Capturing the Heterogeneity of Urban Growth in South Korea using a Latent Class Regression Model

Soyoung Park, Jae Hyun Lee, Keith C. Clarke

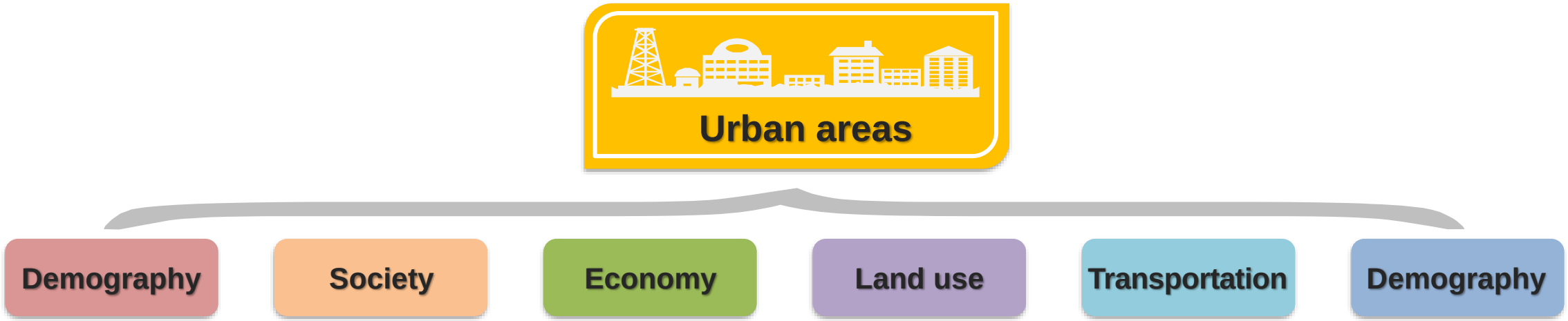
ESRI USER CONFERENCE

Contents

01 Introduction

Introduction

- Urban areas Include a multitude of interdependent measures that embed non-linear feedbacks.



- Spatial analytic methods are now increasingly able to answer following questions:

  Q1. What are these interrelations among the factors that contribute to urban change?

  Q2. Which of the causative relationships are predictable in time and space?

Introduction

Logistic Regression (LR)

- Easy adaptation for studying predictor variables in land change application
- Qualitative exploration of how urban growth and its causative factors interrelate
- Understanding about which are the most influential variables and how to distinguish among them
- Inability to reflect spatial non-stationarity because the global relationships generated show only an average condition

Geographically weighted regression (GWR)

- Making it possible to visualize geographical interaction through maps of coefficients
- More accurate and appropriate reflection on the true situation
- The lack of independence among local estimates, the presence of outliers, and the ineffectiveness of the estimated local coefficient due to the low sampling numbers

Latent class regression (LCR)

- Production of the functionality to identify spatially homogeneous areas (within each latent class) and heterogeneous area (between latent class)

Introduction

Objectives

- Improvement on the understanding of spatial patterns and of the factors underpinning urban growth using the LCR model

Contents

- Analysis of the relationship between dependent and independent variables using the LCR model
- Comparison of the results between the LR and LCR model using various statistical indicators

Question 01 How is the LCR model best applied to land change science?

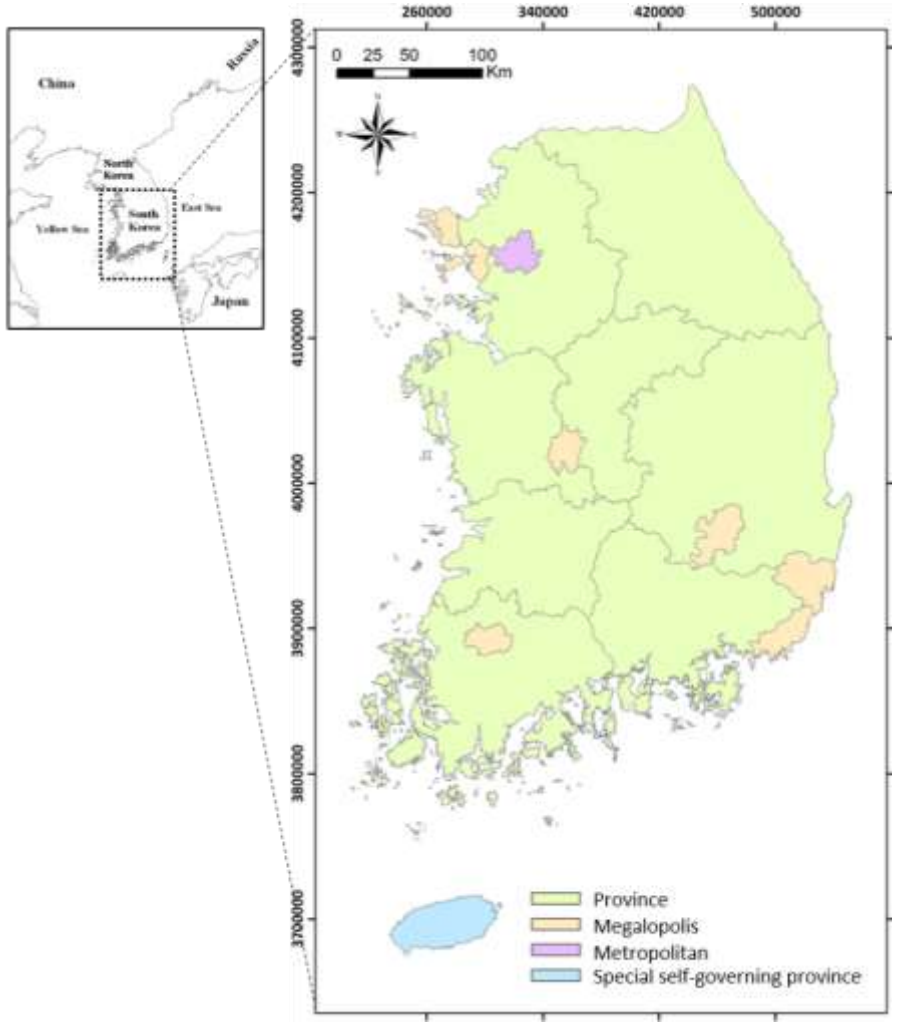
Question 02 How are the coefficients spatially different when using the LCR model?

Question 03 Does the LCR model outperform the LR model?

02

Study area

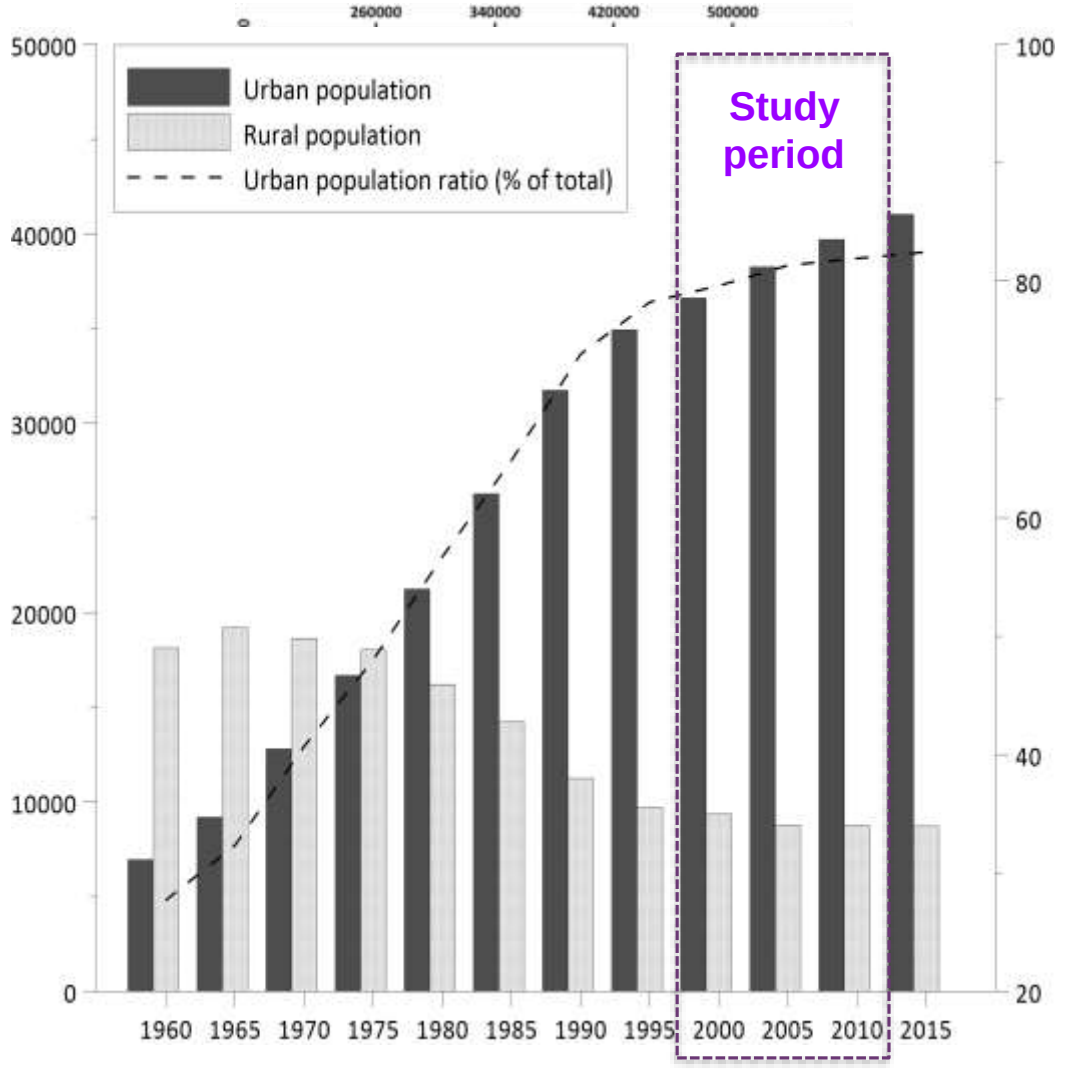
Study area



Location and status of the study area

- **Location**
  - ✓ Latitudes 33˚~39˚N, longitudes 124˚~130˚E
- **Administrative districts**
  - ✓ One special city
  - ✓ Six metropolitan cities
  - ✓ Eight provinces
  - ✓ One special self-governing province
- **Land use**
  - ✓ Mostly mountainous area (64% of the total land area)
- **Topographic conditions**
  - ✓ The west and south-east having the low land with an elevation averaging about 254 meters

Study area



Urban growth

- **Since 1960's**
  - ✓ Acceleration of urban growth by industrial-ization and economic growth
  - ✓ Experience of rapid rural-urban migration
- **Since 1995**
  - ✓ A pause in the trend for urban population to increase and rural population fall
- **Since the mid-1990s**
  - ✓ Population growth in nearby cities then than in major metropolitan cities
  - ✓ Decentralization of the urban society
  - ✓ Societal structural qualitative changes

03 Data and methodology

Data and methodology

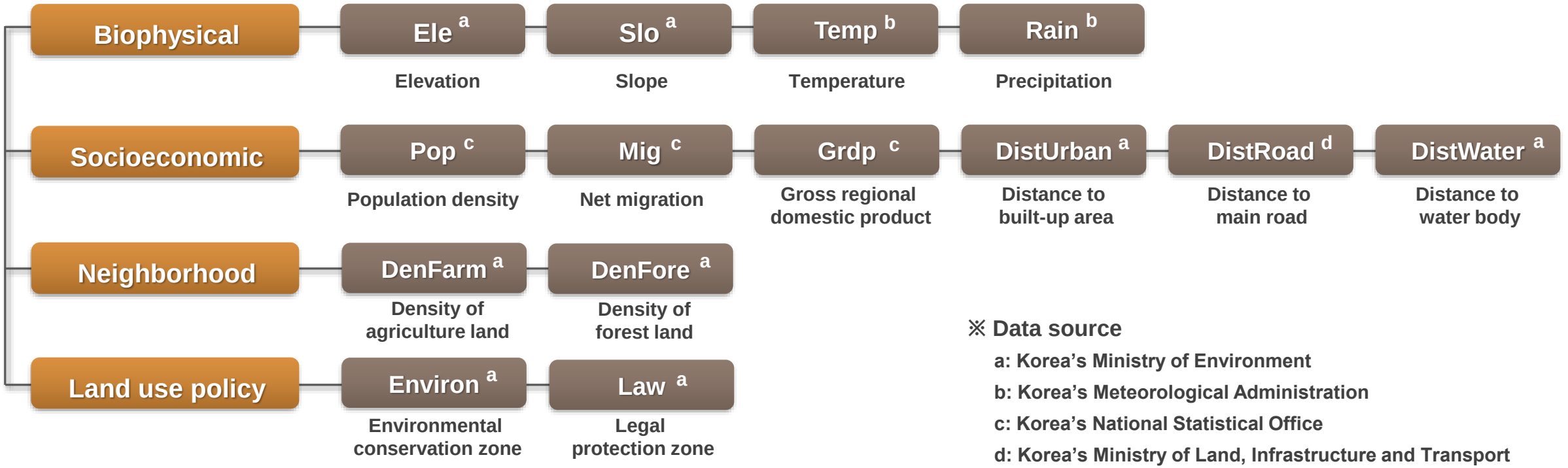
Data preparation

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

Data and methodology

Data preparation

- Production of spatial data **in grid raster format with a 30-m resolution using ArcGIS 10.5 software**
- Usage of **15 factors as explanatory variables** in the later statistical analysis



※ Data source

a: Korea's Ministry of Environment

b: Korea's Meteorological Administration

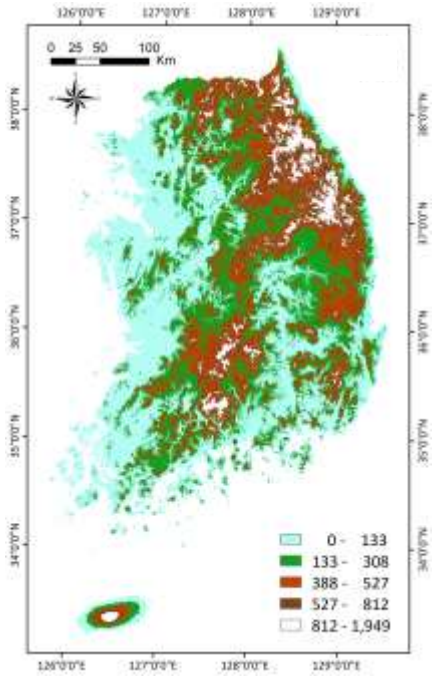
c: Korea's National Statistical Office

d: Korea's Ministry of Land, Infrastructure and Transport

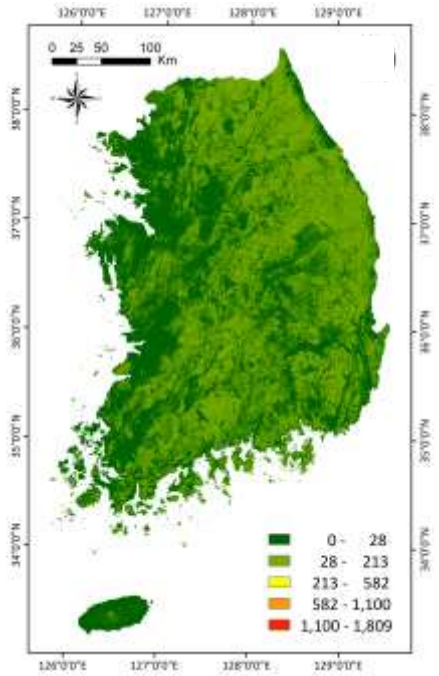
Data and methodology

Biophysical factors

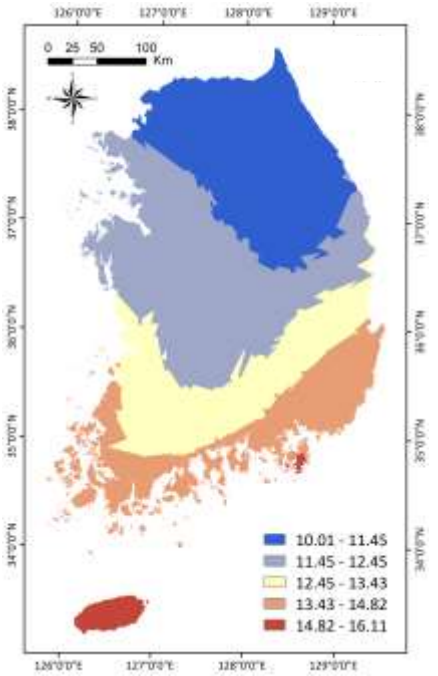
- Elevation and slope produced using digital elevation model with a 30-m resolution
- Temperature and precipitation calculated as the annual mean value between 1981 and 2010 from 73 weather stations



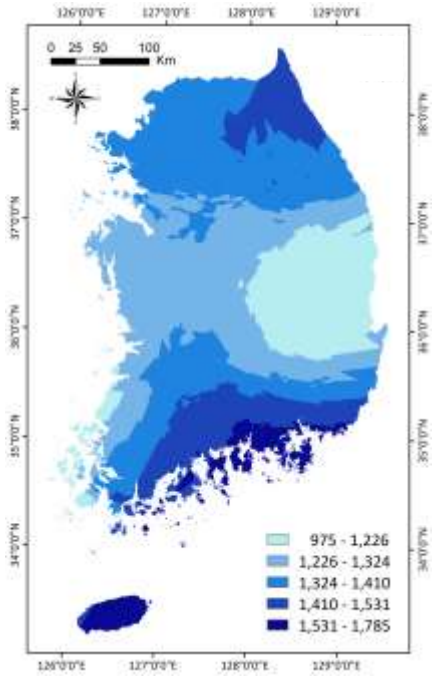
Elevation



Slope



Temperature

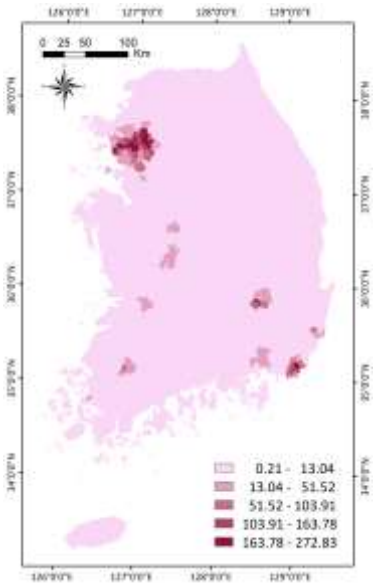


Precipitation

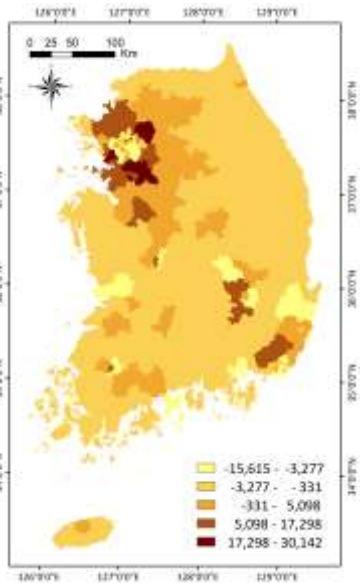
Data and methodology

Socioeconomic factors

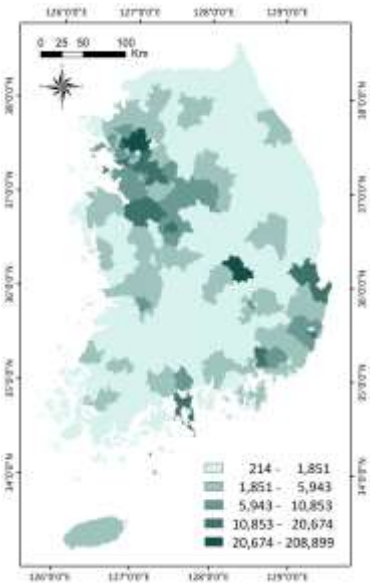
- **Urban areas defined as built-up areas** including residential, industrial, and commercial areas, structures related to transportation, and roads
- **Main roads including highways, national-level roads, and province-level roads** extracted from road maps



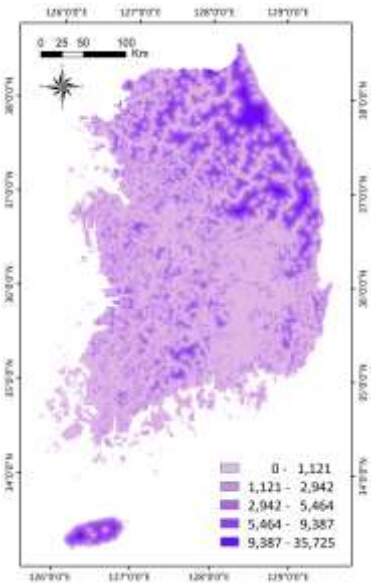
Population density



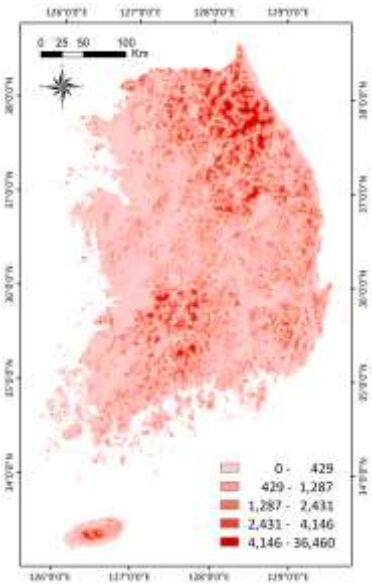
Net migration



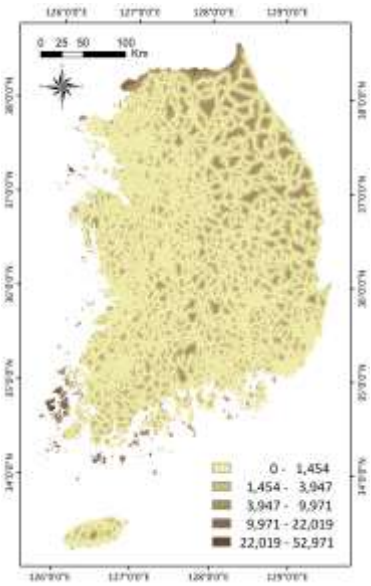
Gross regional domestic product



Distance to built-up area



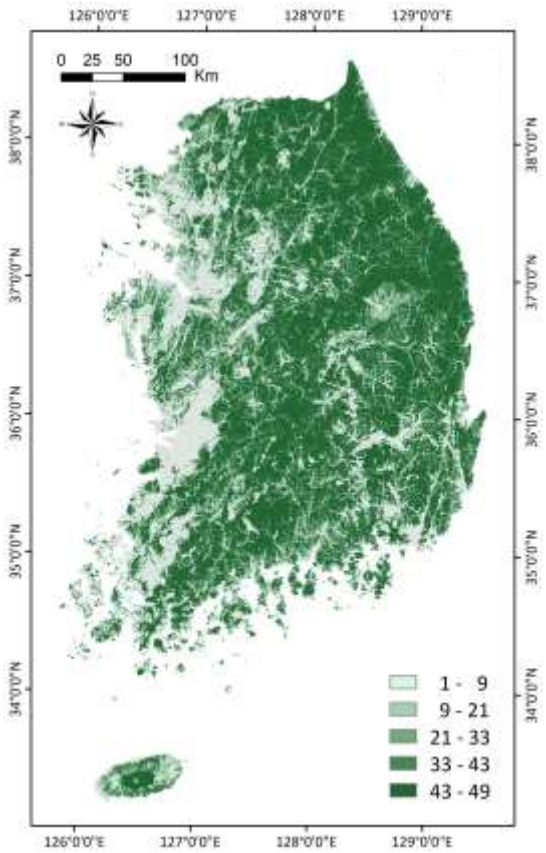
Distance to main road



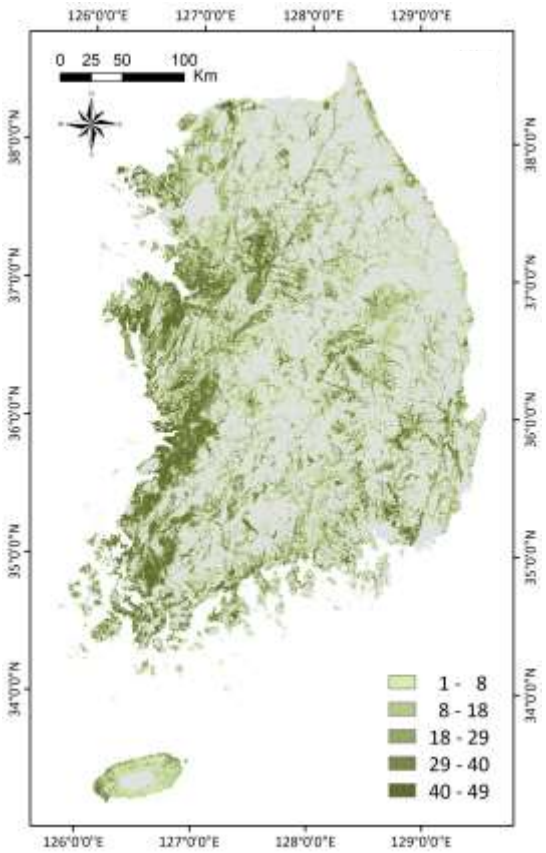
Distance to water body

Data and methodology

Neighborhood factors



Density of agriculture land

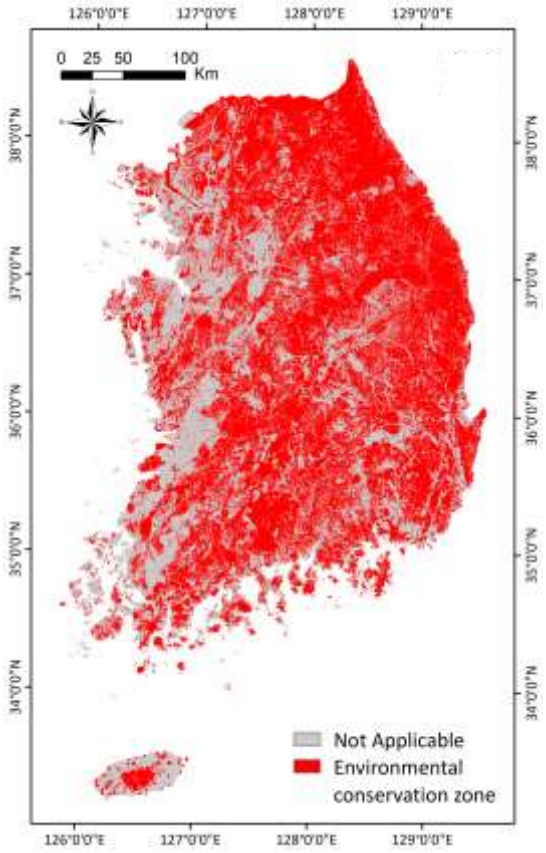


Density of forest land

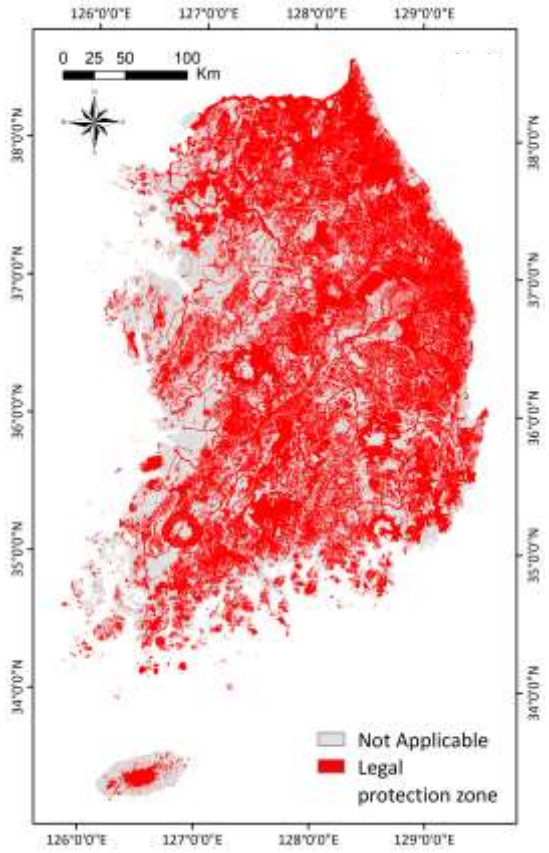
- Application of a **7 X 7 pixel window** with a radius of about a hundred meters to define the neighborhood
- Distance-decay functions and practices that has been used in other studies

Data and methodology

Land use policy factors



Environmental conservation zone



Legal protection zone

- **Extraction from environmental conservation value assessment map** produced by Korea's Ministry of the Environment
- This map for **evaluation about physical and environmental value of land** using eight items related to environmental regulation and 57 items related to legal regulation
- **Usage of the first priority areas assigned to the first and second classes** among total five classes on the map

Data and methodology

Data sampling

- Dependent variable
  - ✓ Production of urban expansion spatial patterns between 2000 and 2010 in a binary map
  - ✓ Assign a value of 1 to cell changed from non-urban to urban
  - ✓ Assign a value of 0 to cell not changed from non-urban to urban and already been open in 2000
- Data sampling
  - ✓ Usage of data sampling for handling and analysis such a large dataset using standard statistical software
  - ✓ Application of a combined systematic and random sampling for minimizing the influence of spatial autocorrelation
  - ✓ The selection of a total of 78,252 about 0 and 1 in the same proportion, with an interval of 10 pixels (300m)

Data and methodology

Logistic regression

- **The logistic regression is used to find empirical relationship between independent continuous and categorical variables and a binary dependent variable.**
- **Logistic regression:**

$$\mathrm{P(Y)} = exp(\boldsymbol{\beta_0} + \sum \boldsymbol{\beta_i x_i})/(1 + exp(\boldsymbol{\beta_0} + \sum \boldsymbol{\beta_i x_i}))$$

↑ Linearization

$$\ln\left(\frac{\mathrm{P}(Y)}{1-P(Y)}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \cdots + \boldsymbol{\beta_k x_k}$$

$\ln\left(\frac{\mathrm{P}(Y)}{1-P(Y)}\right)$ is the log (odds) of the outcome, P is the probability of the dependent variable,

$x_1(i = 1, 2, \cdots, k)$ are the independent variables, $\beta_i(i = 1, 2, \cdots, k)$ is a vector of the coefficient of the estimated parameter

- **The coefficient of the estimated parameters are usually determined using a maximum likelihood (ML) estimation as a convergence criterion.**

Data and methodology

Latent class regression

- The latent class regression tests in a single analysis the differential relationships across a number of latent classes between predictor and output.
- The latent class regression is a random effects model that is not parametrically governed.
- The most popular method to estimate parameters of the latent class model is to maximize likelihood of the log:

$$\ln L = \sum_{i=1}^{n} ln \sum_{w=1}^{W} \pi_w \prod_{i=1}^{I} \pi_{iy^{(i)}|w}$$

- Fitting the function into data is realized by using expectation maximization (EM) algorithm.
- The latent class regression model generalizes the basic latent class model by permitting the inclusion of covariates to predict individuals latent class membership.

$$\ln (\pi_{wi}/\pi_{1i}) = \beta_w X_i + \beta_{0w} \quad (w = 2, \cdots, W)$$

β_w and β_{0w} are the class-specific regression coefficients.

- 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.

04 Results

Results

Logistic regression analysis

- Multicollinearity test
  - ✓ Usage of tolerance (TOL) and a variance inflation factor (VIF) as the standard for multicollonearity diagnosis
  - ✓ Conduction of the LR analysis excepting for the factors having values of TOL < 0.1 and VIF > 10

Variables	TOL [a]	VIF [b]
Ele	0.341	2.934
Slo	0.408	2.451
Temp	0.734	1.362
Rain	0.848	1.179
Pop	0.445	2.246
Mig	0.918	1.089
Grdp	0.486	2.059

Variables	TOL [a]	VIF [b]
DistRoad	0.902	1.108
DistUrban	0.476	2.103
DistWater	0.637	1.570
DenFarm	0.430	2.323
DenFore	0.237	4.221
Environ	0.452	2.212
Law	0.628	1.592

[a] Tolerance

[b] Variance inflation factor

Results

Logistic regression analysis

Statistics	Value
-2 Log (likelihood) of initial	108480.306
-2 Log (likelihood) of final	66721.726 [a]
Cox and Snell *R* Square	0.414
Nagelkerke *R* Square	0.551
Pseudo *R* Square	0.385
Percentage correctly predict (PCP) [b]	80.500

[a] Estimation terminated at iteration number 6 because parameter estimates changed by less than 0.001

[b]The cut off value is 0.5

- **Model summary statistics**
  - ✓ The Psudo R square value indicating logit model/dataset fit with the range from 0 (no relationship) to 1 (perfect fit)
  - ✓ A value greater than 0.2 for the Psedo R square showing a relatively good fit
  - ✓ The value of the Cox and Snell R square indicating that independent variables can explain dependent variables
  - ✓ The predicted accuracy for urban growth and non-urban growth respectively with 87.3% and 73.1%, respectively

Results

Logistic regression analysis

- All independent variables having the significance at the 0.05 level except for the Ele, Grdp, and DistWater
- Ele, Rain, Mig, Grdp, DistRoad, and DistWater having a positive effect on urban growth
- Slo, Temp, Pop, DistUrban, DenFarm, Denfore, Environ, and Law having a negative effects on urban growth

Variables	B [a]	Wald [b]	Exp (B) [c]
Constant	2.891	406.367	18.017
Ele	0.000	1.204	1.000
Slo	-0.002 *	4.099	0.998
Temp	-0.144 *	218.289	0.866
Rain	0.001 *	114.532	1.101
Pop	-0.001 *	4.703	0.999
Mig	0.000 *	439.776	1.000
Grdp	0.000	0.075	1.000

Variables	B [a]	Wald [b]	Exp (B) [c]
DistRoad	0.000 *	241.928	1.000
DistUrban	-0.001 *	1289.414	0.999
DistWater	0.000	0.191	1.000
DenFarm	-0.019 *	455.077	0.981
DenFore	-0.069 *	4497.367	0.934
Environ	-0.645 *	612.137	0.525
Law	-0.493 *	459.633	0.611

* Significant at the 5% level
[a] Logistic coefficient [b] Wald chi-square values [c] Exponential and coefficient

Results

Latent class regression model

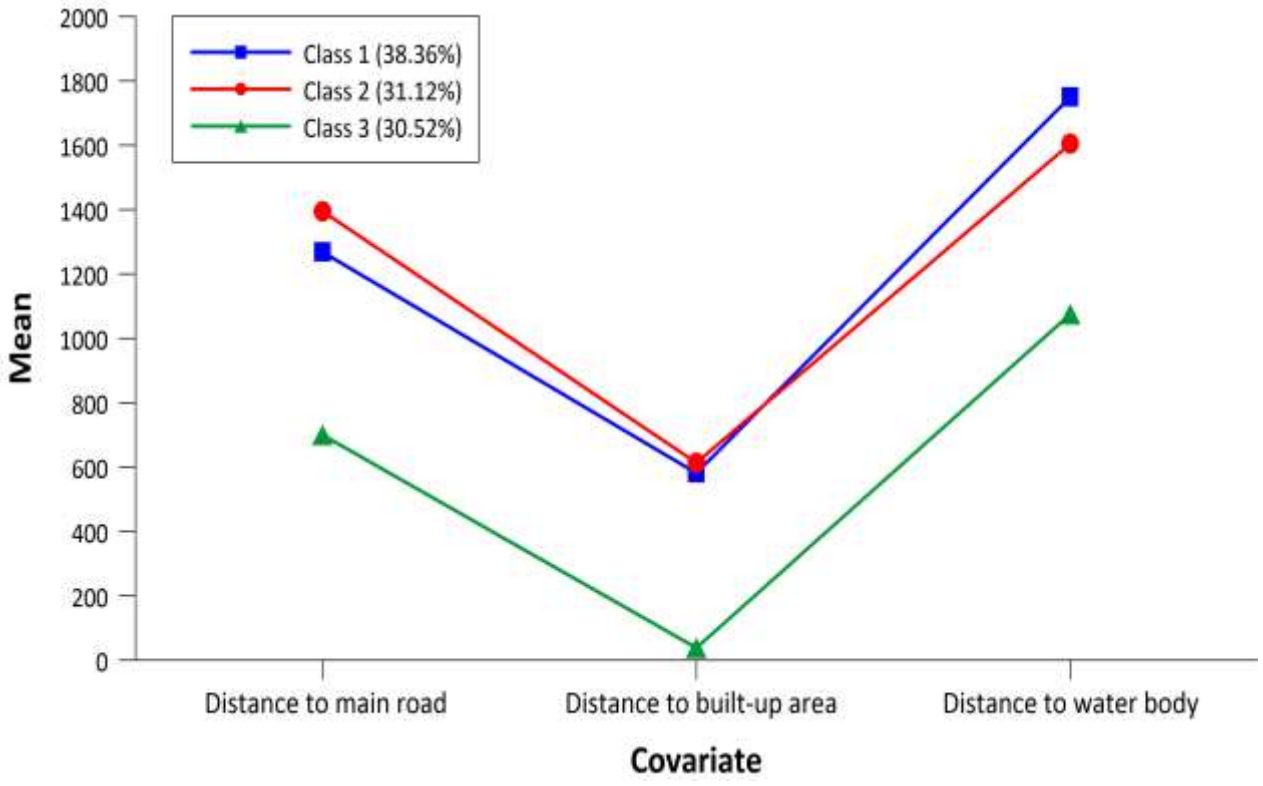
- Starting with one-class and increasing the number of classes until the optimal model is found
- Goodness of fit measurement using the likelihood ratio chi-squared statistic (L^2), Akaike information criterion (AIC), and Bayesian information criterion (BIC)
- The Dramatically decreasing of the magnitudes of improvement after the three-class model, eventually reaching an asymptote

	LL	Difference	BIC (LL)	Difference	AIC (LL)	Difference
1-class model	-33360.863	-	66890.741	-	66751.726	-
2-class model	-30686.287	2674.576	61721.872	5168.869	61434.574	5317.152
3-class model	-29970.315	715.972	60470.211	1251.661	60034.629	1399.945
4-class model	-29771.523	198.792	60252.910	217.301	59669.045	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 proceeding model (eg. 1-class model vs. 2-class model)

Results

Latent class regression model



- Covariates affecting the definition of the latent classes, whereas predictor affecting the dependent variable
- **Usage of the distance variables as the covariates as well as predictors**
  - ✓ The similarity between samples based on DistRoad, DistUrban, and DistWater
  - ✓ The distance between samples
- Classes 1-3 consisting of 38.36% (n=15009), 31.12% (n=12176), and 30.52% (n=11941), respectively

Results

Latent class regression model

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

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

Results

Latent class regression model

- Spatial variables playing important roles as covariates in the latent class regression model
- Based on Wald statistics, DistUrban playing the most important variables for classifying the latent classes

	Class 1			Class 2			Class 3			Wald [d]	p-value
	B [a]	*z*-Stat. [b]	Sig. [c]	B [a]	*z*-Stat. [b]	Sig. [c]	B [a]	*z*-Stat. [b]	Sig. [c]		
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

Results

* Significant at the 5% level

[a] Latent class regression coefficient [b] z-Statistics [c] Significance [d] Wald chi-square values

Comparison of the results between models

① Discrete random effects

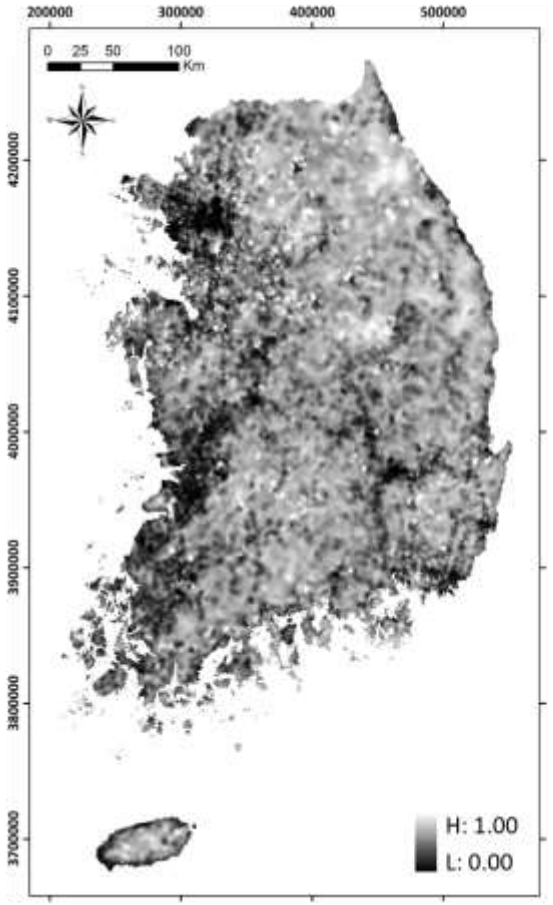
- A discrete distribution is thereby indicated for parameter b, yielding a random-effects, nonparametric modeling approach.
- The LCR model is less intensive, computationally, than parametric models, and that interpreting the results is simplified thanks to the greater consistency of the LCR model with the multiple regression output.

② The flexibility of dependent variables

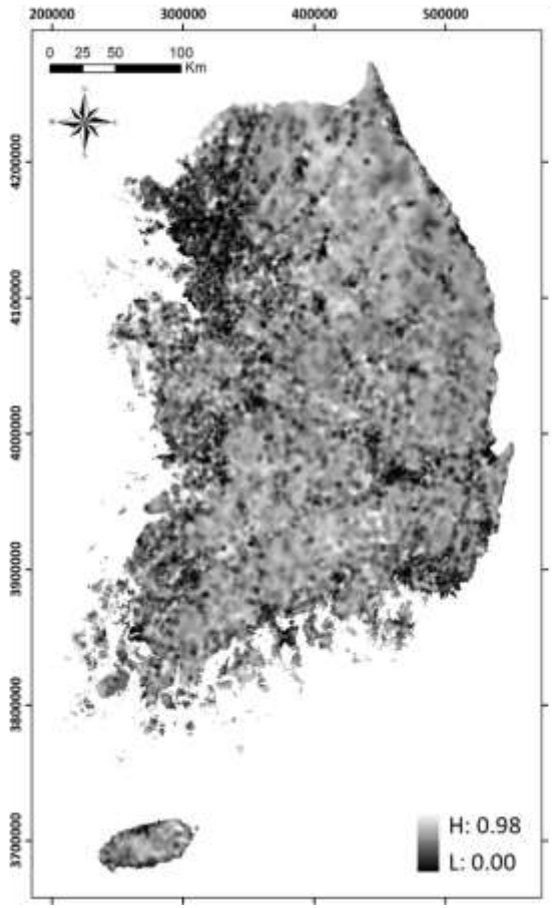
- The dependent variable in the LCR model, though, has greater flexibility since it can be continuous or show Poisson or categorical binomial counts.
- The resulting flexibility being greater than for multiple regression means that models can be tested that are more appropriate for the investigators' data.

Results

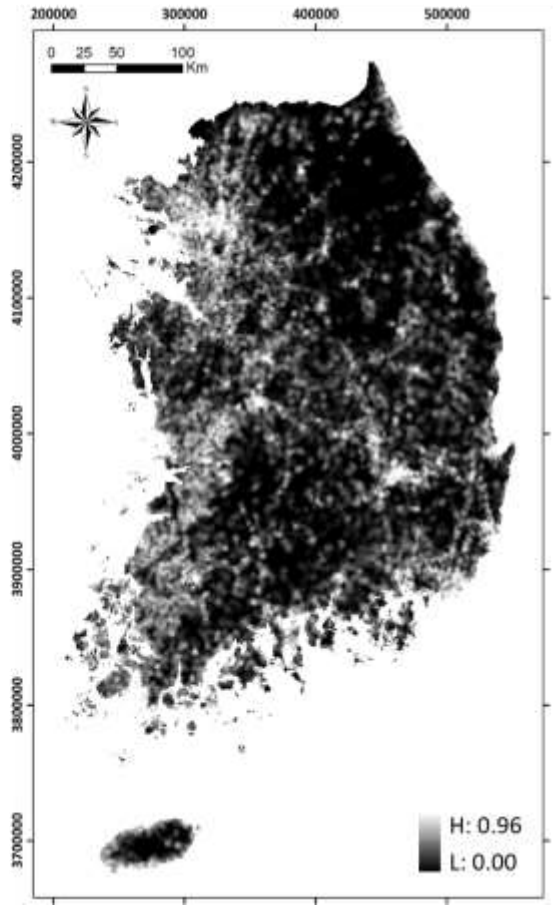
Latent class regression model



Class 1



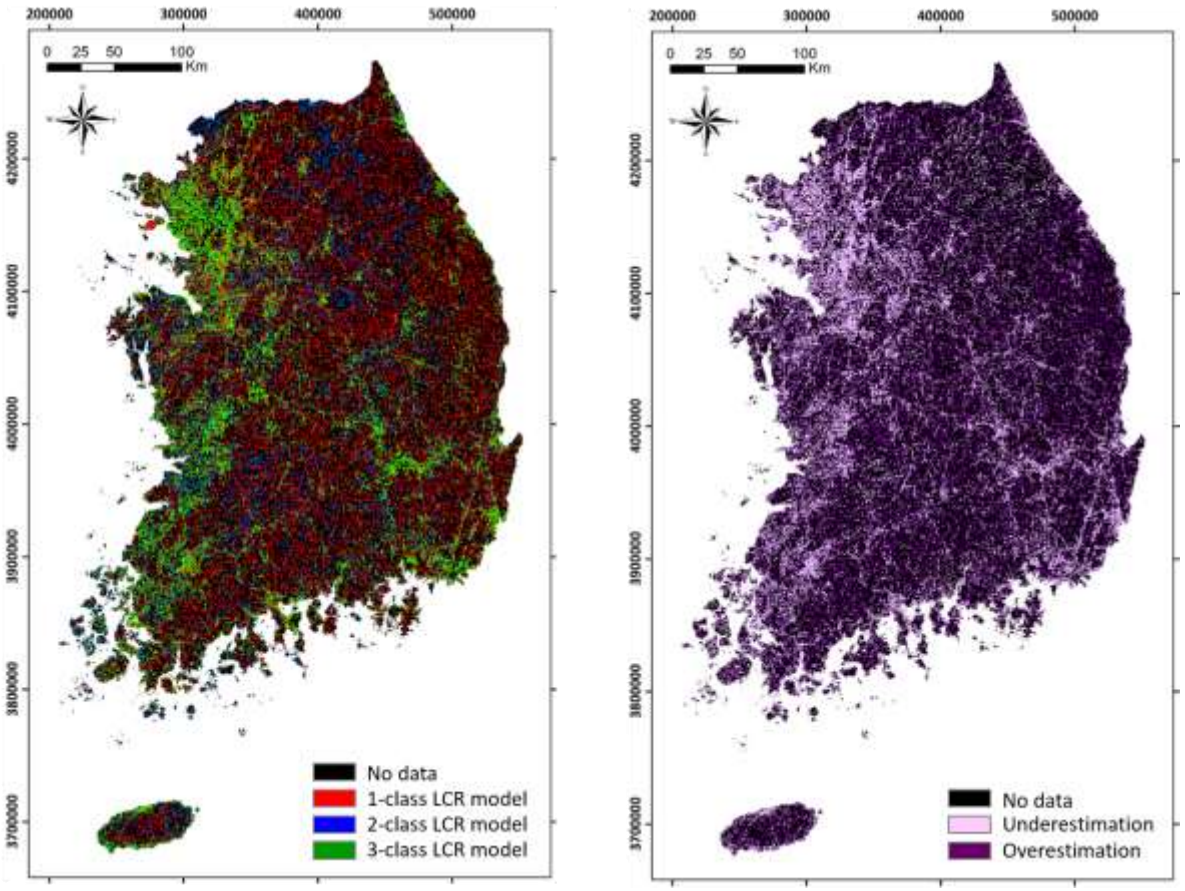
Class 2



Class 3

Results

Comparison of the results between models



③ The derivation of uniqe regression models for each segment

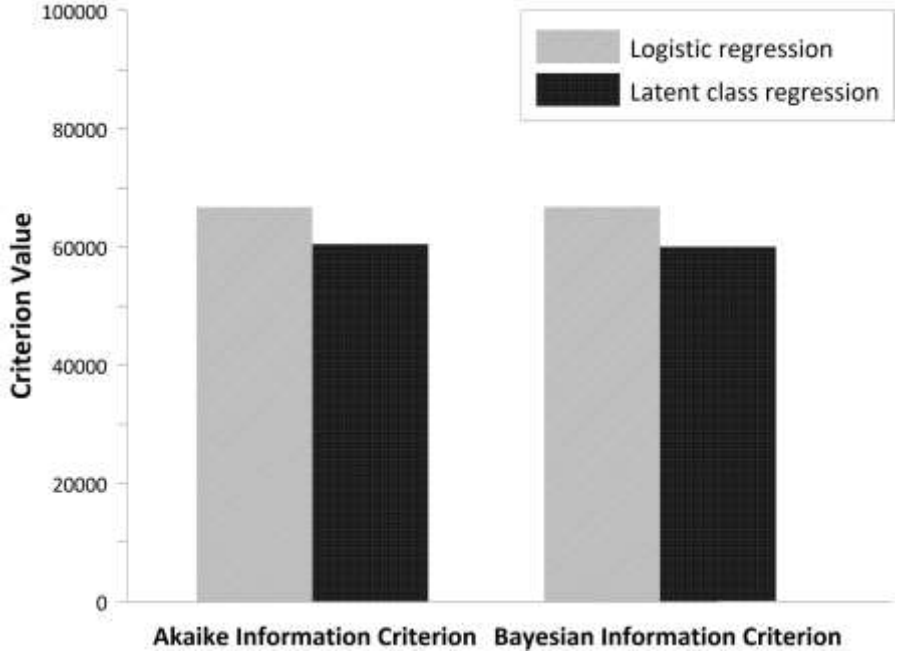
- Homogeneous segments are identified by the LCR model, which derives regression models that are unique for each one.
- For each respondent, probabilities will be provided and a determination made of their likelihood for segment membership.
- The 14 variables affecting urban growth operated in a different way, depending on the underlying spatial structure.

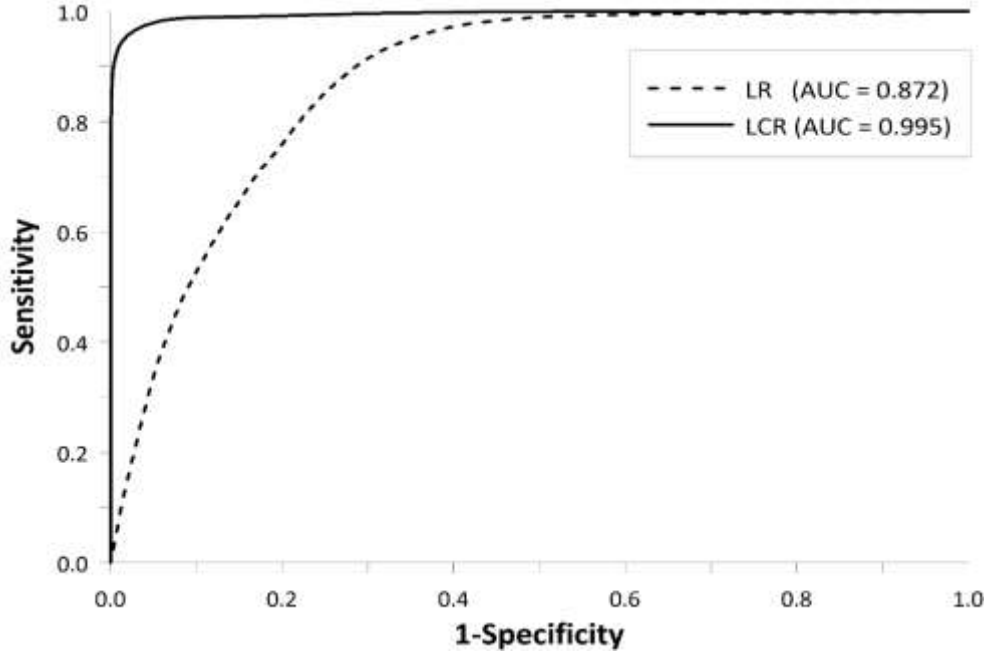
Results

Comparison of the results between models

④ More accurate results

- The values of AIC and BIC for the LCR model were 6279.515, which were 6755.606 lower than for the LR model.
- The interpretability of the model was improved using the LCR model, because toe ROC value was 0.123 higher than that of the LR model.





05 Conclusions

Conclusions

Summary

- The results of the LR and LCR analysis were represented differently in terms of the magnitude and directional effects of coefficients.
  - ✓ 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.
- The LCR analysis could provide insight into the spatial variations of urban growth patterns.
  - ✓ The LCR analysis has a much better goodness of fit with lower values of AIC and BIC.
  - ✓ 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

Conclusions

Limitations and future work

- Since the independent variables used were selected through literature review, they were not be optimized to reflect the exact urban growth patterns in the study area.
  - ✓ Although the three-class LCR model had the highest R-square value, only 6 of the 14 independent variables were significant.
- The spatial factors were used as covariates in the consideration of variation.
- The results may not be representative of the entire study area.
  - ✓ 78,252 data points from a total of 1,116,194 were sampled and used to perform the spatial statistical analysis.
- Future work should determine the definite advantages and disadvantages of LCR analysis by undertaking a comparative study using various statistical methodologies and more data.

Thank you for your attention!!