Journal

Transactions in GIS, 22(3)

ISSN

1361-1682

Authors

Park, Soyoung
Lee, Jae Hyun
Clarke, Keith C

Publication Date

2018-06-01

DOI

10.1111/tgis.12451

Peer reviewed

Received: 30 January 2018 | Revised: 12 April 2018 | Accepted: 15 April 2018

DOI: 10.1111/tgis.12451

RESEARCH ARTICLE

WILEY Transactions in GIS

Capturing the heterogeneity of urban growth in South Korea using a latent class regression model

Soyoung Park[1] | Jae Hyun Lee[2] | Keith C. Clarke[3]

[1]Spatial Information Engineering Department, Pukyong National University, Busan, South Korea

[2]Institute of Transportation Studies, University of California, Davis, CA, USA

[3]Department of Geography, University of California, Santa Barbara, CA, USA

Correspondence
Jae Hyun Lee, Institute of Transportation Studies, 1605 Tilia Street, University of California, Davis, CA 95616, USA.
Email: jaele@ucdavis.edu

Abstract

This study aims to analyze the spatial patterns of urban growth in South Korea between 2000 and 2010. Fourteen suspected causative independent variables were selected and latent class regression (LCR) was used to analyze the relationship between dependent (urban growth) and independent (causative) variables. The goodness-of-fit was assessed by comparison to logistic regression (LR) analysis. The LR analysis produced consistent coefficients for each independent variable across the study area. In contrast, an LCR analysis, with a three-class assumption, resulted in a different magnitude and directional effects of the coefficients for each class. The LCR analysis enabled the identification of both spatially homogeneous and heterogeneous areas. In addition, the LCR analysis performed better than the LR analysis with a lower Akaike information criterion and Bayesian information criterion value, and a higher receiver operating characteristic value. We conclude that LCR analysis should be used to establish causative "driving" factors for efficient urban growth planning and urban spatial policy.

1 | INTRODUCTION

Urban areas are good examples of complex systems: they include a multitude of interdependent measures that embed nonlinear feedbacks reflecting numerous interrelated demographic, social, economic, land-use, transportation, and behavioral subsystems. The big data era now assures us that for many of the world's cities, data about these measures are available in a timely fashion, are reasonably accurate, and are spatially disaggregated to detailed spatial resolutions. Yet what are these interrelations among the factors that contribute to urban change, and which of the causative relationships are predictable in time and space? Spatial analytic methods are now increasingly able to answer these questions.

For the past 20 years, the method of logistic regression (LR) analysis has been among the approaches most often used to identify the drivers of land-use change (Allen & Lu, 2003; Dendoncker, Rounsevell, & Bogaert, 2007; Hu & Lo, 2007; Millington, Perry, & Romero–Calcerrada, 2007; Wu, Huang, & Fung, 2009; Long, Gu, & Han, 2012). LR analysis is easily adapted for studying predictor variables in land-change applications because it is more appropriate for use with categorical variables (e.g. land-cover classes) than is ordinary least-square regression (Lambin, 1997; Overmars &

Verburg, 2005; Wang, Brown, An, Yang, & Ligmann–Zielinska, 2013). What LR analysis adds is the ability to explore qualitatively how urban growth and its causative factors interrelate. This has made it possible to understand which are the most influential variables and how to distinguish among them.

A disadvantage of LR analysis is its inability to reflect spatial non-stationarity. The assumption underlying LR analysis is that relationships remain both steady and unchanging across the area of study, but spatial relationships between independent and dependent variables can change, so that the global relationships generated in fact show only an average condition, and relationships specific to a locality may be hidden (Scott & Janikas, 2010; Su, Xiao, & Zhang, 2012).

Geographically weighted regression (GWR) is a more recently developed spatial analysis technique that provides an alternative way of examining relationships in greater detail (Fotheringham, Brunsdon, & Charlton, 2003). Maps of coefficients specific to a locality can be output by GWR analysis, making it possible to visualize geographical interactions and thus facilitating descriptions and predictions that more accurately and appropriately reflect the true situation (Foody, 2003; Wheeler & Páez, 2010; Su et al., 2012). Due to this advantage, GWR has become widely used in land-change science (Luo, Yu, & Xin, 2008; Partridge, Rickman, Ali, & Olfert, 2008; Shafizadeh-Moghadam & Helbich, 2015).

Since the GWR parameters are estimated from the closest samples (based on the first law of geography), GWR captures the spatial homogeneity between samples using the boundary distance to estimate local parameters and the spatial heterogeneity with geographically different parameters. Therefore, GWR analysis suffers from a lack of independence among local estimates, the presence of outliers, and the ineffectiveness of the estimated local coefficient due to the low sampling numbers (LeSage, 2001). Moreover, just one or two boundary distances are not able to capture all of the similarities between samples in terms of the probability of land-use changes.

Another method that captures spatial homogeneity and heterogeneity is latent class regression (LCR). LCR provides the functionality to identify spatially homogeneous areas (within each latent class) and heterogeneous areas (between latent classes), as well as the same parameters for the former, and different parameters for the latter, with respect to the probability distribution of the dependent variable, but without the low sampling and outlier problems of GWR. Improving the understanding of spatial patterns and of the factors underpinning urban growth is the aim of this study. The LCR model was used to analyze the relationship between dependent and independent variables that were factors in determining urban land-use change. The results were compared to the LR model using various statistical indicators. The following questions were addressed: (1) How is the LCR model best applied to land-change science? (2) How are the coefficients spatially different when using the LCR model? (3) Does the LCR model outperform the LR model?

2 | STUDY AREA

The Republic of Korea (hereafter "Korea") is located on the Southern Korean Peninsula between latitudes 33–39°N and longitudes 124–130°E. Its total area is 100,284 km^2. Korea's terrain is mostly mountainous, accounting for approximately 64% of the total land area (http://kosis.kr). Most of the low land, with an elevation averaging about 254 m, lies in the west and south-east (Figure 1a).

Korea's urban growth since the 1960s has been accelerated by industrialization and economic growth. Most Koreans live in urban areas, because Korea experienced rapid rural–urban migration during its period of rapid economic growth. As of 2010, the total population of Korea was 49,410,000 and the gross domestic product (GDP) was $1,094 billion (http://kosis.kr). Along with rapid socioeconomic development, the urban population as a total count and as a percentage of the national population has increased by 32,754,000 and 54%, respectively, since 1960 (United Nations, Department of Economic and Social Affairs, Population Division, 2014). The years since 1995 have seen a pause in the trend for urban population increasing and rural population falling. The populations of megacities and metropolitan areas have decreased, and the urban population has begun to move to the outer peripheral areas of the megacities (Figure 1b).

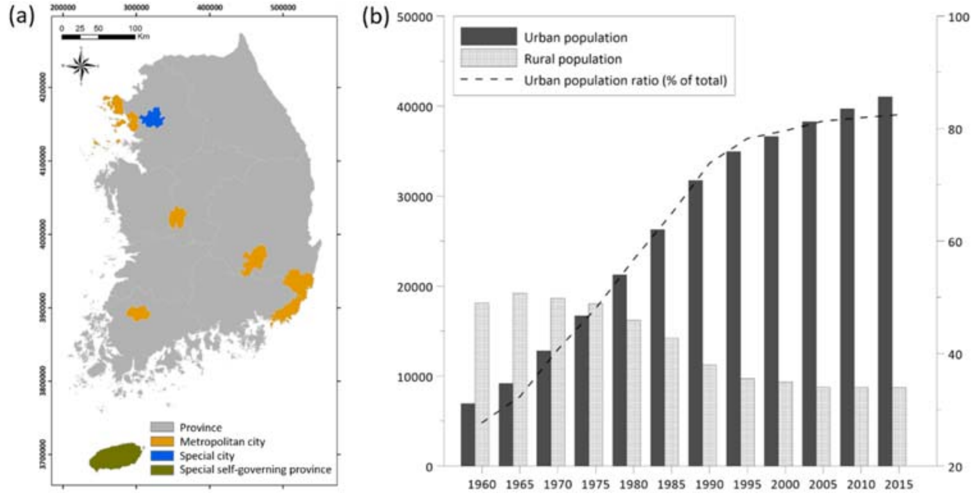

FIGURE 1 Location and status of the study area: (a) the location of the study area and the major administrative districts—eight provinces, one special self-governing province, six metropolitan cities, and one special city; and (b) the status of the urban and rural populations, and the proportion of the total population classified as urban according to the United Nations

Since the mid-1990s, Korea has been experiencing structural change in terms of its urban growth. Population growth began to be higher in nearby cities than in major metropolitan cities. Under the establishment of the local government system, urban society was decentralized and the societal structural qualitative changes that accompany urbanization began (Park, Kim, Ko, Kim, & Park, 2010). Given these representative conditions, the use of Korea as a study area enables an investigation of sustainable urban development planning against a background of persistent urban growth. In addition, by selecting 2000–2010 as the study period, the relationship between urbanization and its causative factors can be assessed.

3 | DATA AND METHODOLOGY

3.1 | Data preparation

There have been a number of approaches to find and interrogate the factors that drive urban expansion (Li, Zhou, & Ouyang, 2013). A literature review permitted identification of five categories into which driving factors for change of land use may be divided: biophysical factors, socioeconomic factors, spatial factors, neighborhood factors, and land-use policy factors (Table 1).

Fifteen variables reflecting these factors were selected, considering the difficulty in obtaining both spatial and aggregated data. The 15 variables were converted to continuous grid raster files with a 30 m resolution using ArcGIS 10.1 software (Esri, Redlands, CA, USA) (Table 2, Figure 2).

3.1.1 | Biophysical factors

Human comfort and the possibility of establishing urban amenities can be seriously restricted by lack of water, lack of vegetation, extreme temperature fluctuations, and high humidity (Portnov & Hare, 2012). In addition, high altitudes and steep slopes increase the cost of constructing buildings and set constraints on building infrastructure, because flatter areas are generally more conducive to urban development.

Four variables were selected to represent biophysical factors. A 30 m resolution digital elevation model produced by Korea's Ministry of the Environment permitted modeling of the elevation and slope, the slope itself being calculated

TABLE 1 Major driving factors affecting urban growth as identified from a literature review

Types of factors	Driving factors
Biophysical factors	**Elevation** (Dendoncker et al., 2007; Li et al., 2013) **Slope** (Dendoncker et al., 2007; Hu & Lo, 2007; Wu & Fung, 2009; Poelmans & Van Rompaey, 2010; Li et al., 2013) **Temperature** (Dendoncker et al., 2007; Millington et al., 2007) **Precipitation** (Dendoncker et al., 2007; Millington et al., 2007)
Socioeconomic factors	**Population density** (Allen & Lu, 2003; Liu & Zhou, 2005; Hu & Lo, 2007; Millington et al., 2007; Wu & Fung, 2009) **Gross domestic product** (Liu & Zhou, 2005) **Migration** (Millington et al., 2007)
Spatial factors	**Distance to socioeconomic center** (Cheng & Masser, 2003; Hu & Lo, 2007; Luo & Wei, 2009; Poelmans & Van Rompaey, 2009) **Distance to roads** (Allen & Lu, 2003; Cheng & Masser, 2003; Liu & Zhou, 2005; Hu & Lo, 2007; Millington et al., 2007; Luo & Wei, 2009; Wu & Fung, 2009; Poelmans & Van Rompaey, 2010; Li et al., 2013) **Distance to built-up land** (Allen & Lu, 2003; Cheng & Masser, 2003; Millington et al., 2007; Poelmans & Van Rompaey, 2010) **Distance to water** (Allen & Lu, 2003; Cheng & Masser, 2003; Dendoncker et al., 2007; Millington et al., 2007; Luo & Wei, 2009)
Neighborhood factors	**Density of built-up land** (Cheng & Masser, 2003; Liu & Zhou, 2005; Dendoncker et al., 2007; Hu & Lo, 2007; Luo & Wei, 2009; Wu & Fung, 2009) **Density of undeveloped land** (Cheng & Masser, 2003; Luo & Wei, 2009)
Land-use policy factors	**Conservation area** (Allen & Lu, 2003; Hu & Lo, 2007) **Master plan** (Cheng & Masser, 2003)

as a percentage. Both temperature and precipitation were obtained from 73 weather stations operated by Korea's Meteorological Administration, calculated as the annual mean value between 1981 and 2010. The point temperature measurements were interpolated to a grid using universal kriging in ArcGIS 10.1.

TABLE 2 Variables used for statistical analysis

Variables	Types	Description	Source
Dependent variables			
Change	Dummy	Land-use change from non-urban to urban	
Independent variables			
Ele	Continuous	Elevation	KME[a]
Slo	Continuous	Slope	–
Temp	Continuous	Temperature	KMA[b]
Rain	Continuous	Precipitation	KMA[b]
Pop	Continuous	Population density	KNSO[c]
Mig	Continuous	Net migration	KNSO[c]
Grdp	Continuous	Gross regional domestic product	KNSO[c]
DistRoad	Continuous	Distance to main road	KMLIT[d]
DistUrban	Continuous	Distance to built-up area	KME[a]
DistWater	Continuous	Distance to water body	KME[a]
DenFarm	Continuous	Density of agriculture land	KME[a]
DenFore	Continuous	Density of forest land	KME[a]
Environ	Dummy	Environmental conservation zone	KME[a]
Law	Dummy	Legal protection zone	KME[a]

[a]Korea's Ministry of Environment.
[b]Korea's Meteorological Administration.
[c]Korea's National Statistical Office.
[d]Korea's Ministry of Land, Infrastructure and Transport.

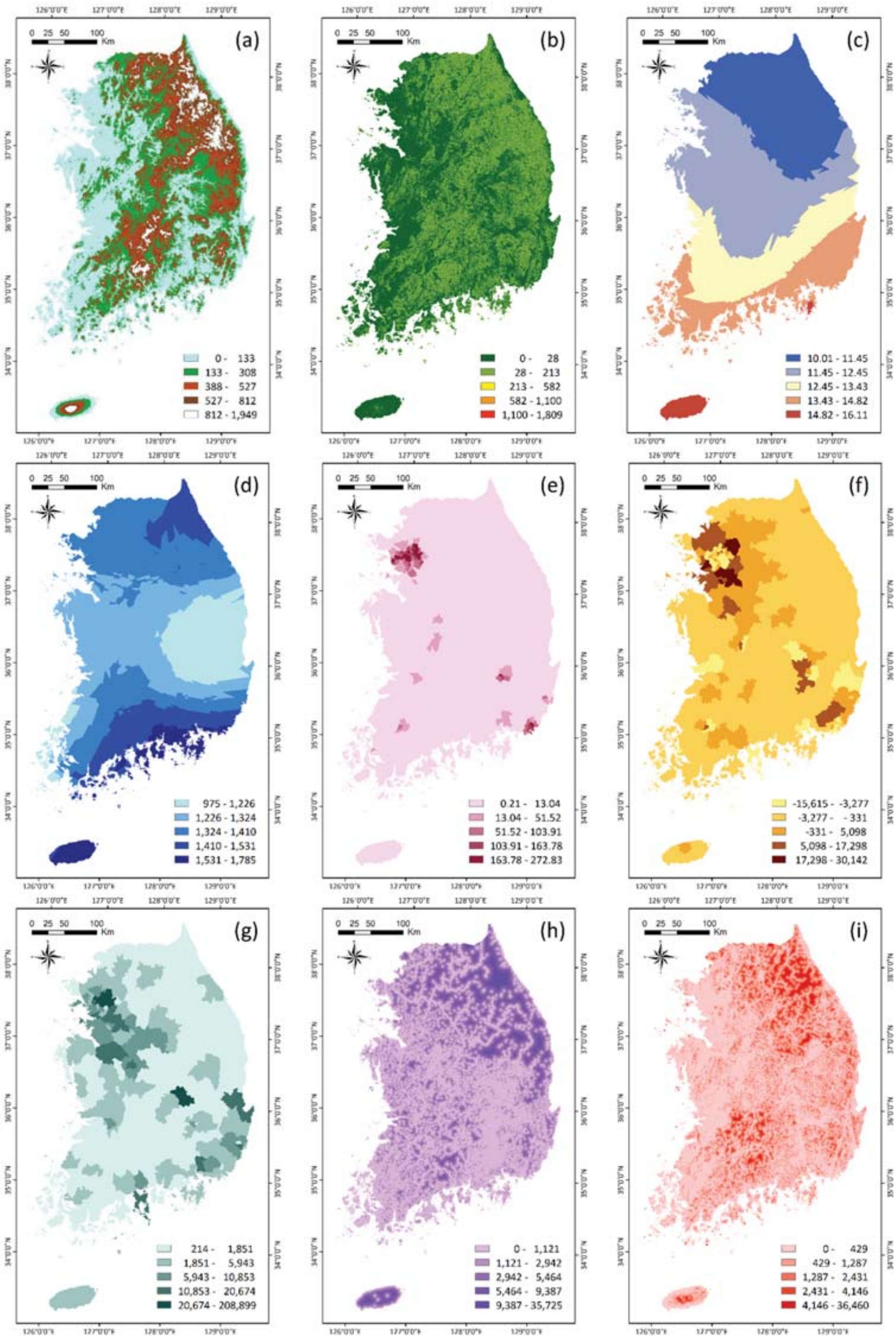

FIGURE 2 Driving factors of urban land expansion classified by natural breaks: (a) elevation; (b) slope; (c) temperature; (d) precipitation; (e) population density; (f) net migration; (g) gross regional domestic product; (h) distance to main road; (i) distance to built-up area; (j) distance to water body; (k) density of agriculture land; (l) density of forest land; (m) environmental conservation zone; and (n) legal protection zone

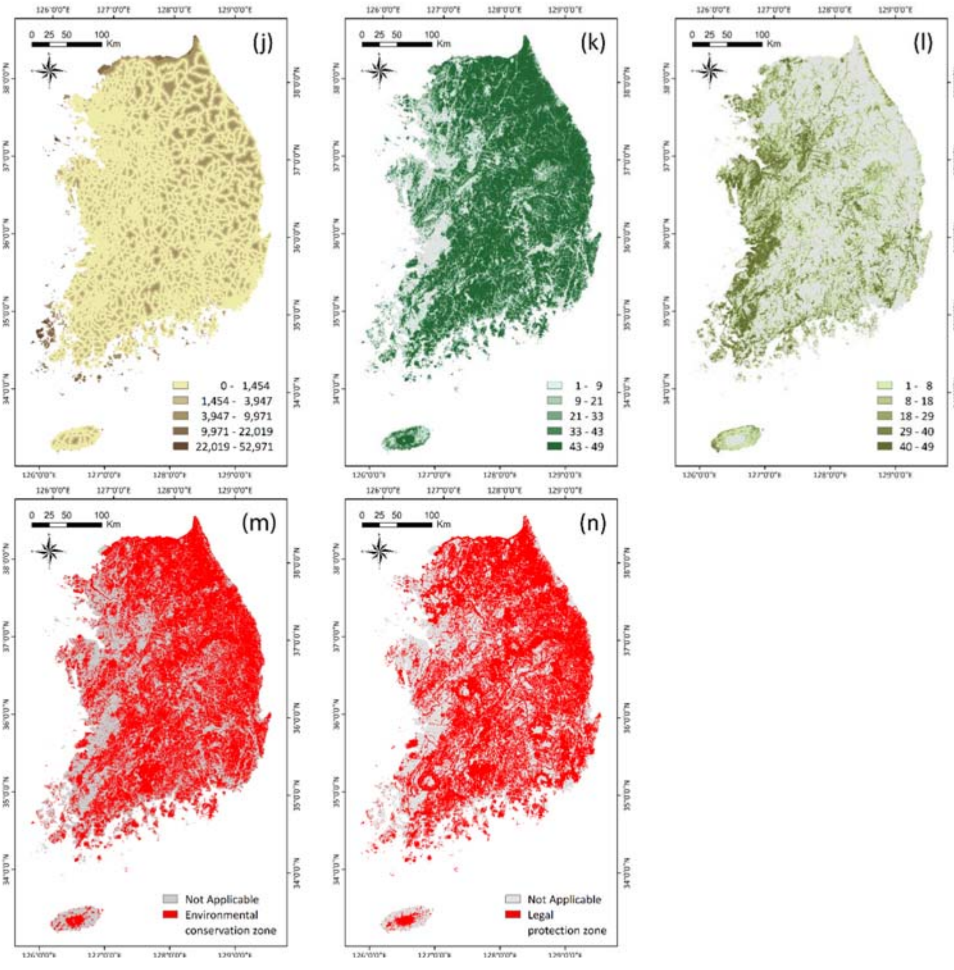

FIGURE 2 (Continued)

3.1.2 | Socioeconomic factors

Census-based socioeconomic factors, such as population, migration, employment rate, and GDP, are significant determinants in the rates and spatiotemporal patterns of urban growth (Wu & Zhang, 2012).

Six variables (three census-based and three accessibility variables) were selected to represent the socioeconomic factors. The census-based variables were based on city and county data, such as population, net migration, and gross regional domestic product (GRDP), and were collected using the statistical yearbook of Korea's National Statistical Office. The distances to main roads, existing urban areas, and water bodies were selected as accessibility variables. Here, urban areas are defined as built-up areas including residential, industrial, and commercial areas, structures related to transportation, and roads. Main roads including highways, national-level roads, and province-level roads were extracted from road maps produced by Korea's Ministry of Land, Infrastructure and Transport in 2000. The existing urban area and water bodies were obtained using a land-cover map produced by Korea's Ministry of Environment. The accessibility variables were calculated as the Euclidean nearest distance using Spatial Analyst in ArcGIS 10.1.

3.1.3 | Neighborhood factors

Forces, both centrifugal and centripetal, focused on the city's downtown are responsible for spatial autocorrelation in patterns of land use. Cities can be regarded as self-organizing systems (Batty & Longley, 1994; Verburg, van Eck, de

Nijs, Dijst, & Schot, 2004; Poelmans & Van Rompaey, 2010). It follows that the development of land use will depend largely on conditions of neighborhood land use, while the orientation of neighborhood variables is usually toward density (White & Engelen, 1997; Wu & Yeh, 1997; Cheng & Masser, 2003). The neighborhood variables can indicate the availability of land for development, or possible constraints to new development (Luo & Wei, 2009).

The amount of agricultural and forest land was used as a proxy of land-use conditions. A 7×7 pixel window with a radius of about 100 m was used to define the neighborhood. Distance–decay functions and practices that had been used in other studies underpinned this choice (Cheng & Masser, 2003; Verburg et al., 2004; Luo & Wei, 2009). The neighborhood factors were generated using the neighborhood statistic operation in ArcGIS 10.1.

3.1.4 | Land-use policy factors

Urban growth has often been controlled and restrained by land-use policies that designate specified areas as protected (e.g. water source protection areas, natural parks, wildlife protection areas, greenbelt zones, agricultural development regions, etc.) at the national or regional level. Therefore, land-use policy factors were evaluated and used as land-use determinants.

Environmental conservation and legal protection zones were selected to represent land-use policy factors. These variables were obtained using an environmental conservation value assessment map produced by Korea's Ministry of the Environment. This map was used to evaluate the physical and environmental value of land using 8 items related to environmental regulation and 57 items related to legal regulation. In addition, this map highlighted environmental conservation and legal protection zones on a grade of 1 to 5. The first priority areas, which were assigned to the first and second classes on the map, were coded as 1 and other grades were coded as 0.

3.2 | Data sampling

Fourteen variables were used as explanatory variables in the later statistical analysis. Urban expansion spatial patterns between 2000 and 2010, as shown in a binary map, formed the dependent variable. Land-cover maps obtained from the Korea Ministry of Environment were used and reclassified using values of 0 and 1. A value of 1 indicated that in a non-urban cell, the land use had changed to urban between 2000 and 2010, whereas cells that had already been open in 2000, or that did not change use to urban from non-urban between 2000 and 2010, were assigned a value of 0.

It was not possible to handle and analyze such a large dataset using standard statistical software. Also, there was spatial autocorrelation among variables, both dependent and explanatory (Li et al., 2013). A combined systematic and random sampling was conducted to minimize the influence of spatial autocorrelation. The points were extracted regularly, with an interval of 10 pixels (300 m). From this result, we selected 39,126 points coded as 1, and then randomly selected the same number of points coded as 0. This was necessary so that estimates of the model coefficient β_j are not affected by unequal sampling rates, but the intercept α is (Allison, 1999). Consequently, there was a total of 78,252 sample points. Spatial overlay was used for observations at each point of the explanatory variables, and also in the statistical analysis.

3.3 | LR model

The LR statistical modeling technique is in widespread use as a way of finding empirical relationships between independent continuous and categorical variables and a binary dependent variable (McCullagh & Nelder, 1989). The assumption underlying the technique is that there is a logistic curve that indicates how likely a dependent variable is to have a value of 1 (positive response), and that the curve's value can be ascertained by use of the following formulae (Mahiny & Turner, 2003; Arsanjani, Helbich, Kainz, & Boloorani, 2013):

$$\mathrm{P}\ (Y=1 \mid x_1, x_2, \ldots, x_m)=\exp\left(\beta_0+\sum\beta_i x_i\right)/\left(1+\exp\left(\beta_0+\sum\beta_i x_i\right)\right) \tag{1}$$

$$\text{logit P}\ (Y=1 \mid x_1, x_2, \ldots, x_m) = \ln\ (p/(1-p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \tag{2}$$

where P is the probability of the dependent variable; x_i $(i=1, 2, \ldots, m)$ are the independent variables, β_0 is a constant, and β_i $(i=1, 2, \ldots, m)$ is a vector of the coefficients of the estimated parameter.

The model in Equation (1) was linearized using Equation (2) (the standard model for linear regression) and the logit transformation was the means of removing 0/1 boundaries for the original dependent variable. Transforming binary data through a logit transformation allows certainty of a continuous dependent variable; the dependent variable newly created by the logit transformation of the probability is boundless. This approach also ensures a continuous probability surface in the range 0 to 1 (Arsanjani et al., 2013).

The coefficients of the estimated parameters are usually determined using a maximum likelihood (ML) estimation as a convergence criterion. After standardizing the variables, a β_i value so created may show how each independent variable relatively influences the dependent variable. The impact on logit P will increase in line with the absolute value of β_i (Menard, 2004; Shu, Zhang, Li, Qu, & Chen, 2014).

All of the operations for the LR analysis were executed using SPSS® Statistics software version 21.0 (IBM-SPSS, Chicago, IL, USA), and the probability for all of the pixels in the research area was then calculated within a GIS.

3.4 | LCR model

LCR analysis tests in a single analysis the differential relationships across a number of latent classes between predictor and output, thereby combining the strengths of regression models and cluster analysis (Wedel & DeSarbo, 1994). In concept, by combining a number of independent variables to predict the dependent variable, LCR resembles multiple regression (Garver, Williams, & Taylor, 2008).

An LCR model of this sort used repetitively resembles other models which deal with the dependent observation problem by including random effects: multilevel (two-level), mixed, and random-coefficient models. In fact, LCR is a random effects model that is not parametrically governed (Aitkin, 1999; Vermunt & Van Dijk, 2001; Agresti, 2002; Skrondal & Rabe-Hesketh, 2004). ML is the usual method for estimating LCR, and the following equation illustrates the likelihood contribution of a level-two unit j:

$$f(Y_j|X_j, W_j) = \sum_{k=1}^{k} \pi_k f_k\ (Y_j|X_j, W_j) = \sum_{k=1}^{k} \pi_k \prod_i f_k(y_{ij}|X_j, W_j) \tag{3}$$

where k is the number of latent classes and $f_k(y_{ij}|X_j, W_j)$ is a class-specific density. Any exponential function can constitute this density (Vermunt & van Dijk, 2001). The expectation maximization (EM) algorithm is used more often than any other to solve the ML estimation problem.

Among the various software packages, latent GOLD 5.0 software (Statistical Innovation, Belmont, MA, USA) was used to estimate the LCR models in this study. Latent GOLD 5.0 begins with a series of EM iterations; a small relative change in parameters will cause it to transfer to the Newton–Raphson method. Multiple sets of random starting values allow the avoidance of local optima.

4 | RESULTS AND DISCUSSION

4.1 | LR analysis

A multicollinearity test was performed to avoid a failure to converge, using a tolerance (TOL) and a variance inflation factor (VIF), as is standard for multicollinearity diagnosis. After a multicollinearity test, values of TOL $<$ 0.1 and VIF $>$ 10 were considered to indicate serious multicollinearity between independent variables, and these variables are excluded from the LR analysis (Ozdemir, 2011). The result of the multicollinearity analysis confirmed that for all variables the TOL was larger than 0.1, and the VIF was smaller than 10. These results indicated that there was no significant collinearity among the independent variables used in the analysis; thus, LR analysis could be performed, with the inclusion of all the independent variables (Table 3).

T3

TABLE 3 Multicollinearity diagnosis indexes for the independent variables used in the analysis

	Collinearity statistics			Collinearity statistics	
Variables	TOL[a]	VIF[b]	Variables	TOL[a]	VIF[b]
Ele	0.341	2.934	DistRoad	0.902	1.108
Slo	0.408	2.451	DistUrban	0.476	2.103
Temp	0.734	1.362	DistWater	0.637	1.570
Rain	0.848	1.179	DenFarm	0.430	2.323
Pop	0.445	2.246	DenFore	0.237	4.221
Mig	0.918	1.089	Environ	0.452	2.212
Grdp	0.486	2.059	Law	0.628	1.592

[a]Tolerance.
[b]Variance inflation factor.

The fit of a logistic model with a dataset can be evaluated using pseudo R^2 measures. The pseudo R^2 value, which indicates logit model/dataset fit, ranges from 0 (no relationship) to 1 (perfect fit). A value greater than 0.2 for the pseudo R^2 shows a relatively good fit (Clark & Hosking, 1986; Menard, 2002). The value of the pseudo R^2 in this study was 0.385. The regression function's coefficients permit an assessment of the independent variables' relative importance.

That independent variables can explain dependent variables was shown by the value of the Cox and Snell R^2 (0.414) and the Nagelkerke R^2 (0.385); dependent variables in these instances explained, respectively, 41.4% and 38.5% of the variance. In addition, the predicted accuracy was 87.3% for urban growth and 73.1% for non-urban growth. The overall predicted accuracy was 80.5% (Table 4).

All independent variables except the elevation (Ele), GRDP (Grdp), and distance to a water body (DistWater) were significant at the 0.05 level. The factors of Ele, precipitation (Rain), net migration (Mig), Grdp, distance to a main road (DistRoad), and DistWater had a positive effect on urban growth. Rain made the highest contribution. Thus, these results indicated that urban growth in Korea was largely dependent on Mig and reflect the characteristics of urban growth during the period from 2000 to 2010, a period of development of satellite towns and planned city-based new towns. DistRoad was significantly related to the probability of change. This is presumably because there are two types of urban development in Korea: (1) expansion of existing urban areas and (2) new city developments planned by central or local governments in rural areas. In addition, although it was not significant at the 0.05 level, a positive value of Ele still reflects policy change in South Korea. In the late 2000s, the regulations regarding the height of developments in mountainous areas were relaxed by land-use deregulation (Table 5).

TABLE 4 Model summary statistics

Statistics	Value
−2 Log(likelihood) of initial	108,480.306
−2 Log(likelihood) of final	66,721.726[a]
Cox and Snell R^2	0.414
Nagelkerke R^2	0.551
Pseudo R^2	0.385
Percentage correctly predicted (PCP)[b]	80.500

[a]Estimation terminated at iteration number 6 because parameter estimates changed by less than 0.001.
[b]The cutoff value is 0.5.

TABLE 5 Logistic regression model results

	B[a]	S.E.[b]	Wald[c]	Df[d]	Sig.[e]	Exp(B)[f]	95% C.I. for Exp (B)[g]	
							Lower	Upper
Constant	2.891	0.143	406.367	1.000	0.000	18.017		
Ele	0.000	0.000	1.204	1.000	0.272	1.000	1.000	1.000
Slo	−0.002	0.001	4.099	1.000	0.043	0.998	0.996	1.000
Temp	−0.144	0.010	218.289	1.000	0.000	0.866	0.849	0.882
Rain	0.001	0.000	114.532	1.000	0.000	1.001	1.001	1.001
Pop	−0.001	0.001	4.703	1.000	0.030	0.999	0.997	1.000
Mig	0.000	0.000	439.776	1.000	0.000	1.000	1.000	1.000
Grdp	0.000	0.000	0.075	1.000	0.784	1.000	1.000	1.000
DistRoad	0.000	0.000	241.928	1.000	0.000	1.000	1.000	1.000
DistUrban	−0.001	0.000	1,289.414	1.000	0.000	0.999	0.999	0.999
DistWater	0.000	0.000	0.191	1.000	0.662	1.000	1.000	1.000
DenFarm	−0.019	0.001	455.077	1.000	0.000	0.981	0.980	0.983
DenFore	−0.069	0.001	4,497.367	1.000	0.000	0.934	0.932	0.935
Environ	−0.645	0.026	612.137	1.000	0.000	0.525	0.498	0.552
Law	−0.493	0.023	459.633	1.000	0.000	0.611	0.584	0.639

[a]Logistic coefficient.
[b]Standard error of estimate.
[c]Wald chi-square values.
[d]Degree of freedom.
[e]Significance.
[f]Exponential and coefficient.
[g]95% confidence interval for Exp(B).

However, the variables of slope (Slo), temperature (Temp), population density (Pop), distance to built-up area (DistUrban), density of agriculture land (DenFarm), density of forest land (Denfore), environmental conservation zone (Environ), and legal protection zone (Law) had negative effects on urban growth. Temp made the least contribution to urban growth. The results for Slo and DistUrban were not surprising, because urban growth is more easily facilitated in areas with a gradual slope and/or that are close to existing built-up areas. The fact that Pop had a positive effect reflects urban growth at the fringe of existing built-up areas. The results for DenFarm, DenFore, Environ, and Law indicate that Korea is protective of areas with a high conservation value, as well as major agricultural and forest land areas. This explains why development policies have aimed to implement eco-friendly urban development (Table 5).

4.2 | LCR model

Latent class models were also used to identify the optimal regression model describing the dataset of this research. This methodology starts with one class and increases the number of classes until the optimal model is found. Comparing the models' relative fit allows the optimal number of latent classes to be selected. The likelihood ratio chi-squared statistic (L^2), Akaike information criterion (AIC), and Bayesian information criterion (BIC) are most commonly used to measure goodness-of-fit. Being relative measures, these indices have no threshold value. The lower the value, the better the fit; lower values also indicate a more parsimonious model (McLachlan & Peel, 2000; Guerrero, Egea, &

TABLE 6 Model fitting statistics for latent class models

	LL	Difference[a]	BIC (LL)	Difference[a]	AIC (LL)	Difference[a]
1-class model	−33,360.863	—	66,890.741	–	66,751.726	—
2-class model	−30,686.287	2,674.576	61,721.872	5,168.869	61,434.574	5,317.152
3-class model	−29,970.315	715.972	60,470.211	1,251.661	60,034.629	1,399.945
4-class model	−29,771.523	198.792	60,252.910	217.301	59,669.045	365.584

[a]The value of difference refers to the difference between the value of each index for a particular model and that of the preceding model (e.g. 1-class model vs. 2-class model).

González, 2007). Sparse data indicates a greater suitability for AIC and BIC measures of goodness-of-fit; their use is indicated for comparisons and evaluations of models having different numbers of segments (Garver et al., 2008).

No explanatory variables were added in this step, and four models were identified (Table 6). The model fit gradually improved as the number of classes increased. Among the four models, although the four-class model had the lowest values of goodness-of-fit indices, the magnitudes of improvement were dramatically decreased after the three-class model, eventually reaching an asymptote. The three-class model was shown to provide the optimal representation of the data, and was the most parsimonious. Therefore, the LCR model was estimated with the three-class assumption.

The covariates and predictors had different roles in estimating the LCR model. The covariates affect the definition of the latent classes, whereas the predictors affect the dependent variable. There is no consensus in the literature to define exogenous variables for urban growth analysis. However, the spatial factors displayed the greatest variation among the explanatory variables, which also reflects the distances between samples, because the distance variables for closer samples had similar values. Therefore, it is possible to capture spatial homogeneity and heterogeneity in two ways: (1) the similarity between samples based on DistRoad, DistUrban, and DistWater; and (2) the distance between samples. Therefore, these factors were used as covariates as well as predictors.

Classes 1–3 consisted of 38.36% ($n = 15{,}009$), 31.12% ($n = 12{,}176$), and 30.52% ($n = 11{,}941$), respectively. Each class accounted for a similar proportion of the total, although the one-class LCR model made the largest contribution. With regard to the covariates for each class, the two-class LCR model had the largest amount of urban growth, while the three-class LCR model had the least. The one-class LCR model was similar to the two-class LCR model. The amount of urban growth for the one-class LCR model was lower than that for the two-class LCR model, except for DistWater. The value of the DistWater coefficient was higher than in the two-class LCR model (Figure 3).

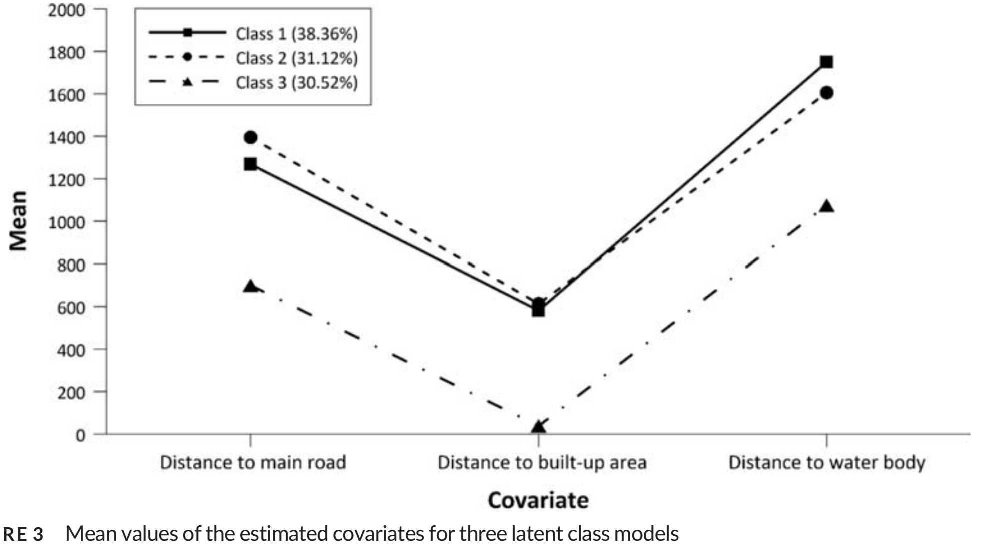

FIGURE 3 Mean values of the estimated covariates for three latent class models

TABLE 7 Latent class regression results for each class

	Class 1			Class 2			Class 3			Between class			
	B[a]	z Stat.[b]	Sig.[c]	B[a]	z Stat.[b]	Sig.[c]	B[a]	z Stat.[b]	Sig.[c]	Wald (=)[d]	Sig.[c]	Mean	Std. dev.[e]
Constant	2.704*	5.504	0.000	0.298	0.410	0.682	−8.306*	−4.302	0.000	32.126	0.000	−1.406	4.682
Ele	−0.001	−3.461	0.001	0.000	0.580	0.562	−0.001	−0.712	0.476	3.637	0.160	−0.001	0.001
Slo	−0.010*	−4.078	0.000	0.002	0.372	0.710	0.050*	2.776	0.005	14.070	0.001	0.012	0.026
Temp	−0.430*	−13.234	0.000	0.136*	2.405	0.016	0.054	0.511	0.609	68.679	0.000	−0.106	0.258
Rain	0.003*	7.622	0.000	0.000	0.625	0.532	0.001	0.892	0.372	10.921	0.004	0.001	0.001
Pop	0.019*	5.278	0.000	−0.003	−0.842	0.400	0.003	0.410	0.682	13.826	0.001	0.007	0.010
Mig	0.000*	9.108	0.000	0.000*	7.491	0.000	0.000*	2.107	0.035	0.686	0.710	0.000	0.000
Grdp	0.000*	8.011	0.000	0.000	−1.258	0.208	0.000	−0.216	0.829	60.370	0.000	0.000	0.000
DistRoad	0.000	−0.589	0.556	−0.018*	−11.668	0.000	0.000	0.714	0.475	136.984	0.000	−0.006	0.008
DistUrban	0.000*	−2.954	0.003	0.000*	2.822	0.005	0.452*	13.646	0.000	197.252	0.000	0.138	0.208
DistWater	0.000*	−2.973	0.003	0.000	−0.430	0.667	0.000	0.048	0.962	1.753	0.420	0.000	0.000
DenFarm	0.071*	11.616	0.000	−0.052*	−9.786	0.000	−0.094*	−5.979	0.000	246.606	0.000	−0.017	0.072
DenFore	−0.067*	−21.861	0.000	−0.030*	−5.144	0.000	−0.159*	−9.614	0.000	60.475	0.000	−0.083	0.052
Environ	−0.524*	−13.665	0.000	−0.239*	−3.022	0.003	0.063	0.372	0.710	17.139	0.000	−0.256	0.242
Law	−0.580*	−15.365	0.000	−0.171*	−2.660	0.008	−0.288*	−2.423	0.015	25.369	0.000	−0.363	0.177
R^2	0.738			0.532			0.847						

*Significant at the 5% level.
[a]Latent class regression coefficient.
[b]z Statistics.
[c]Significance.
[d]Wald chi-square values.
[e]Standard deviation.

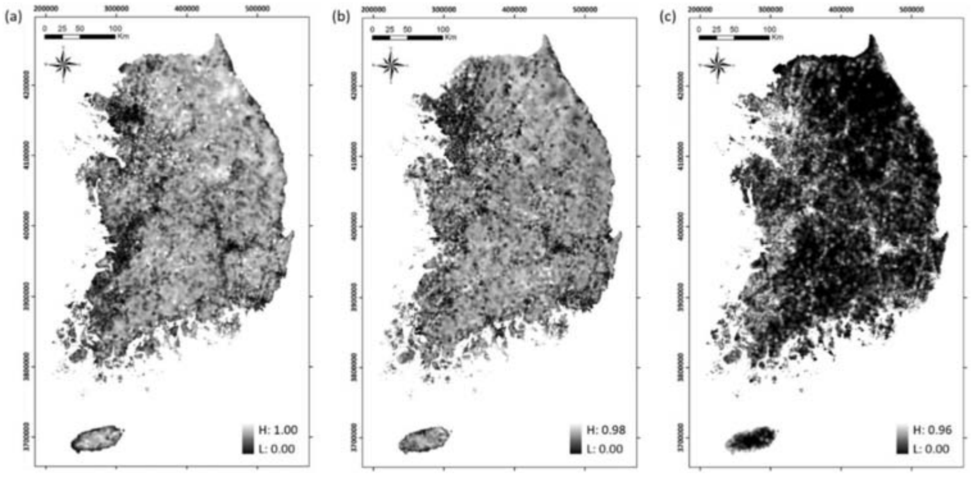

FIGURE 4 Spatial distribution of probability values for classifying as each cluster: (a) class 1; (b) class 2; and (c) class 3

The estimated parameters for the three-class LCR models are shown in Table 7 and Figure 4. From the Wald test statistics, our results indicated that all of the predictors had different coefficients among the latent groups, except for Ele, Mig, and DistWater. The coefficients of DistRoad, DistUrban, and DenFarm were markedly different for each class of LCR model. For example, the smallest unit contribution for DenFarm was found in the two-class model, but the largest was found in the three-class representation. In addition, the directional effects of the coefficients, such as positive or negative coefficients, were different depending on the class membership.

In the one-class LCR model, all of the predictors except for DistRoad were significant. The R^2 was slightly higher (0.7383). The number of significant predictors was the largest among the three classes of LCR models. The next largest number of significant predictors was found in the two-class LCR model, with the lowest R^2 value (0.5324). The three-class LCR model was estimated, with an R^2 value of 0.8473, which was larger than for the other classes. However, fewer significant predictors were found with the six predictor variables. This is presumably related to the model having the smallest sample size among the three classes of LCR models (Table 7). In addition, the spatial variables played important roles as covariates in this model. Based on Wald statistics, DistUrban was the most important variable for classifying the latent classes (Table 8). This result can also be verified by the data presented in Figure 3, which shows a chart of mean covariates.

TABLE 8 Class-specific coefficients of covariates

	Class 1			Class 2			Class 3				
	B[a]	z-Stat.[b]	Sig.[c]	B[a]	z-Stat.[b]	Sig.[c]	B[a]	z-Stat.[b]	Sig.[c]	Wald[d]	p Value
Constant	−1.308*	−37.711	0.000	−1.499*	−40.149	0.000	2.806*	45.884	0.000	2106.966	3.0e−458
DistRoad	0.000	−0.941	0.347	0.000	2.795	0.005	0.000	−0.825	0.409	8.059	0.018
DistUrban	0.019*	31.559	0.000*	0.019*	31.251	0.000	−0.037*	−31.431	0.000	1001.833	0.000
DistWater	0.000	1.418	0.156	0.000*	−4.438	0.000	0.000	1.822	0.069	20.930	0.000

*Significant at the 5% level.
[a]Latent class regression coefficient.
[b]z Statistics.
[c]Significance.
[d]Wald chi-square values.

4.3 | Comparison of the results between LR and LCR

LCR offers methodological benefits compared with LR in terms of discrete random effects, the flexibility of dependent variables, the derivation of unique regression models for each segment, and more accurate results. Garver et al. (2008) gave a detailed review of the advantages of the LCR model. A brief summary is as follows.

First, a coefficient b may be assumed to take value $b1$, probability $p1$, in the LCR model. A discrete distribution is thereby indicated for parameter b, yielding a random-effects, nonparametric modeling approach (Garver et al., 2008). It follows that the LCR model is less intensive, computationally, than parametric models, and that interpreting the results is simplified thanks to the greater consistency of the LCR model with the multiple regression output (Vermunt & Van Dijk, 2001).

Second, the LR model usually accommodates dependent variables, with categorical binominal counts. The dependent variable in the LCR model, though, has greater flexibility since it can be continuous or show Poisson or categorical binomial counts (Vermunt & Magidson, 2005). The resulting flexibility being greater than for multiple regression means that models can be tested which are more appropriate for the investigators' data (Garver et al., 2008).

Third, from a conceptual viewpoint, beta coefficients across respondents in the data are averaged by multiple regression models, but homogeneous segments are identified by the LCR model, which derives regression models that are unique for each one. For each respondent, probabilities will be provided and a determination made of their likelihood for segment membership (Vermunt & Magidson, 2005; Garver et al., 2008). In this study, the probabilities of the LR and the LCR model were shown to be different in terms of spatial distribution (Figure 5). Therefore, the 14 variables affecting urban growth operated in a different way, depending on the underlying spatial structures. This shows the necessity of using the LCR model, which is able to reflect the heterogeneous nature of the geography.

Finally, the results of the LCR model were more accurate than those of the LR model, as shown by testing using AIC, BIC, and the receiver operating characteristic (ROC). The values of AIC and BIC for the LCR model were 6,279.515, which were 6,755.606 lower than for the LR model. Also, the ROC values for the LR and LCR models were

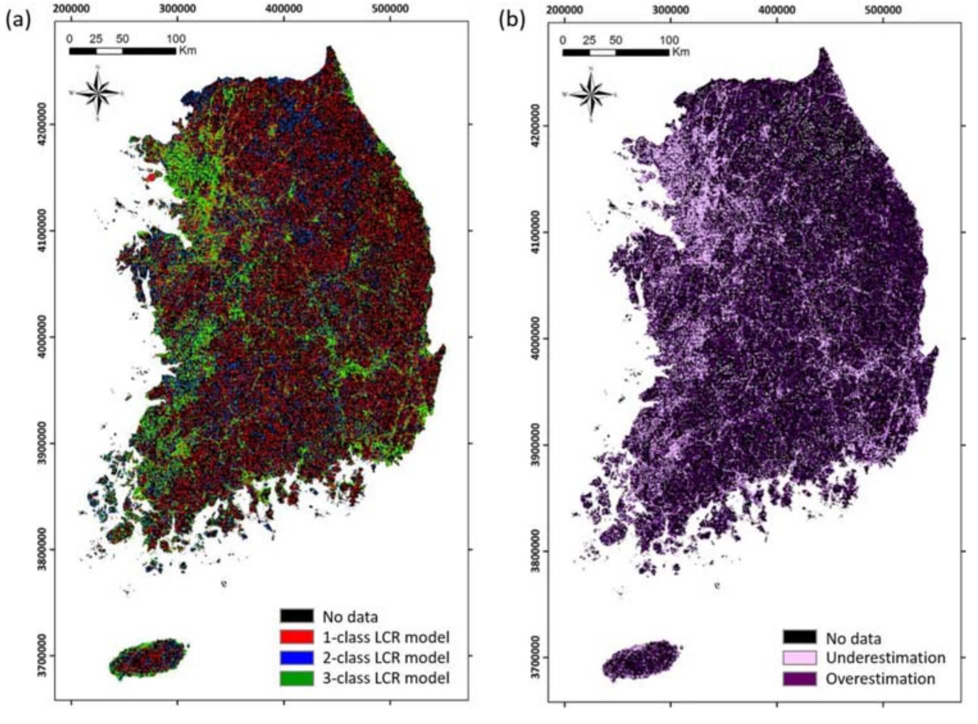

FIGURE 5 Spatial distribution of each class through the latent class regression model (a) and differences of probabilities between the logistic regression and latent class regression models (b)

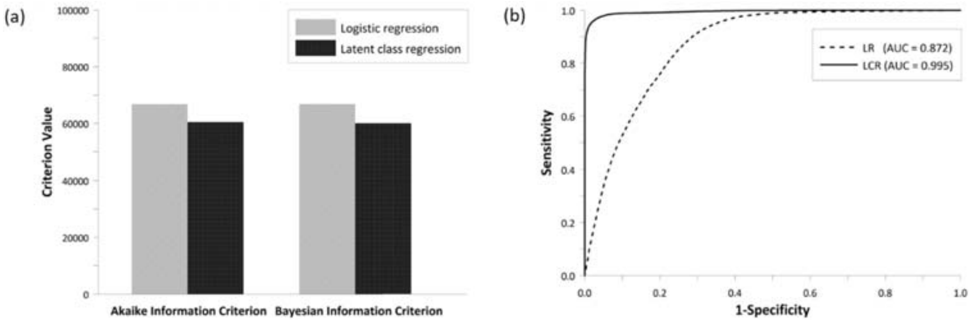

FIGURE 6 Comparison of the results between the two models using statistical indexes: (a) Akaike information criterion and Bayesian information criterion; and (b) receiver operating characteristic

0.872 and 0.995, respectively. The interpretability of the model was improved using the LCR model, because the ROC value was 0.123 higher than that of the LR model (Figure 6).

5 | CONCLUSIONS

The aim of this study was to achieve an analysis of the underlying factors of urban growth in Korea from 2000 to 2010, including their spatial patterns. LCR analysis was used and the results were compared to those using LR analysis. Fourteen independent variables were selected (through a literature review).

The results of the LR and LCR analysis were represented differently in terms of the magnitude and directional effects of coefficients. The results indicated that the different segments had different predictor relationships for urban growth in the study area. These results suggest that spatial non-stationarity has an important role in analyzing urban growth patterns. Most importantly, the LCR analysis could provide insight into the spatial variations of urban growth patterns. The LCR analysis can significantly improve the LR through an analysis using statistical indexes. The LCR analysis has a much better goodness-of-fit with lower values of AIC and BIC. Also, the LCR analysis can be seen to have done a better job of interrogating the relationships between independent variables and urban growth than the LR, since ROC values were 0.123 higher than for the LR analysis.

There are, however, a number of limitations to this study. First, because the independent variables used were selected through literature review, they were not optimized to reflect the exact urban growth patterns in the study area. As a result, although the three-class LCR model had the highest R^2 value, only 6 of the 14 independent variables were significant. Second, the spatial factors were used as covariates in the consideration of variation. The methodology used to select suitable covariates should be given further consideration. Third, 78,252 data points from a total of 1,116,194 were sampled and used to perform the spatial statistical analysis. Therefore, the results may not be representative of the entire study area. In particular, they do not capture the strong nature of spatial autocorrelation in the data that implies new urbanization is usually at the edge of existing urban areas.

The conceptual similarity between LCR analysis and multiple regression means that there are no great difficulties in interpreting the results of LCR. Therefore, LCR can be used easily and may become a powerful tool for the analysis of urban growth patterns. Future work should determine the definite advantages and disadvantages of LCR analysis by undertaking a comparative study using various statistical methodologies and more data. In addition, the results of this study could be used as a probability map for further urban growth modeling, and the predicted results should be compared.

ORCID

Jae Hyun Lee http://orcid.org/0000-0003-4007-7048

REFERENCES

Agresti, A. (2002). *Categorical data analysis*. New York, NY: Wiley.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*(1), 117–128.

Allen, J., & Lu, K. (2003). Modeling and prediction of future urban growth in the Charleston region of South Carolina: A GIS-based integrated approach. *Ecology & Society*, *8*(2), 2.

Allison, P. D. (1999). *Logistic regression using SAS: Theory and application*. Cary, NC: SAS Institute.

Arsanjani, J. J., Helbich, M., Kainz, W., & Boloorani, A. D. (2013). Integration of logistic regression, Markov chain and cellular automata models to simulate urban expansion. *International Journal of Applied Earth Observation & Geoinformation*, *21*, 265–275.

Batty, M., & Longley, P. A. (1994). *Fractal cities: A geometry of form and function*. London, UK: Academic Press.

Cheng, J., & Masser, I. (2003). Urban growth pattern modeling: A case study of Wuhan city, PR China. *Landscape & Urban Planning*, *62*(4), 199–217.

Clark, W. A., & Hosking, P. L. (1986). *Statistical methods for geographers*. New York, NY: Wiley.

Dendoncker, N., Rounsevell, M., & Bogaert, P. (2007). Spatial analysis and modelling of land use distributions in Belgium. *Computers, Environment & Urban Systems*, *31*(2), 188–205.

Foody, G. M. (2003). Geographical weighting as a further refinement to regression modelling: An example focused on the NDVI–rainfall relationship. *Remote Sensing of Environment*, *88*(3), 283–293.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester, UK: Wiley.

Garver, M. S., Williams, Z., & Taylor, G. S. (2008). Employing latent class regression analysis to examine logistics theory: An application of truck driver retention. *Journal of Business Logistics*, *29*(2), 233–257.

Guerrero, M. M., Egea, J. M. O., & González, M. V. R. (2007). Application of the latent class regression methodology to the analysis of Internet use for banking transactions in the European Union. *Journal of Business Research*, *60*(2), 137–145.

Hu, Z., & Lo, C. P. (2007). Modeling urban growth in Atlanta using logistic regression. *Computers, Environment & Urban Systems*, *31*(6), 667–688.

Lambin, E. F. (1997). Modelling and monitoring land-cover change processes in tropical regions. *Progress in Physical Geography*, *21*(3), 375–393.

LeSage, J. P. (2001). A family of geographically weighted regression models. In L. R. Anselin, J. G. M. Florax, & S. J. Rey (Eds.), *Advances in spatial econometrics* (pp. 240–264). Berlin, Germany: Springer-Verlag.

Li, X., Zhou, W., & Ouyang, Z. (2013). Forty years of urban expansion in Beijing: What is the relative importance of physical, socioeconomic, and neighborhood factors? *Applied Geography*, *38*, 1–10.

Liu, H., & Zhou, Q. (2005). Developing urban growth predictions from spatial indicators based on multi-temporal images. *Computers, Environment & Urban Systems*, *29*(5), 580–594.

Long, Y., Gu, Y., & Han, H. (2012). Spatiotemporal heterogeneity of urban planning implementation effectiveness: Evidence from five urban master plans of Beijing. *Landscape & Urban Planning*, *108*(2), 103–111.

Luo, J., & Wei, Y. D. (2009). Modeling spatial variations of urban growth patterns in Chinese cities: The case of Nanjing. *Landscape & Urban Planning*, *91*(2), 51–64.

Luo, J., Yu, D., & Xin, M. (2008). Modeling urban growth using GIS and remote sensing. *GIScience & Remote Sensing*, *45* (4), 426–442.

Mahiny, A. S., & Turner, B. J. (2003). Modeling past vegetation change through remote sensing and GIS: A comparison of neural networks and logistic regression methods. In *Proceedings of the 7th International Conference on GeoComputation*. Southampton, UK.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London, UK: Chapman & Hall.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Menard, S. (2002). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage Publications.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *American Statistician*, *58*, 218–223.

Millington, J. D., Perry, G. L., & Romero-Calcerrada, R. (2007). Regression techniques for examining land use/cover change: A case study of a Mediterranean landscape. *Ecosystems*, *10*(4), 562–578.

Overmars, K. P., & Verburg, P. H. (2005). Analysis of land use drivers at the watershed and household level: Linking two paradigms at the Philippine forest fringe. *International Journal of Geographical Information Science*, *19*(2), 125–152.

Ozdemir, A. (2011). Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). *Journal of Hydrology*, *405*, 123–136.

Park, J., Kim, D., Ko, Y., Kim, E., & Park, K. (2010). *Urbanization and urban policies in Korea*. Gyeongi-do, South Korea: Korea Research Institute for Human Settlements.

Partridge, M. D., Rickman, D. S., Ali, K., & Olfert, M. R. (2008). The geographic diversity of U.S. non-metropolitan growth dynamics: A geographically weighted regression approach. *Land Economics*, *84*(2), 241–266.

Poelmans, L., & Van Rompaey, A. (2010). Complexity and performance of urban expansion models. *Computers, Environment & Urban Systems*, *34*(1), 17–27.

Portnov, B. A., & Hare, A. P. (2012). *Desert regions: Population, migration and environment*. Heidelberg, Germany: Springer.

Scott, L. M., & Janikas, M. V. (2010). Spatial statistics in ArcGIS. In M. Fischer & A. Getis (Eds.), *Handbook of applied spatial analysis* (pp. 27–41). Heidelberg, Germany: Springer.

Shafizadeh-Moghadam, H., & Helbich, M. (2015). Spatiotemporal variability of urban growth factors: A global and local perspective on the megacity of Mumbai. *International Journal of Applied Earth Observation & Geoinformation*, *35*, 187–198.

Shu, B., Zhang, H., Li, Y., Qu, Y., & Chen, L. (2014). Spatiotemporal variation analysis of driving forces of urban land spatial expansion using logistic regression: A case study of port towns in Taicang City, China. *Habitat International*, *43*, 181–190.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. London, UK: Chapman & Hall.

Su, S., Xiao, R., & Zhang, Y. (2012). Multi-scale analysis of spatially varying relationships between agricultural landscape patterns and urbanization using geographically weighted regression. *Applied Geography*, *32*(2), 360–375.

United Nations, Department of Economic and Social Affairs, Population Division (2014). *World Urbanization Prospects* (2014 Rev., CD-ROM Ed.). New York, NY: United Nations, Department of Economic and Social Affairs, Population Division.

Verburg, P. H., van Eck, J. R. R., de Nijs, T. C., Dijst, M. J., & Schot, P. (2004). Determinants of land-use change patterns in the Netherlands. *Environment and Planning B*, *31*(1), 125–150.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Van Dijk, L. (2001). A nonparametric random-coefficients approach: The latent class regression model. *Multilevel Modelling Newsletter*, *13*(2), 6–13.

Wang, N., Brown, D. G., An, L., Yang, S., & Ligmann-Zielinska, A. (2013). Comparative performance of logistic regression and survival analysis for detecting spatial predictors of land-use change. *International Journal of Geographical Information Science*, *27*(10), 1960–1982.

Wedel, M., & DeSarbo, W. S. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Eds.), *Advanced methods of marketing research* (pp. 352–388). Cambridge, UK: Blackwell.

Wheeler, D. C., & Páez, A. (2010). Geographically weighted regression. In M. M. Fischer & A. Getis (Eds.), *Handbook of applied spatial analysis: Software tools, methods and applications* (pp. 461–486). Heidelberg, Germany: Springer.

White, R., & Engelen, G. (1997). Cellular automata as the basis of integrated dynamic regional modelling. *Environment & Planning B*, *24*(2), 235–246.

Wu, B., Huang, B., & Fung, T. (2009). Projection of land use change patterns using kernel logistic regression. *Photogrammetric Engineering & Remote Sensing*, *75*(8), 971–979.

Wu, F., & Yeh, A. G. O. (1997). Changing spatial distribution and determinants of land development in Chinese cities in the transition from a centrally planned economy to a socialist market economy: A case study of Guangzhou. *Urban Studies*, *34*(11), 1851–1879.

Wu, K. Y., & Zhang, H. (2012). Land use dynamics, built-up land expansion patterns, and driving forces analysis of the fast-growing Hangzhou metropolitan area, eastern China (1978–2008). *Applied Geography*, *34*, 137–145.

How to cite this article: Park S, Lee J H, Clarke K C. Capturing the heterogeneity of urban growth in South Korea using a latent class regression model. *Transactions in GIS*. 2018;22:789–805. https://doi.org/10.1111/tgis.12451