Jurnal Ekonomi Malaysia 51(1) 2017 145 - 158
https://doi.org/10.17576/JEM-2017-5001-12

Segmenting Agricultural Land Market According to Development Potential: A Latent Class Approach

(Mengasingkan Pasaran Tanah Pertanian Menurut Potensi Pembangunan: Pendekatan Kelas Terpendam)

Haniza Khalid
International Islamic University Malaysia

ABSTRACT

Not all farmlands are purchased for farming. Where development pressures are strong and urban boundaries still fluid, some farmlands are purchased for non-agricultural purposes. However, since the future development use is not evident or pre-determined at the time of transaction, the farmland market may appear to operate as one albeit with latent segments. Analyses of land price determinants should involve some measures to ascertain the cause and the degree of functional segmentation in the market, so that the shadow prices of different land attributes can be differentiated by market segments. Using an extensive dataset of over 2,000 Malaysian farmland sales, our Latent Class Analysis confirms that there are two underlying distinct distributions and that within each distribution, relationships between variables display considerable local independence. Strength of potential drivers of farmland price is proven to differ according to segments. In addition, we are able to show that the segment classification results based on the parcel's 'developability' was fairly accurate when compared to the classification given by official land valuation documents. This exercise proves that unobserved segmentation can be predicted with a reasonable degree of accuracy simply by letting the data 'speak for itself'. In terms of agricultural support funding, the segmentation may allow for the country's better targeting of recipients and refinement of farm support programs.

Keywords: Latent class models (C38); farmland prices (Q1); hedonic price model (Q13); functional segmentation

ABSTRAK

Tidak semua tanah pertanian dibeli untuk tujuan pertanian. Di kawasan yang mana tekanan pembangunan adalah kuat dan sempadan antara bandar dan luar bandar masih tidak jelas, banyak tanah kebun dibeli untuk tujuan bukan pertanian. Guna-tanah masa depan boleh berubah-ubah dan tidak dapat dijangka dengan tepat pada tempoh urusan jual beli. Oleh kerana ini, kita boleh perhatikan bahawa pasaran tanah pertanian selalunya mempunyai pelbagai segmen terpendam, iaitu berdasarkan perbezaan niat sipembeli. Analisis penentu harga tanah haruslah melibatkan langkah untuk menentukan punca dan tahap segmentasi pasaran yang berlaku, supaya harga bayangan bagi setiap atribut tanah boleh dibezakan mengikut segmen pasaran masing-masing. Dengan menggunakan set data yang besar iaitu lebih 2,000 data jualan tanah pertanian di Malaysia, Analisa Kelas Pendam kami telah mengesahkan bahawa terdapat dua taburan yang berbeza dan bahawa dalam kedua-dua taburan/segmen, hubungan antara pemboleh ubah adalah berlainan. Kekuatan setiap faktor yang mempengaruhi harga tanah pertanian terbukti berbeza mengikut segmen. Selain itu, kami dapat menunjukkan bahawa hasil pengasingan segmen berdasarkan 'daya maju' sesebuah plot tanah adalah hampir sama dengan pengelasan yang dibuat Jabatan Penilaian Harta. Kajian ini membuktikan bahawa segmen terpendam masih boleh diramal dengan tahap ketepatan yang munasabah dengan kaedah membiarkan data 'bercakap untuk dirinya sendiri'. Menerusi pengasingan kelas terpendam di dalam pasaran tanah pertanian, pihak pembuat polisi dapat membuat sasaran polisi dan penghalusan program yang lebih tepat untuk memajukan sektor pertanian.

Kata kunci: Model kelas terpendam (C38); harga tanah pertanian (Q1); model harga hedonik (S13); segmentasi mengikut fungsi

INTRODUCTION

One of the most defining characteristic of land market studies is sample heterogeneity; basically due to the fact that land parcels seldom exist with standardised characteristics. For farmlands, heterogeneity in a sample can easily emerge from differences in the land's development potential, structural and spatial attributes as well as legal and fiscal constraints on land usage and so forth. Hence, analyses of land price determinants typically involve measures to ascertain the cause and degree of functional segmentation in a given market, so that the effects of each explanatory variable on price can be adjusted to the respective segments' profiles. By

demonstrating heterogeneity in the farmland market and in particular, predicting developability of farmland parcels, agricultural support programs can be suitably directed at specific spatial areas. Existing policy coverage do not fully take into account the potential for an area to deliver development benefits to its owners, making agricultural program suitability and variation in their effectiveness rather difficult to explain. Where there is strong development pressure coupled with somewhat poor efficacy of planning controls to regulate conversion of farmland to commercial, industrial and residential lands, it is even more critical that the market is properly segmented and that land parcels with strong 'development potential' is clearly identified. In order to optimise scarce resources (especially labour, technology and good farm management), the results of this study could suggest that over a long time horizon there is some merit to increasing the agricultural support to land with lower development potential.

Heterogeneous goods are normally characterised by a set of all its utility-bearing attributes or characteristics, which Rosen (1974) calls a "tied package of characteristics". It follows that its price should be estimated as a function of a vector of its attributes' values corresponding to their individual economic scarcity and worth. However, this approach of using a hedonic price function to estimate marginal values of each attribute assumes that buyers' valuation of them is constant across all categories of land. If market segmentation is suspected, this can be proven using *a priori* separation of the sample observations (i.e. along geographical, administrative, land-use lines) and then the model is tested for interaction effects.[1] However, this approach is not entirely without problems. Firstly, imposing too many interactive terms and intercepts can compromise model parsimony and this is particularly problematic in small samples studies. Secondly, if the model involves higher degrees of interactions, the estimated interaction parameters can become rather impossible to interpret meaningfully. Thirdly, interactive models for land also require that variables interacted must be of the same spatial unit, e.g. parcel or district or state, in order to make sense.

Magidson & Vermunt (2001) argue that in regression, discriminant and log-linear analyses, traditional models mainly involve parameters describing relationships between observed (or manifest) variables. If functional groupings cannot be established by looking at the observed variables, then some form of latent class modeling (LCM) is recommended. The model, which comes under the finite mixture model family, allows the inclusion of one or more discrete unobserved variables to determine segmentation in the data. Compared to interactive OLS models which assumes only certain parameters differ across market segments, the LC approach is less restrictive in that all moments and parameters are allowed to differ according to its 'latent' class. In addition, it lets the number of latent classes to be statistically determined, which is very useful when there are no segmentation principles suggested by theory. Latent variable approaches have been recognized as an effective tool to explore unobserved segments or groupings through observed attributes or attitudinal responses. For instance, in environmental economics, LCA helps segment distinct preference groups according to their willingness to pay for contaminated site clean-up (Patunru & Braden 2007) and landscape preservation (Morey, Thiene, De Salvo & Signorello 2008).[2]

The latent segmentation approach has been used in sociological (Eid, Langeheine & Diener 2003), health economics (Thacher, Morey & Craighead 2005) research, environmental economics (Boxall & Adamowicz 2002; Morey, Thacher & Breffle 2006). Only recently it is used to study buyer preferences in real estate markets (Patunru & Braden 2007; Sevenant & Antrop 2007; Morey et al., 2008; Rid & Profeta 2011). Our study builds upon existing literature but differs slight in that (i) it uses sales prices instead of preference survey data and (ii) it studies undeveloped land i.e. farmland, rather than completed residential properties. The LC method is used to confirm that the farmland market data has distinct underlying distributions and that the segmentation principles uncovered through the latent class approach (statistical method) is consistent with those determined *a priori* through government valuation exercises. Hence, LC can be used to provide statistical support to pre-set classifications (through dummy variables) used in hedonic price modeling. In our study of the Malaysian farmland sales values, the Property Market Report (PMR) (by National Institute of Valuation) data is pre-divided according to the land's development potential determined through the mandatory land valuation process, even though the actual type of development, and consequently its present net value to the landowner is unobservable at the point of time. Moreover, development 'potential' of a land changes *ad infinitum* until actual and irreversible investment on the land is made. By having the crucial 'development potential' information in the raw dataset, we are in position to conduct a unique natural experiment with respect to the following. Firstly, the whole dataset is tested without any 'developability' or other segmentation indicators to check the existence of any possible latent or 'hidden' segments. Secondly, if there are, we want to predict the classes from which each observation belongs to using conditional values of the explanatory variables derived in the regression. Thirdly, we compare the predicted latent segments to the given 'developability' segmentation in the PMR dataset. Finally, using the revealed latent classes, within-segment hedonic price regressions is carried out.

To achieve these ends, the paper is organized as follows: Section II introduces briefly conceptual aspects of latent variable analysis and provides the mathematical and empirical notions for the estimation exercise. Section III describes data and model specification and

Section IV discuss the results. The paper closes with some final remarks.

LITERATURE REVIEW

For a developing country pursuing vigorous economic transformation, the election of "development-friendly" land-use approach can be viewed simply as reflecting Society's preferences. Where the rewards to development are strong, interests of lobby groups, corporations and corrupt officials may increase the rate of farmland development. Even with comprehensive and fully-gazetted land-use plans, over-development is still possible if the land authorities are indifferent to long term land-use goals. In such a scenario, when empirically investigating the factors that drive farmland prices, the existence of different land purchasing motivations simply cannot be ignored. Conceptually, the market value of farmland that is subject to preservation should reflect the net present value (NPV) of future agricultural returns and very little else, since its development potential should be nullified by the program. However, Nickerson and Lynch"s (2001) study of land sales in the U.S. found little evidence of this, a result they attributed to an expectation of policy reversal when sufficient political and economic pressures emerged. In their study of Canadian farmland, Cotteleer, Stobbe and van Kooten (2008) also concluded that development speculation cannot be averted entirely and that its degree largely depends on perceived credibility of the land preservation programs' terms. Furthermore, conversion of small pockets of farmland for development has the effect of encouraging price speculation for other lands in the locality (Coughlin & Keane 1981) and the erosion of agricultural viability in general, especially where the parcels of land first converted are strategically important for access and water resources. Within such ambiguity, there is a strong potential for functional segmentation in the farmland market, which may be revealed through the use of latent class analysis.

Latent Class Analysis (LCA) stipulates that the distribution of the observed data is a mixture of a finite number of underlying distributions (Greene 2008: 558). Such circumstances can arise in any of these contexts:

1. The observed data is drawn from a mix of distinct underlying populations which collide or intersect in a study, such that the resulting parameters are heterogeneous (i.e. discrete) across the different sub-populations.
2. The distribution of the observed data is constructed from a mixture of two or more different underlying distributions, for instance two normal distributions with different parameters.

It is assumed that the responses on the manifest variables are the result of the unit's values with respect to the latent variables. According to McCutcheon (1987), by introducing the latent variable, independence is restored in the sense that within classes, variables are independent because the association between the observed variables is explained by the classes of the latent variable. Therefore, by controlling for latent variables in the model, the observed variables will be "conditionally independent" or in other words, there is clear local statistical independence.

LCA allows any combination of continuous, ordinal or nominal variables.[3] The presence of latent classes is detected from patterns of association among characteristics of the observed units. Hence, LCA is not really different from factor analysis and cluster analysis in that one can classify cases according to their maximum likelihood (ML) "class membership". However, in LCA, both tasks of classifying and estimating parameters are done simultaneously, that there is no need for a second-stage analysis following the group identification process (Magidson & Vermunt 2001).[4]

METHODOLOGY

MODEL

The following illustration of the model is based on Greene (2008) and Deb (2008). Assume that there are two underlying distributions for a given set of data. The probability that an observation is drawn from the first distribution, $N[\mu_1, \sigma_1^2]$, is unknown and denoted as λ_1 and the probability that the observation is drawn from the second is $\lambda_2 = (1 - \lambda_1)$. The density of the observed dependent variable, y, is therefore a linear combination of the C different densities. For this example of a two-class ($C = 2$) mixture model, the density function is

$$f(y) = \lambda_1 N[\mu_1, \sigma_1^2] + \lambda_2 N[\mu_2, \sigma_1^2] = \frac{\lambda_1}{\left(2\,\pi\sigma_1^2\right)} e^{-\frac{1}{2}[(y-\mu_1)/\sigma_1^2]^2} + \frac{\lambda_1}{\left(2\,\pi\sigma_2^2\right)} e^{-\frac{1}{2}[(y-\mu_2)/\sigma_2^2]^2} \quad (1)$$

Assuming that we know which population an observation comes from, we have for the i^{th} observation,

$$f(y_i|class_i = 1) = N[\mu_1, \sigma_1^2] = \frac{\exp\left[-\frac{1}{2}(y_i-\mu_1)^2/\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}} \quad (2)$$

and,

$$f(y_i|class_i = 2) = N[\mu_2, \sigma_2^2] = \frac{\exp\left[-\frac{1}{2}(y_i-\mu_2)^2/\sigma_2^2\right]}{\sigma_2\sqrt{2\pi}}$$

The contribution to the likelihood function is $f(y_i|class_i = 1)$ for an observation in class 1 and $f(y_i|class_i = 2)$ for an observation in class 2. Therefore, the unconditional marginal density for observation i is the probability-weighted additive density function

$$f(y_i) = f(y_i|class_i = 1) + \lambda_2 f(y_i|class_i = 2)$$

We estimate λ_1, μ_1, μ_2, σ_1 and σ_2 from the model using the log-likelihood function for a sample of n observations

$$\ln L = \sum_{i=1}^{n} \ln\left(\frac{\lambda_1 \exp\left[-½(y-\mu_1)/\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}} + \frac{\lambda_2 \exp\left[-½(y-\mu_2)/\sigma_2^2\right]}{\sigma_{21}\sqrt{2\pi}} \right) \quad (3)$$

To improve estimation of class probabilities, covariates i.e., information that help predict group probabilities are added to the model. Assuming that both distributions are normal, a mixture of normal model can be derived as follows

$$f(y_i|\mathbf{z}_i) = \left(\frac{\text{Prob}(class = 1|\mathbf{z}_i)\exp\left[-½(y_i-\mu_1)^2/\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}} + \frac{\left[1 - \text{Prob}(class = 1|\mathbf{z}_i)\right]\exp\left[-½(y_i-\mu_2)^2/\sigma_2^2\right]}{\sigma_2\sqrt{2\pi}} \right) \quad (4)$$

where z_i is a vector of variables that help to determine class probabilities. It is possible to estimate the respective class probabilities using a logit probability function whereby

$$\text{Prob}(class = 1|\mathbf{z}_i) = \frac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{2 + \exp(\mathbf{z}_i'\boldsymbol{\theta})}, \quad (5)$$

$$\text{Prob}(class = 2|\mathbf{z}_i) = 1 - \text{Prob}(class = 1|\mathbf{z}_i)$$

The respective probabilities in (5) represent the unconditional or "prior" probabilities, in a Bayesian sense, for an observation y_i to belong to a specific class. Substituting this into the log likelihood expression in (4), the following equation will be maximized

$$\ln L = \sum_{i=1}^{n} \ln\left(\left(\frac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})} \right) \frac{\exp\left[-½(y-\mu_1)/\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}} + \left(\frac{1}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})} \right) \frac{\exp\left[-½(y-\mu_2)/\sigma_2^2\right]}{\sigma_{21}\sqrt{2\pi}} \right) \quad (6)$$

The log-likelihood function is now maximised with respect to μ_1, μ_2, σ_1, σ_2 and θ. Consequently, (6) can be summarised to accommodate any C-number of classes,

$$\max_{\pi,\theta} \ln L = \sum_{j=1}^{N} \left(\ln\left(\sum_{j=1}^{C} \pi_j f_j(y|\boldsymbol{\theta}_j) \right) \right) \quad (7)$$

where π_j is the mixing probabilities, or the unknown probability of an observation classified into j^{th} class ($j = 1, 2, 3, ..., C$) such that $0 < \pi_j < 1$ and $\sum_{j=1}^{C} \pi_j$.

Explanatory variables can then be introduced into the analysis to improve accuracy of class membership probabilities. The resulting "posterior" or conditional probability, Pr$(class_j : \mathbf{z}_i : \mathbf{x}_i)$, is the probability of an observation i belonging to class j given its covariates and conditional upon the values of x, the vector of exogenous variables. More specifically, a conditional LC regression model can be derived by converting the marginal mean in j^{th} class, μ_j, into a conditional mean

$$\mathbf{E}(y_i|\mathbf{x}_i) = \sum_{j=1}^{C} \pi_j \mu_j \quad \text{where} \quad \mu_j = \mathbf{E}_j(y_i|\mathbf{x}_i) \quad (8)$$

All observations with the same observed characteristics would have the same conditional probability of belonging to the same segment. For the two-class case, the conditional probability function of categorized in Class 1 is

$$\text{Prob}(class = 1|\mathbf{z}_i, y_i) = \frac{f(y_i, class = 1|z_i)}{f(y_i)} = \frac{\left(\frac{\exp(\mathbf{z}_i'\boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i'\boldsymbol{\theta})} \right) \frac{\exp\left[-½(y-\mu_1)/\sigma_1^2\right]}{\sigma_1\sqrt{2\pi}}}{L_i} \quad (9)$$

where L equals the density function of y, as given in (4). The j^{th} class for an observation is simply the class associated with the largest estimated/predicted posterior probability.

$$\text{Prob}(class = j|\mathbf{z}_i, y_i) > \text{Prob}(class = k|\mathbf{z}_i, y_i) \quad (10)$$

where $j \neq k$. The conditional LC regression model for a C-component with predictor variables and covariates can be summarised as follows,

$$f(y|\mathbf{x}; \mathbf{z}; \theta_1, \theta_2, ..., \theta_C; \pi_1, \pi_2, ..., \pi_C) = \sum_{j=1}^{C} \pi_j(\mathbf{z}) f_j(y|\mathbf{x};\theta_i) \quad (11)$$

Marginal effects of predictors within each component or in the global model can be estimated using the respective functions.

$$\frac{\partial \mathbf{E}_j(y_i|\mathbf{x}_i)}{\partial \mathbf{x}_i} = \frac{\partial \gamma_i}{\partial \mathbf{x}_i} \text{ and } \frac{\partial \mathbf{E}(y_i|\mathbf{x}_i)}{\partial \mathbf{x}_i} = \sum_{j=1}^{C} \pi_j \frac{\partial \gamma_i}{\partial \mathbf{x}_i} \quad (12)$$

The procedure described above is based on the Expectation-Maximisation (EM) algorithm introduced by Dempster, Laird and Rubin in 1977. The algorithm is designed for ML estimation with missing information, i.e. the latent class variable. The iterative procedure finds the local maxima of the log likelihood function (McLachlan and Peel, 2000, Bartholomew and Knott, 1999) in two steps. In the E-step, the log-likelihood is replaced by its expected value, conditional on available information and initial values of the parameters as described by (9). In the M-step, the modified likelihood for each class is estimated separately using posterior probabilities as weights. The class associated with the observation is the one that yields the highest posterior probability value, as per (10).[5] Once the number of LC

that gives the best fit to the data is established, the next step is to examine parameter differences across classes for data segmentation principles. This task is relatively easy if the number of classes is small, say, 2, and the means of the class distribution are far apart. However, unambiguous discrimination can often occur especially where there is overlapping of distributions with a high number of classes - with or without significant outlier problems. Cameron and Trivedi (2009) suggest some warning signs of potential failure of the data segmentation process:

1. The log-likelihood increases slightly when additional classes are added
2. The log-likelihood "falls" when additional classes are added, indicative of a multimodal objective function.
3. One or more mixture distributions are too small in terms of the numbers of observations, such that latent classes cannot be distinguished through the exercise
4. Convergence is slow, indicative of a flat log likelihood.

It is therefore imperative that specification and evaluation of LCA functions must be done in full awareness of contextual and external information associated with the subject or commodity studied. We have so far described the underlying conceptual and mathematical substance of the LCA approach. The following section describes the application of the LCA to the Malaysian farmland data.

DATA AND VARIABLES

Using LCA for farmland price analysis can help confirm the presence of sub-markets of land from unobserved or latent sources, which could be prevalent given the nature of the good and usual data availability constraints. In principle, price of land parcels in the same segment should naturally be more correlated with each other than with parcels from other segments by virtue of them being derived from the same statistical distributions. Our database consist 2,222 actual farmland sales of various types of farmland from four states in the Central West coast of Peninsular Malaysia during a period of 7 years (between 2001 and 2007). The four states, Selangor, Perak, Negeri Sembilan and Melaka are selected because of their relatively higher growth rates of non-agricultural investment and population compared to the rest of the country. Data is obtained from the annual issues of the Property Market Report (PMR) published by the National Institute of Valuation, Valuation and Property Services Division, Ministry of Finance Malaysia. Demographic information is added using Malaysian Population Census reports of 2000 and 2010.

Table 1 gives the data description and summary statistics for all observations in the dataset. The single outcome measure is price per unit of land adjusted for inflation using year 2000 as the base year , *rprice*. The mean price in our data set is RM106,028 per hectare. Mean size of parcels within the sample, size, is 2.74 hectares. Approximately 20% of the sample parcels have road frontage. Road frontage, *rdfnt*, is hypothesized to give positive value to parcel price, irrespective of parcel's potential use. If a parcel is under land-transfer or land-use constraint, the relevant restriction dummy variables will take the value of one (zero otherwise). The two land restrictions examined are (i) Malay Reserve Land, *mrl*, introduced by the state authorities to bar sales of land in designated Malay-majority areas to non-Malay buyers and (ii) land under agrarian reform schemes, *gsa*, enacted through Group Settlement Act (1960). Around 22% of our land samples have Malay Reserve Land restrictions and another 22% subjected to GSA restrictions. Mean distance of the land parcels to the nearest town is 40.5km. Proximity of the land parcel to the nearest town area, *distown*, is expected to be positively related to unit price of land. Demographic information sheds light on changes occurring in the surrounding areas of a parcel, and is often employed to signal urban demands on the existing overall supply of land. Mean values for Population growth, *popgro*, and population density, *popden* are 1.96% per annum and 228.7 persons. Both variables are hypothesized to be positively associated with land price.

TABLE 1. Data Description and Summary Statistics

Variable	Description	Mean	Std Deviation	Min	Max
rprice	Sale Value per hectare (in RM) in 2000 prices	106,028	146,490	4,753	1,254,197
size	Parcel size (in hectares)	2.74	16.855	0.02	500
rdfront	1=Parcel with Road Frontage; 0=otherwise	0.202	0.402	0	1
distown	Euclidian distance to nearest urban centre (in km)	40.54	24.32	1.81	126.62
popden	District's population density based on 2000 Census	228.78	303.61	13.09	2516.08
popgro	Annualised district population growth based on 1991 & 2000 Census (in %)	1.96	2.66	-0.41	13.47
gsa	*gsa*=1 if located in Group Settlement Schemes	0.22	0.42	0	1
mrl	*mrl*=1 if located in Malay Reserve Land areas	0.22	0.41	0	1

Source: Author's estimation

The *rprice, size, distown* data have been transformed to log form to obtain normally distributed continuous variables. As a result, it is possible to impose a restriction that *f(y)* follows a normal or Gaussian distribution created from the mixture of two or more normal distributions with different parameters.[6]

EMPIRICAL STEP

Identifying the Optimal Number of Latent Classes The EM algorithm assumes that the appropriate number of classes, C, is known. If otherwise, the optimal number of segments is ascertained by comparing the log likelihood for different values of C and with different combinations of predictors and covariates. The best fitting model is determined using standard model selection principles, Akaike's Information criterion (AIC) and Schwarz's Bayesian information criterion (BIC). For non-nested ML models, information criteria measures commonly used are Akaike's Information criterion (AIC) and Schwarz's Bayesian information criterion (BIC). Stata uses the following scaling (Cameron & Trivedi 2009: 346).

$$\text{AIC} = -2 \ln L + 2k$$

$$\text{BIC} = -2 \ln L + k \ln L$$

where k is the number of free parameters to be estimated and $\ln L$ is the log-likelihood of the model at convergence. Smaller AIC and BIC are preferred, because they stand for higher log likelihood values. The quantities *2k* and *(k ln L)*, respectively, are AIC and BIC's penalties for model size.

The exercise begins by estimating the model with only the dependent variable, and a single class model (Table 2). Then, this is repeated by specifying two classes, following (3). Although the raw PMR data hints at a two-component finite mixture of normals (farmland with and without development potential), we proceed with cases of three, four and even classes to see if there are other roots of heterogeneity. The process was repeated as covariates and/or predictors are incrementally added to the model.

By convention, predictors are defined as exogenous factors that influence the dependent variable or outcome; while covariates are defined as attributes that influence the probability of an observation's classification into a particular latent segment. In this study, we are inclined to use land attributes as predictors rather than covariates, because following the hedonic valuation framework, it is the attributes' respective values that collectively give the land its overall value (dependent variable), regardless of development probability.[7] Nevertheless, (11) was tested for latent segments using both definitions. As Table 2 shows, the function is first tested without any covariate or predictor (specification A), with all OLS regressors included as predictors (specification B). Median splits on the two demographic variables, *hipopgro* and *hipopden* are then used to substitute *popgro* and *popden* (specification C). In specification D, all OLS regressors are set as covariates. In E, the original demographic variables, *popden* and *popgro* along with, *ldistown* and *rdfnt* are tested as covariates; while another two variables, *gsa* and *mrl* are retained as predictors. Specification F differs only slightly whereby *rdfnt* is moved to the predictor's group. Since *rdfnt* can affect price regardless whether a parcel is purchased for agricultural or development use, it may not predict development probability very well; therefore it could not be a covariate. For the final specification, G, we removed observations which are considered outliers and influential observations causing the sample size to fall to 1,901; and ran the model with the full list of original regressors as predictors.

i. Latent Class Regression
   Values of the mixing probabilities, $\hat{\pi}_j$ for all classes $j = 1,2$ are estimated. The algorithm parameterises π as a logistic function to constrain it to have a positive value. After the algorithm converges, $\hat{\pi}_j$ is recovered by transformation.[9]
ii. Comparing Marginal Effects Across the Latent Segments
   The marginal effects of predictor variables on land prices across latent classes are calculated at sample mean of the regressors, as per (12).

TABLE 2. Specification of the models are as follows:

A.	No covariate or predictor
B.	All OLS regressors included as predictors
C.	Median splits on the two demographic variables, *hipopgro* and *hipopden* are used to substitute *popgro* and *popden*
D.	All OLS regressors are set as covariates
E.	Original demographic variables, *popden and popgro* along with *ldistown* and *rdfnt* are tested as covariates; while another two variables, *gsa* and *mr*l are retained as predictors
F.	Variable *rdfnt* is moved to the predictor's group[8]
G.	Observations which are considered outliers and influential observations removed causing the sample size to fall to 1901 from 2222; but the model contains the full list of original regressors as predictors

Source: Author's own

iii. Class Membership from Estimation of Posterior Probability
For each observation the conditional or posterior probabilities of coming from specific distribution/ segments is then predicted.

iv. Within-Segment Hedonic Price Regression
For internal consistency, classifications based on predicted conditional probability is used to re-estimate the hedonic price function.

RESULTS AND DISCUSSION

This section describes the results from the empirical analysis according to the steps listed in the previous section.

NUMBER OF LATENT CLASSES

The estimates for a single-class model ($C = 1$) without covariates or predictors, is simply the sample mean and standard deviation of the dependent variable. The differences in log–likelihood values for specification A are very small, suggesting that the dependent variable, *rprice,* alone may not be sufficient to explain market heterogeneity. By adding predictors and covariates to the baseline model, the separation of log-likelihood values between different numbers of classes became increasingly clearer (Table 3). The results provide overwhelming support to the two-class model without outliers (specification G, $C = 2$), which gave the smallest AIC and BIC outcomes, 3075 and 3202 respectively, compared to all other specifications. The next best is the three-class model with full data and complete list of predictors (specification B, $C = 3$). However, since both

TABLE 3. Measures of fit for the Latent Class Models in Various Specifications

Model	Log-likelihood	AIC	BIC	No. of Parameters
Without Predictors or Covariates				
C = 1	-3114.9	6231	6237	2
C = 2	-3014.5	6039	6067	5
C = 3	-2993.1	6002	6047	8
C = 4	-2991.4	6004	6067	11
C = 5	-2988.5	6005	6085	14
With all OLS regressors as Predictors				
C = 2	-2248.3	4542	4673	23
C = 3	-2186.2	4442	4642	35
C = 4	-2155.2	4404	4672	47
C = 5	-2107.1	4332	4668	59
With all Predictors including *hipopden* and *hipopgro*				
C = 2				
C = 3	-2335.6	4717	4848	23
C = 4	-2311.8	4693	4893	35
C = 5	-2261.7	4617	4855	47
	a	a	a	a
With all OLS regressors as Covariates				
C = 2	-2413.0	5013	5093	14
C = 3	-2276.0	4604	4752	26
C = 4	-2206.3	4489	4706	38
With 2 Predictors and 6 Covariates				
C = 2	-2413.2	4860	4957	17
C = 3	-2226.6	4511	4676	29
C = 4	a	a	a	a
With 4 Predictors and 4 Covariates				
C = 2	-2384.7	4805	4908	18
C = 3	-2217.6	4495	4666	30
C = 4	-2194.0	4472	4711	42
With all OLS regressors as Predictors without Outliers				
C = 2				
C = 3	-1514.5	3075	3202	23
	-1481.2	3032	3226	35

[a] refers to failed estimation attempts after large number of iterations
Source: Author's estimation

specifications are essentially the same, except that in G outlier observations were removed, the study will focus only on the two-class model with specification G.

LATENT CLASS REGRESSION

Table 4 shows that the dataset can be segmented into two classes: 47 percent of the parcels belongs to LC1 and the rest LC2. The mean log price is higher in LC2, while standard deviations in both classes are small, in fact, lower than the full sample's. Figure 1 shows fitted values distribution by latent classes. The combination graph reveals that fitted values in LC1 has a density that is mildly right-skewed; and the observations are more concentrated between 9.8 and 10.8. On the other hand, LC2 is more widely dispersed, although a high proportion of the members fall between 10.6 and 11.7.

MARGINAL EFFECTS ACROSS DIFFERENT LATENT SEGMENTS

Overall, the table shows that marginal effects of explanatory variables differ across latent segments calculated at sample mean of the regressors, as per (12). Overall, marginal effects of explanatory variables differ across latent segments of the data for the 2-class model without outliers (Table 5). Notable observations are:

1. Mean log of price changes by a larger extent in LC2, when there are changes in any of these variables: *lsize*, *rdfnt, gsa* and *ldistown*.
2. Marginal effects of *popgro* and *popden* are similar in the two LC's.

CLASS MEMBERSHIP FROM ESTIMATION OF POSTERIOR PROBABILITY

For each observation the conditional or posterior probabilities of coming from specific distribution/ segments is then predicted, as per (10). When computing the percentage of deviation of fitted values from actual values in absolute terms, we found for the 2-class model, average percentage of deviation is 0.54 percent in LC1 and 0.01 in LC2.

WITHIN-SEGMENT HEDONIC PRICE REGRESSION

For internal consistency, the classification based on predicted conditional probability given in the preceding section is used to re-estimate the hedonic price function.

TABLE 4. Estimated Mixing Probabilities and Descriptive Statistics

Fitted Values	Mixing Probabilities	Standard Error	Mean	Standard Deviation	Minimum	Maximum
Pooled Model	-	-	11.02	0.983	8.467	14.042
Latent Class 1	0.47	0.402	10.46	0.555	9.411	12.729
Latent Class 2	0.53	0.402	11.21	0.643	9.862	13.425

Source: Author's estimation

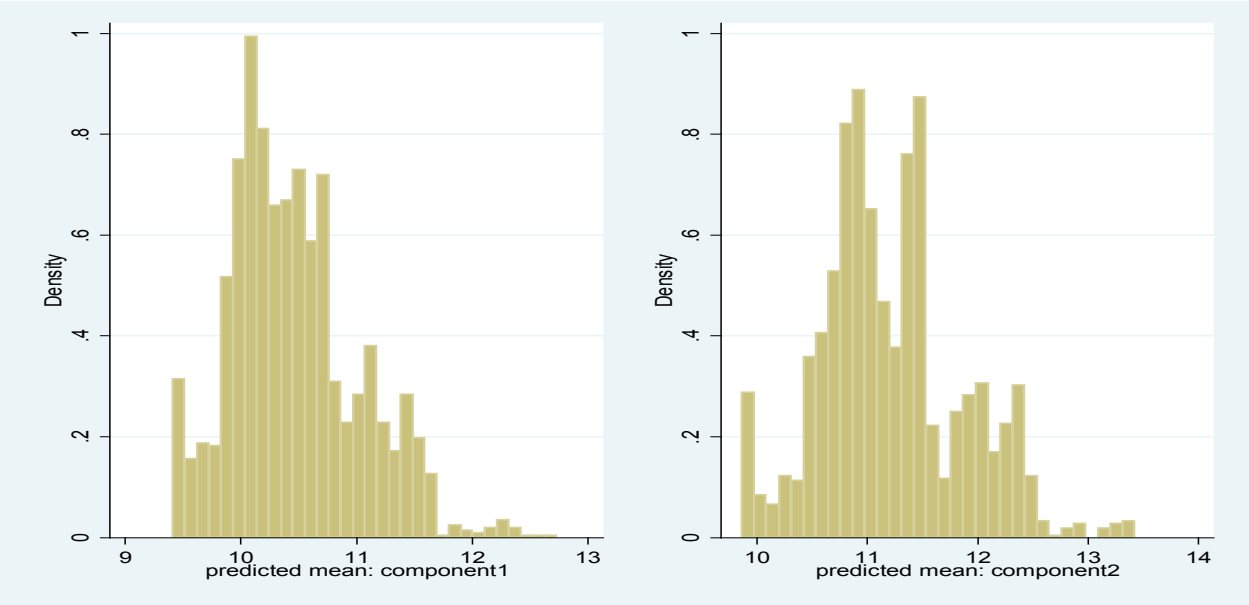

FIGURE 1. Histograms of Distributions of Fitted Means for the 2-class Model

TABLE 5. Marginal Effects based on Latent Classes

Predictors	*dy/dx* LC1	*dy/dx* LC2
$E(y \mid x)$	10.461	11.213
lsize	0.04 (0.024)	-0.05 (0.041)
rdfnt	0.79 (0.057)	0.90 (0.094)
gsa	-0.15 (0.112)	-0.40 (0.101)
mrl	-0.24 (0.151)	-0.24 (0.058)
popgro	0.20 (0.065)	0.20 (0.053)
lpopden	0.12 (0.036)	0.11 (0.029)
ldistown	-0.04 (0.051)	-0.15 (0.075)

(*) *dy/dx* is for discrete change of dummy variable from 0 to 1; Standard error is given in parentheses
Source: Author's estimation

The distribution of data following the posterior probabilities classifications is given in Table 6. The 2-class model without outliers follows a rather balanced ratio of 47:53 which is consistent with the earlier LCA estimation.

TABLE 6. Frequency Distribution according to Latent Classes

Latent Class	Frequencies	Percent
Class1	900	47.34
Class2	1001	52.67

Source: Author's estimation

The LC descriptive statistics table (Table 7) verified that the latent classes differ substantially from each other not only with respect to the mean price but also to land restrictions and urbanization variables. Noteworthy findings are:

1. Mean price of parcels in LC1 is less than half of that from LC2. The range of prices also indicate that parcels in LC2 are generally higher-priced than LC1.
2. There is higher percentage of parcels with road frontage in LC1.
3. Proportion of restricted land (*gsa* and *mrl*) are equal in both LC's.
4. More than half of LC2 parcels are located in districts with above sample average rate of growth and density.
5. Estimates for *distown* are not substantially different between the two classes.

From the table, it is possible to deduce that LC2 has higher development potential compared to LC1. The observations in LC2 are generally higher priced, located in high growth districts and have cities as the nearest town. The kernel density plot visualises the distribution of data over a continuous interval, as a variation of a Histogram that uses kernel smoothing to plot values to obtain smoother distributions. The peaks of a density plot help display where values are concentrated over the data interval. Figure 2 plots the distinct distributions for LC1 and LC2 where both are found to have different means and variances, although there is some overlapping of values at the tails of the distributions.

TABLE 7. Descriptive Statistics by Class

Variable	Latent Class 1			Latent Class 2		
	Mean	Min	Max	Mean	Min	Max
Real price (per ha)	40,026 (32,513)	5,363	341,684	106,659 (98,849)	15,748	678,843
Parcel size	1.9 (2.7)	0.14	40.34	2.1 (3.2)	0.08	41.07
Road Frontage	0.19 (0.39)	0	1	0.16 (0.36)	0	1
GSA land	0.25 (0.43)	0	1	0.24 (0.43)	0	1
MRL land	0.21 (0.41)	0	1	0.21 (0.41)	0	1
Population growth	1.34 (1.3)	-0.41	6.82	1.53 (1.43)	-0.41	10.9
Population density	194.8 (249.4)	13.1	1307	200.1 (259)	13.1	2516
Distance to Nearest town	32.2 (17.1)	4.0	90.6	32.9 (18.2)	2.9	91

* Standard deviations are given in parentheses
Source: Author's estimation

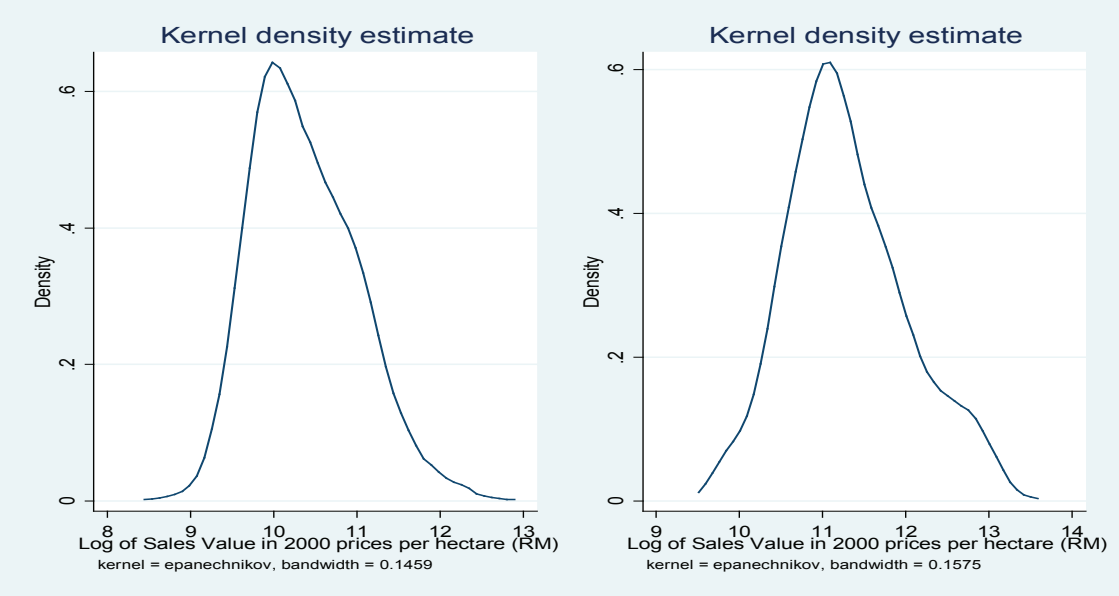

FIGURE 2. Kernel Density Plot of the Dependent Variable, *lprice*

1. The same baseline double log model is re-estimated separately for each LC of land and results are shown in Table 8 along with results of the original baseline OLS model:
2. The two classes are much better determined than the original baseline model: they each gave better goodness of fit measures. All parameters are statistically significant and have the expected signs
3. Generally, restrictions contributed negatively to price. However, the effect of MRL is less damaging than GSA, the former is perceived as less of a barrier to development of the land compared to the latter.
4. All of the proxies for urbanisation pressure (proximity to town, population growth and density) contribute positive marginal gains in price.

The difference between the two LC's essentially lies in the magnitude of the parameters. In almost all of the variables, the marginal effect is relatively stronger in LC2. For instance, using the natural scale for *lprice,* price of land with road frontage in LC1 is higher by 132 percent compared to parcels without; but the premium is 154 percent in LC2. Another example is elasticity of price with respect to distance to nearest town, *ldistown*, which is two times larger in LC2 compared to LC1.

In summary, regression results for the 2-class model lend ample support to the 'developability' definition of the two classes. LC1 appears to comprise land parcels with low development potential while LC2 are more liable to include parcels with high development potential. Figure 3 shows scatter plots of residuals on the y axis and fitted values (estimated responses) on the x axis for each of the Latent classes. These residuals versus fitted values plots are used to detect non-linearity, unequal error variances, and outliers within the respective group. Apparently there is no particular pattern in either Latent Class to indicate major statistical problems.

TABLE 8. OLS Regression Results by Latent Classes

VARIABLES	Full Sample	Latent Class	Latent Class
		(1)	(2)
Constant	10.14***	9.73***	10.73***
	(0.178)	(0.115)	(0.122)
lsize	-0.07***	0.01	-0.05**
	(0.016)	(0.014)	(0.015)
rdfnt	0.83***	0.84***	0.91***
	(0.041)	(0.025)	(0.032)
gsa	-0.39***	-0.19***	-0.33***
	(0.033)	(0.022)	(0.026)
mrl	-0.13***	-0.06**	-0.21***
	(0.036)	(0.023)	(0.029)
popgro	0.12***	0.17***	0.18***
	(0.009)	(0.011)	(0.010)
lpopden	0.18***	0.13***	0.13***
	(0.022)	(0.015)	(0.017)
ldistown	-0.10**	-0.08**	-0.15***
	(0.038)	(0.027)	(0.026)
Observations	2222	900	1001
R-squared	0.5055	0.801	0.782
Adj.R-squared	0.5035	0.799	0.780

Robust standard errors in parentheses *** p<0.001, ** p<0.01, * p<0.05
Source: Author's estimation

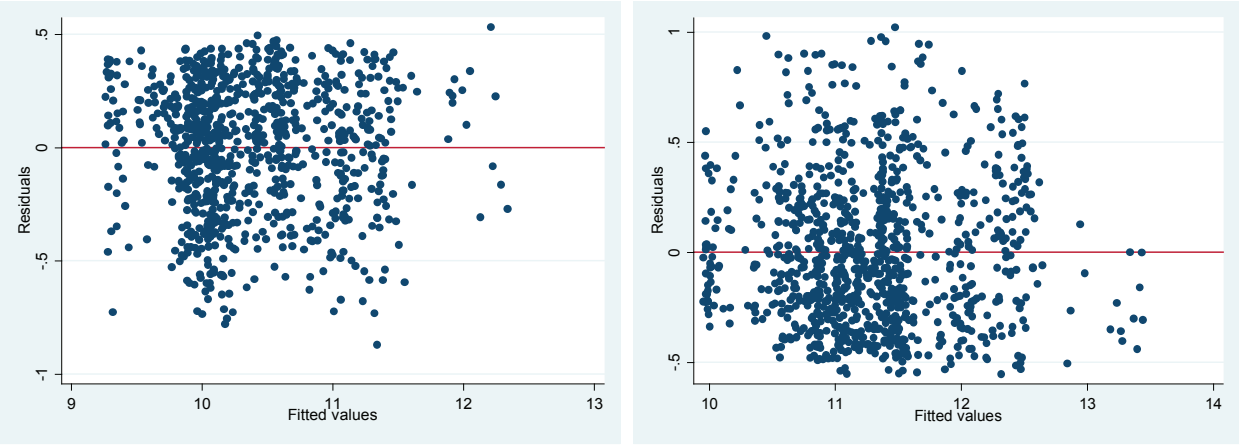

FIGURE 3. Residuals plotted against Fitted Values after OLS Regression

Table 9 shows favorable results for model diagnostics concerning heteroscedasticity, specification and autocorrelation for the LCA regressions. Through the Breusch-Pagan/Cook-Weisberg test, the null hypothesis of homoskedasticity cannot be rejected since the p-values > 0.05 in both Latent Classes. In other words, there may not be heteroskedasticity in either Latent class. The Ramsey Regression Specification Error Test (RESET) statistics are significant for both Latent classes, which means that at 5% significance level there is no evidence of misspecification of the functional form. Finally the mean VIF which is basically used to test for multicollinearity is close to 1 for both Latent Classes (1.57 and 1.67), which is very good since a mean VIF value of 1 indicates no correlation between predictors.

TABLE 9. Diagnostic results of OLS regression by latent class

Latent Class	Breusch-Pagan/Cook-Weisberg test	Ramsey RESET Test	Mean VIF
1	0.3186	0.0008	1.57
2	0.0950	0.0021	1.67

Source: Author's estimation

PREDICTING DEVELOPMENT CLASS

This section compares the segmentation suggested by the LCA with the segmentation by development potential of the land which is suggested in the PMR document.

Table 10 shows almost 86 percent of PMR developable farmlands are 'correctly' assigned to LC2. On the other hand, 54 percent of the PMR's non-developable farmlands are assigned to LC1 and the rest into LC2. This indicates that there are parcels assessed by the official valuers as not very 'developable' were actually perceived differently by the market.

Among the 306 observations which PMR identified as 'developable', the LC segmentation 'wrongly' classified only 43 land parcels into LC1. Explanation for this prediction error can be found by looking at Table 11

TABLE 10. Distribution of latent classes by category = developable land

	Latent Class 1		Latent Class 2		Total
	Frequency	% of Category	Frequency	% of Category	
Developable	43	*14.1*	263	*85.9*	306
Non-developable	857	*53.7*	738	*46.3*	1595

Source: Author's estimation

TABLE 11. Distribution of latent classes by category = development potential

	Latent Class 1		Latent Class 2		Total
	Frequency	% of Category	Frequency	% of Category	
Commercial	0	*0.0*	2	*100*	2
Industrial	0	*0.0*	10	*100*	10
Mixed Dev.	7	*38.9*	11	*61.1*	18
Residential	36	*13.0*	240	*87.0*	276

Source: Author's estimation

which shows a more detailed breakdown of the specific development envisaged for the parcel according to the PMR. The group of wrongly-classified parcels in LC1 comprises 7 with possible mixed development potential and 36 possible residential development potential. The parcel prices are generally less than half of the mean price of PMR's developable land category; mostly are located more than 15km away from the nearest town and in districts that register annualised population growth rate of less than 7 percent. These characteristics can explain why the model recommends that they be assigned to LC1 instead of LC2. In other words, these parcels may be perceived to have high development potential in the local market, but not by buyers in the larger land market due to their 'moderate' characteristics.

CONCLUSION

The objective of the paper is to test the likelihood of significant (but unobserved) heterogeneity in the farmland market. Using LCA, information on observable attributes of the land parcel is used to determine the number of hidden sub-markets as well as predict the sub-market membership probabilities. The full dataset was tested for the existence of latent or 'hidden' segments. Subsequently we predict for each observation the latent class it should belong to using conditional values of the explanatory variables derived in the regression. Once the data was organised into their respective latent classes, we re-estimated the hedonic pricing model for each sub-market separately. The estimation results are consistent with our expectations and goodness-of-fit measures improved substantially compared to the baseline regression on the full sample. The results for the Malaysian farmland market dataset confirm that there are two distinct distributions from which the land transfer data came from. The two classes vary in the level of importance attached to land restrictions and urban demand on land. Within each segment, the relationships between variables are constant/stable and display local independence.

We also compared the predicted latent segments from the LCA exercise to segments recommended in the Property Market Report, i.e. based on the land parcel's development potential. Essentially, the predicted classification concurs very well with the report's classification.10 The LCA exercise supports the notion that Malaysian farmland market is somewhat vertically differentiated i.e., parcels with certain characteristics favorable to development are more inclined to be sold at higher market prices as compared to parcels with lower values of the same characteristics.

The payoff from the LCA method is better convergence speed and statistically identifiable (or distinct) segments. This exercise proves that unobserved segmentation can indeed be predicted with a fair degree of accuracy by letting the data 'speak for itself'. It is possible for the researcher to pre-empt overly complicated statistical models and restrictive assumptions not to mention inaccurate *a priori* structures on a dataset suspected of being heterogeneous. The LCA is particularly appealing in situations where a study lacks quality data or suffers from other sorts of data constraints, as land price models are prone to, as well as in situations where there are more than one outcome or response measures. This is the first paper that we are aware of that uses LCA to explore market latent classification for land as an extension of the hedonic price modeling approach. Nevertheless robust checks are needed to validate the segmentation results; and perhaps Monte Carlo and bootstrapping procedures can be considered to augment the analysis.

What are the policy implications of understanding farmland market segmentation? Firstly, ad-hoc approvals that are given for farmland-conversion (and may be against urban containment policies of the country) leads to inappropriate pressure on farmland prices as 'development' and 'agricultural' buyers converge and compete in the same market. If left uncontrolled, it is inevitable that there will be unwarranted development pressures in green-field areas. Some of these pockets of successfully converted farmland will in turn, encourage pre-mature development speculation in the surrounding locality, eroding the critical mass and profitability of extant agricultural areas. Secondly, if market pressures cannot be effectively contained, then it is only wise to incorporate the future development potential of the land into urban planning exercise. This will ensure that, moving forward, the development of the urban region are planned and managed efficaciously. On the other hand, agricultural support programs and funding can be concentrated on farmland with low developability index, ensuring more sustainable and ardent effort by farmers receiving the support. The recent food crisis in 2007-2008 has shown that there is real need in protecting available farmland resources as part of a broader set of objectives to plan for food production, protect open space and the rural/agrarian character of the countryside, and particularly so in developing countries.

NOTES

1. More data-reliant techniques for segmenting involve methods such as clustering, factor analysis as well as many graphical applications.
2. Magidson and Vermunt, 2001, Morey et al. (2008) as well as Deb (2008) provide more elaborate explanation and examples of LCA applications.
3. Technically, an LC model is called Latent Profile Analysis if the independent variables are continuous and Latent Class Model if they are discrete. The manifest variables in latent profile analysis are continuous and in most cases, their distribution is assumed to be normal.
4. A more thorough comparison of cluster analysis and latent class analysis is provided in Thacher, Morey and Craighead (2005) and Aldrich et al. (2007); while Atella, Brindisi,

Deb and Rosati (2004), make an interesting case for latent class models over multivariate probit methods.

5. Despite its prevalent use, the EM algorithm has several notable drawbacks. The method is well-known to be computationally intensive and as a result, convergence is typically slow. According to Cameron and Trivedi (2009), the method is more tedious if the log-likelihood function is multi-modal and not log-concave. For instance, the presence of outliers normally causes the likelihood function to have more than one local maxima (Deb, 2008). Heckman and Singer (in Greene, 2008) noted that when the number of classes tested is larger than appropriate, the estimation breaks down and it is no longer possible to obtain meaningful parameters.
6. Depending on the nature of data used (count data or categorical dependent variable), other popular mixture class densities are Poisson, Gamma, negative Binomial, Student-t and Weibull.
7. For instance, where a state has a large stock of land available for agriculture and development relative to its population growth, price of farmlands would be more dependent on returns to farming than on its development potential.
8. Since rdfnt can affect price regardless whether a parcel is purchased for agricultural or development use, it may not predict development probability very well; therefore it could not be a covariate
9. Cameron and Trivedi, p. 580
10. Granted some corrections are necessary to remove outliers and influential points from the dataset.

REFERENCES

Aldrich, G., Grimsrud, K., Thacher, J. & Kotchen, M. 2007. Relating environmental attitudes and contingent values: How robust are methods for identifying preference heterogeneity? *Environmental and Resource Economics* 37(4): 757–775.

Atella, V., Brindisi, F., Deb, P., Rosati, F. C. 2004. Determinants of access to physician services in Italy: a latent class seemingly unrelated probit approach. *Health Economics* 13: 657–668.

Bartholomew, D. J., & Knott, M. 1999. *Latent Variable Models and Factor Analysis*. 2nd edition. London: Arnold. (Kendall's Library of Statistics 7).

Boxall, P. C., & Wiktor L. A. 2002. Understanding heterogeneous preferences in random utility models: A latent class approach. *Environmental and Resource Economics* 23(4): 421–446.

Cameron, A. C. and Pravin K. T. 2009. *Microeconometrics Using Stata*. TX: Stata Press.

Coughlin, R. E., & Keane, J. C. 1981. *The Protection of Farmland: Report to the National Agricultural Land Commission*. Washington, DC: US Government Printing Office.

Cotteleer, G., Stobbe, T. & Van Kooten, G. C. 2008. *A Spatial Bayesian Hedonic Pricing Model of Farmland Values*. Paper presented at the 12th European Association of Agricultural Economists, August 2008, Ghent, Belgique.

Deb, P. 2008. Finite Mixture Models. Summer North American Stata Users' Group Meetings 2008 7, Stata Users Group, revised 28 Aug 2008.

Dempster A. P., Laird N. M., Rubini D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statisitics Society* 3(9):1–38.

Eid, M., Langeheine, R., & Diener, E. 2003. Comparing typological structures across cultures by multigroup latent class analysis a primer. *Journal of Cross-Cultural Psychology* 34(2): 195-210.

Greene, H. W. 2008. *Econometric Analysis*. 6th edition. Pearson Education.

Magidson, J., Vermunt, J. K. 2001. Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology* 31: 223–264.

McCutcheon, A. L. 1987. *Latent Class Analysis, Sage University Paper*. Newbury Park: Sage Publications.

McLachlan, G. J. & Peel, D. 2000. *Finite Mixture Models*. New York: Wiley.

Morey, E., Thacher, J., & Breffle, W. 2006. Using angler characteristics and attitudinal data to identify environmental preference classes: a latent-class model. *Environmental and Resource Economics* 34(1): 91–115.

Morey, E., Thiene, M. De Salvo, M & Signorello, G. 2008. Using attitudinal data to identify latent classes that vary in their preference for landscape preservation. *Ecological Economics* 68(1-2): 536–546.

Nickerson, C. J. & Lynch, L. 2001. The Effect of Farmland Preservation Programs on Farmland Prices. *American Journal of Agricultural Economics* 83: 341–351.

National Institute of Valuation. Various years. *Valuation and Property Services Division. Property Market Report*. Kuala Lumpur: Ministry of Finance Malaysia.

Patunru, A. A., Braden, J. B. & Chattopadhyay, S. 2007. Who cares about Environmental Stigmas and does it matter? A latent segmentation analysis of stated preferences for real estate. *American Journal of Agricultural Economics* 89(3):712–726.

Rid, W. & Profeta, A. 2011. Stated preferences for sustainable housing development in Germany-a latent class analysis. *Journal of Planning Education and Research* 31(1): 26–46.

Rosen, S.1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy* 82: 32–55.

Sevenant, M., & Antrop, M. 2010. The use of latent classes to identify individual differences in the importance of landscape dimensions for aesthetic preference. *Land Use Policy* 27(3): 827–842.

Thacher, J. A., Morey, E. & Craighead, W. E. 2005. Using patient characteristics and attitudinal data to identify depression treatment preference groups: a latent-class model. *Depression and Anxiety* 21(2):47–54.