model compared to OLS, it gives a smaller AICc (Akaike's Information Corrected Criterion) value and its output is more spatially explainable.

1. Introduction

Urbanization accommodates the driving factor and serves as a stimulant for growth and development in an area or a country. The three main factors that control the continual growth of an urban area are urban commerce, population increase and chains of goods and information networks. Urban growth and commerce are related to each other and inseparable. Urban areas serve as a platform for economic growth. Commercial and industrial activities usually concentrate in urban areas because of the convenience these areas offer. Furthermore, urban areas also had a more organized road network which leads to a fast and efficient transportation and highly productive labor markets.

Moreover, urban areas also aid the spread of products, ideas and human resources between urban, sub-urban and the rural areas. In turn, industries and commercial activities in cities attract other associated services to support them and this interdependency provides the urban areas with more competitive advantages. In addition, the population increased and leads to urban expansion. Concurrently, in most developing countries, natural increase contributes as much as that by rural-urban migration to the increase in urban population. Rearrangement of city boundaries also contributes to urban growth. The increase in the urban population has various social and economic implications,

[1] To whom any correspondence should be addressed.

which in turn has an effect on the environment. Rural-urban migration itself conveys many problems, closely tied to employment, housing and other basic amenities [1]. The rate of urban growth cannot be measured directly. Therefore, in this study regression modeling technique is used to model it.

Research works that use the stochastic regression model technique such as Cellular Automata, SLEUTH Model and Neural Network. Overall, GWR is the most used model for urban growth prediction. Some of the examples OLS and GWR method to model urban land use changes in Penang Island [2] and Sungai Petani [3], Minimum Density (MD) Procedure and GWR [4], Classic Logistic Regression and GWR [5], logistic regression and GWR model [6].Meanwhile, Multiple Linear Regression and GWR are used for modeling the spatial structure of urban heat island in the city of Wroclaw , Poland [7]. Moreover, Global Regression method [8], Principal Components Analysis and GWR are applied for Northern Ireland [9]. GWR are also applied in other field such as deforestation in Mexico [10] and hotel room price [11].

2. Urban Growth Model

Regression model is a model that has both deterministic and stochastic components. It can be expressed by equation $Y \sim p(y|x)$ which states that, for a given x, Y is generated at random from a probability distribution whose mathematical form is $p(y|x)$. It also allows you to make a "what-if" prediction as to the value of Y. In the deterministic components, these predictions will depend on the value of x but stochastic components do not allow precise value of y *to* be determined [12]. A probabilistic / stochastic model use ranges of values for variables in the form of probability distributions. The words "Stochastic" means being or having a random variable. A stochastic model is a tool for estimating probability distributions of potential outcomes by allowing random variation in one or more inputs over time. The random variation is usually based on fluctuations observed in historical data for a selected period using standard time-series techniques. Distributions of potential outcomes are derived from a large number of simulations (stochastic projections) which reflect the random variation in the input(s) [13]. The relationship can be derived by $Y \sim p(y)$.

Two stochastic regression model techniques are adapted in this study, namely OLS and GWR. OLS are a statistical technique that uses sample data to estimate the true population relationship between two variables. OLS can be derived by two equations,

$$E(Y_i / X_i) = \beta_o + \beta_1 X_i \text{ is the population regression line and} \quad (1)$$

$$Y_{i(hat)} = b_o + b_1 X_i \text{ is the sample regression equation.} \quad (2)$$

Where, Y = dependent variable, X = independent variable or the regressor. β_o =intercept of this line and β_1 is the slope. The intercept $b0$ and the slope $b1$ are the coefficients of the population regression line, also known as the parameters of the population regression line.

OLS allows us to find = b_o and b_1. OLS also produces a line that minimizes the sum of the squared vertical distances from the line to the observed data points (it minimizes $\Sigma e_i^2 = e_1^2 + e_2^2 + e_3^2 + \ldots\ldots\ldots + e_n^2$, where n is the sample size) [14]. e = limit of $(1+1/n)n$ as n approaches infinity, i= unit imaginary number

GWR is a local statistical technique that can evaluate spatial variations in a relationship. It is based on the "First Law of Geography": everything is related with everything else, but closer things are more related [15]. It also addresses the non-stationarity (unpredictable, cannot be modelled or forecast) directly and allows the relationships to vary over space, i.e., βs do not need to be the same everywhere. In the linear form GWR can be explained by :

$$y_i = \beta_{i0} + \beta_{i1} x_{1i} + \beta_{i2} x_{2i} + \ldots\ldots + \beta_{in} x_{ni} + \varepsilon_i, \quad (3)$$

instead of remaining the same everywhere, βs now vary in terms of locations (i)[15].

Based on the review, statement and equation above these two models are found to be necessary in assisting analytical process.

3. Study Area

Kuala Langat is one of the districts in Selangor, Malaysia. Situated at the south-western of Selangor, it embraces a total area of 885 m², and its population is 222,261 people according to 2010 Census Data. It is in between the districts of Klang to the north and Sepang to the east. The southern border separated Selangor from Negeri Sembilan state. The Strait of Malacca is it western border. The major towns in Kuala Langat are Banting, Bandar Jugra, Teluk Datok (district capital) and Morib. Morib is famous among tourist for its sandy beach.

4. Methodology

Overall, this research employs four stages of work. Figure 2 illustrates the steps involved.

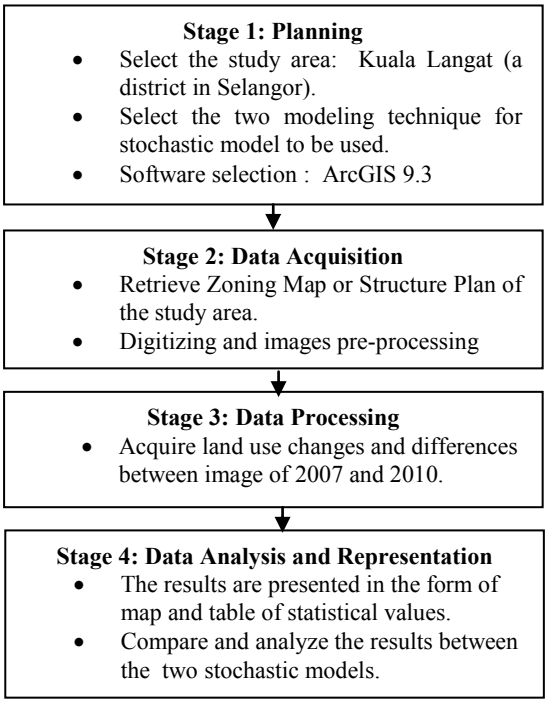

Figure 1. The flow chart of methodological process.

Referring to Figure 1, the first stage involved site identification, Kuala Langat is selected because of its rapid development and it's the Second Klang Valley (WLK II) Development Phase in the Selangor Structure Plan (RSN).Two stochastic modelling approaches are then applied to the selected area which are OLS and GWR.

In stage 2, the zoning areas are then identified from the structure plan and then digitized in ArcGIS 9.3, which is then converted later into shapefiles format. The next stage, OLS and GWR modelling are applied for processing. The two regression models are run consecutively in ArcGIS 9.3. Data processing at stage 3, the data value in the map is set to standard residuals for both years (2007 and 2010). Statistical measurements and performances are also calculated in order to verify that the calculations involved are acceptable and accurate. These calculations are performed on stage 4.

Finally, the results are presented in tables and maps. Tables showing the result of the two modeling technique are then distinguished.

5. Results & Discussion

GWR is called a local model because its implement a regression equation to every single feature in the dataset meanwhile OLS is classed as a global model because its implement a regression equation to an overall average of the features in the dataset. The statement "global models" derived the processes which are assumed to be stationary and as such are location independent but "local models" in contrast are spatial disaggregation of global models, the results of which are location-specific. Nevertheless, not all the spatial relationship indicators between OLS and GWR are the same. In OLS the spatial relationship indicators are no. of observation, degrees of freedom, Multiple R^2, Joint F- Statistic, AICc, Jarque-Bera Statistic, R^2 Adjusted and Joint Wald Statistic meanwhile in GWR the spatial relationship indicators are bandwidth/neighbor, Residual Squares, Effective Numbers, Sigma, AICc, R^2 and R^2 Adjusted. The values that are compared between the two models are R^2 Adjusted and AICc. AIC is a measure of spatial collinearity within the model data [2]. AICc is AIC with a correction for finite sample size.The AICc value is more preferable if it is smaller (the model with the lower AICc value provides a better fit to the observed data) meanwhile R^2 and R^2 Adjusted prefer higher value.

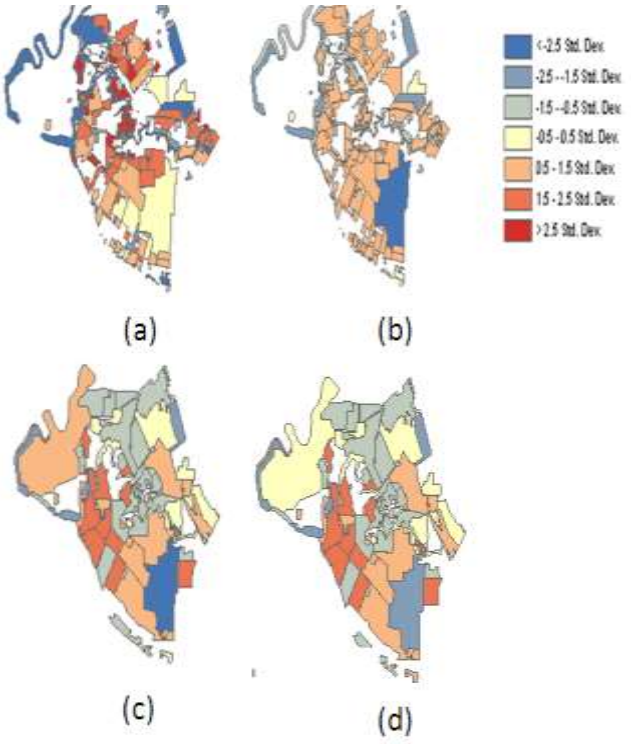

Figure 2. Map showing (a) GWR year 2007, (b) OLS year 2007, (c) GWR year 2010, (d) OLS year 2010.

Figure 2 shows the result map of GWR and OLS modelling. The value is in standard deviation (the variation or expansion exists for the average or expected value). A low standard deviation signifies that the data points very close to the average meanwhile high standard deviation signifies that the data points are spread out over a large range of values. The blue color indicates the over-predicted area while the red color indicates under-predicted area. The map for year 2007 is (a) for GWR and (b) for OLS meanwhile for the year 2010 the result is (c) for GWR and (d) for OLS. The results are supported by the statistic result table below:

Table 2. Statistic Result of OLS.

Method Summary	2007	2010
AICc	1607.2	477.49
R^2 Adjusted	0.017	0.018

Table 3. Statistic Result of GWR.

Method Summary	2007	2010
AICc	1504.961	475.060
R^2 Adjusted	0.288	0.077

Table 2 and 3 are the results obtained from OLS and GWR modelling using ArcGIS 9.3. The result that gained from the two regressions modelling is different. The AICc value for OLS is 1607.2 in 2007 and 477.49 in 2010 meanwhile for GWR the AICc value is 1504.96 in 2007 and 475.06 in 2010. The result showed that GWR is a better stochastic regression modelling because it gives a smaller AICc value which means that it provides a better fit to the observed data (the lower is the value of AIC; the better the fit is the model to observed data). The result is consistent with R^2 Adjusted value which in OLS is smaller 0.017 in 2007 and 0.018 in 2010 compared to GWR 0.288 in 2007 and 0.077 in 2010 (R^2 Adjusted prefer higher value or a value that nearing to 1). This suggests that the GWR model for Kuala Langat is better than the OLS model based on the AICc value. The study area and the dataset used in this study is smaller but the result is consistent with the result of previous works [2] [3].

Acknowledgement
The authors would like to thank the Kuala Langat District Council for providing dataset used in this study and Research Initiative Funding, UiTM for sponsorship.

References

[1] Khairulmaini U M, Salleh O and Ab Ghaffar F 2004 *Urban Environmental Hazards And Its Impact On The Urban* (Malaysia:University of. Malaya Fundamental Research) 43–57
[2] Shariff N M, Gairola S and Talib A 2010 *International Congress on Environmental Modelling and Software, Modelling for Environment's Sake, Fifth Biennial Meeting* (Ottawa:Canada)
[3] Noresah M S 2009 *18th World IMACS / MODSIM Congress* 13–17 July (Cairns:Australia) 1950–56
[4] Lee B 2006 Dissertation Abstracts University Of Southern California
[5] Luo J 2006 ProQuest Dissertations and Theses The University of Winconsin- Milwaukee
[6] Liao F H F 2012 *Spatial Determinants Of Urban Growth In Dongguan, China* (USA:Department of Geography, University of Utah)
[7] Szymanowski M and Kryza M 2011 *Procedia Environmental Sciences* **3** 87–92
[8] Platt R V *Agriculture, Ecosystems & Environment* **101** 207–18
[9] Lloyd C D 2010 *Computers, Environment and Urban Systems* **34** 389–99
[10] Pineda Jaimes N B, Bosque Sendra J, Gómez Delgado M and Franco Plata R 2010 *Applied Geography* **30** 576–91
[11] Zhang H, Zhang J, Lu S, Cheng S and Zhang J 2011 *International Journal of Hospitality Management* **30** 1036–43
[12] Deterministic-model @ www.businessdictionary.com homepage http://www.businessdictionary.com/definition/deterministic-model.html. [retrieved: 24-May-2013]
[13] Abiden M Z Z, Arshad S H M, Jaafar J, Latif Z A and Abidin S Z Z 2013 *9th IEEE Colloquium on Signal Processing and its Applications* 8-10 March (Kuala Lumpur:Malaysia)
[14] Hoyt G 2003 *Lecture Handout Over 15.3* (UK:Gatton College of Business & Economics, University of Kentucky)
[15] Fotheringham A and Brunsdon C 2002 *Geographically weighted regression: the analysis of spatially varying relationships* (England:John Wiley & Sons Ltd.)