JOURNAL OF ENVIRONMENTAL DESIGN (JED)

A Journal of the Faculty of Environmental Studies, University of Uyo, Uyo, Nigeria
Vol. 6, No. 1, May, 2011

TABLE OF CONTENT

GIS CELLULAR AUTHOMATA FOR LAND USE CHANGE PREDICTION USING LOGISTIC REGRESSION

Onuwa Okwuashi[1], Etim Eyo[2], & Aniekan Eyoh[3]
(Onuwaokwuashi@yahoo.com, etim.eyo@newcastle.ac.uk, & ani_eyo@yahoo.com)

Department of Geoinformatics & Surveying Faculty of Environmental[1 & 3] Studies, University of Uyo, Uyo, Nigeria.

Department of Geomatics[2] University of Newcastle, Newcastle, United Kingdom.

GIS cellular automata for land use change prediction using logistic regression

ABSTRACT

Since researchers have found the conventional geographic information systems based models not well suited for land use change modelling, in favour of the cellular automata techniques, this research therefore explored the use of the logistic regression for cellular automata calibration. The logistic regression based cellular automata model was loosely coupled with the geographic information systems. Both the non cellular automata and cellular automata techniques were explored. The training of the logistic regression model was based on the k-fold crossvalidation technique. The simulation was validated with the kappa statistic, receiver operating characteristics, and McNemar's test. Results from the cellular automata based technique were better than those from the non cellular automata. The results of the modelling showed substantial agreement between the predicted and the reference data

INTRODUCTION

The GIS-based Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR) are both linear regression models subject to linear regression assumptions of: multicollinearity, linearity, normality, autocorrelation, and homoscedasticity. Unfortunately the nature of land use data for land use change modelling makes it impossible for these linear regression assumptions to be met (Overmars, Koning, & Veldkamp, 2003; Aguiar, 2006; Okwuashi, 2011). Other demerits of conventional GIS models are that: they are not flexible enough (Wagner, 1997), they are not well suited for dynamic modelling (Longley & Batty, 2003), and that they cannot easily be adjusted to perform complex numerical analysis (Wagner, 1997; Couclelis; 2002). Cellular Automata (CA) have therefore been adopted due to their simplicity, dynamic properties, and inventive bottom-up approach (Clarke & Gaydos, 1998). Another advantage of CA models is their compatibility with remote sensing and GIS (Torrens & O'Sullivan, 2001). Coupling GIS and CA models has helped improve dynamic spatial modelling (Park & Wagner, 1997). This research explores the loose coupling of the GIS and Logistic Regression (LR) based CA and non-CA models for modelling land use change in Lagos, Nigeria. The objective of this research is to compare the non-CA and CA methods.

I. LR based CA calibration

This section presents a brief mathematical illustration on how an LR based CA model can be derived. LR is the linear regression model usually used in cases where the dependent variable is dichotomous [0, 1]. Given a linear function,

$$q = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \,, \quad (1)$$

where q is a binary dependent variable, $\beta_0, \ldots \beta_i$ are logistic regression coefficients to be estimated, while x_i are independent variables. An LR model can therefore be expressed as:

$$q = \ln\left(\frac{P(y=1/x)}{1-P(y=1/x)}\right), \quad (2)$$

where P is the probability that $q = 1$, given x_i independent variables; $\frac{P(y=1/x)}{1-P(y=1/x)}$ is called the odds, while $\ln\left(\frac{P(y=1/x)}{1-P(y=1/x)}\right)$ is called the logit. Therefore,

$$P = \frac{e^q}{1+e^q} = \frac{e^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}} = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}$$

(Pohlmann & Dennis, 2003). (3)

Equation 3 is the LR-based non-CA model; and P is the development probability. By introducing the Moore neighbourhood function $\Omega_{3\times3}$ (Wu, 2002), a coefficient Q, constraints contributions $cons_{ij}$, and a stochastic function $1+(-\ln\gamma)^{\alpha}$ (White & Engelen, 1993), equation 3 can be revised to derive the final development probability (Okwuashi, 2011):

$$P_{ij}^{t} = Q * \left(\frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}} \right) * \left(1+(-\ln\gamma)^{\alpha}\right) * \Omega_{3\times3}^{t-1} * \prod_{i=1}^{m} cons_{ij}$$

(4)

where γ is a uniform random variable within the range of [0, 1]; α is a constant that controls the magnitude of the perturbation; $\Omega_{3\times3}^{t-1}$ is an updated function that determines the values of P_{ij}^{t} in each iteration; Q is a coefficient that ensures the values of P_{ij}^{t} are confined to [0, 1]; and $\prod_{i=1}^{m} cons_{ij}$ are the immutable cells that are not affected by the simulation (water and developed cells are considered immutable).

Equation 4 is the LR-based CA model. A threshold probability value (ψ) is set as a benchmark for determining undeveloped cells that are eligible to transit to developed cells:

$$\begin{cases} P_{ij}^{t} \geq \psi & developed \\ Otherwise & undeveloped \end{cases}$$

(5)

Q can also be used to regulate the value of P_{ij}^{t} with respect to ψ, in order to either decrease or increase the number of iterations required for the simulation.

2. Data preparation

The study area for this experiment is Lagos, Nigeria. Lagos is a littoral environment, has a relatively flat terrain, an area of about 2910km^2, and lies between latitudes 6° 26′ and 6° 50′ N, and between longitudes 3° 09′ and 3° 46′ E (Braimoh & Onishi, 2007). Substantial land use change has occurred in Lagos between 1963 and 2000 (Figure 1)

Twelve salient land use drivers were selected for Lagos. The selected land use drivers influencing land use change in Lagos are: water, residential structures, industrial and commercial centres, major roads, railway, Lagos Island, international airport, international seaport, University of Lagos, Lagos State University, income potential, and population potential.

The land use data of Lagos consist of two remotely sensed Landsat Thematic Mapper images, acquired in 1984 and 2000 respectively; and two analogue base maps acquired in 1963 and 1978. The analogue base maps were sourced from the Lagos State Ministry of Lands.

The Landsat images acquired in 1984 and 2000 were classified with the *k*-means algorithm using MATLAB 7.8 software. The analogue base maps acquired in 1963 and 1978 were scanned and digitised using ArcGIS 9.3. The analogue and remote sensing data were geo-referenced to ensure both data were in the same coordinate system. Digitising analogue maps reduces their accuracy. It is therefore important to enhance the accuracy of the analogue data to ensure they approximate that of the satellite maps. The enhancement of the analogue base maps was done in MATLAB, by first obtaining *n*-classifications of the satellite data and overlaying the resulting classified satellite data with the digitised base maps. A digital editing procedure was then used to remove errors from the digitised maps.

The land use independent variables (Table 1) are grouped into two categories: (i) proximity variables, and (ii) weighted variables. The proximity variables were extracted with the GIS while the weighted variables were extracted in MATLAB.

The proximity variables were extracted by calculating the Euclidean distances from the proximity variables to all cells. For the weighted variables, income potential was estimated by ranking major towns in Lagos, using a ranking r_i of a production/service centre c_i. The ranking was normalized to weight w_i using the formula (Braimoh & Onishi, 2007),

$$w_i = \frac{r_i}{\sum_{c_i=1}^{n} r_i}.$$

(6)

Therefore the weighted inverse distance formula for calculating the income potential i_S of a location S is given as (Braimoh & Onishi, 2007),

$$i_s = \frac{\sum_{x=1}^{n} \frac{1}{|S-S_x|^2} w_i}{\sum_{x=1}^{n} \frac{1}{|S-S_x|^2}}.$$

(7)

where $x = 1,2,...,n$ are settlements in the study area.

The final ranking was based on government documents, interviews, and personal knowledge of the author. The population estimate was obtained using the formula (Braimoh & Onishi, 2007),

$$P_p = \frac{P_r}{e^{r(Y_r - Y_p)}} \quad (8)$$

where, r = intercensal population growth rate, P_p = previous population, P_r = present population, Y_r = recent year, and Y_p = previous year.

An inverse distance weighting formula for calculating an unknown population potential p_S, at a given location S, is given as (Braimoh & Onishi, 2007),

$$p_s = \frac{\sum_{x=1}^{n} \frac{1}{|S - S_x|^2} P(S_x)}{\sum_{x=1}^{n} \frac{1}{|S - S_x|^2}} \quad (9)$$

where $P(S_x)$ is the population of a settlement located at S ; $x = 1,2,...,n$ are settlements in the study area.

The original land use maps (in 1963, 1978, 1984, and 2000) had a cell size of 100m x 100m before they were gridded to increase their cell size to 500m x 500m (see the overlaid gridded maps in Figure 1). The essence of increasing the cell size of the maps is to reduce the matrix size of the maps for modelling convenience. The original values of the independent variables were scaled to [0, 1] using the transformation formula (Gong, 1996; Li & Yeh, 2002),

$$x_i' = (x_i - \min)/(\max - \min) \quad (10)$$

where x_i' is the scaled land use variable; min is the lowest value in the land use vector; max is the highest value in the land use vector; and x_i represents the land use variables.

This scaling technique is effective in ensuring that all the independent variables are equally weighted (Gong, 1996; Li & Yeh, 2002).

3. Modelling

All the land use independent variables were prepared in the GIS using the ArcGIS before being imported into MATLAB. Two stages of land use development were used to assess the evolution of land use: (i) non-CA land use change modelling, and (ii) land use change modelling using CA. These two techniques were run through a land use transition module that finally determined the future state of the undeveloped cells. Only undeveloped cells can change their state to developed. The developed cells are immutable in the simulation. The non-CA land use change module is non-iterative while the CA land use change model is iterative. For the non-CA modelling, the transition from undeveloped to developed is when undeveloped cells have development probabilities greater than 0.5 (P>0.5). For CA modelling, cells with probabilities greater or equal to the threshold probability became developed cells while cells with probabilities below the threshold probability were re-admitted into the land use processing module. This section describes the modelling of land use change between 1963-1978, 1978-1984, and 1984-2000. Dummy variables were used to represent the land use dependent variables. The dependent variable is dichotomous (0=undeveloped and 1=developed). The independent variables consist of the explanatory variables for periods 1963-1978, 1978-1984, and 1984-2000 given in Table 1. As mentioned in the previous section, all the data for the three periods 1963-1978, 1978-1984, and 1984-2000 were scaled to [0, 1]. Each pair of land use map was overlaid to determine the change regions, to ensure that the data from the change region were excluded from the training data, thereby extracting only points common to both developed and undeveloped cells.

The non-CA modelling experiment is a classification experiment. The outcome of the classification is either developed or undeveloped. Cells with probabilities >0.5 are classified as developed while cells with probabilities <0.5 are classified as undeveloped. To predict the non-CA maps for periods 1963-1978, 1978-1984, and 1984-2000, the stratified random sampling approach was used to extract 1000 training points from each combination of periods 1963-1978, 1978-1984, and 1984-2000. The two objectives of this modelling were to: (i) predict land use change between the periods, and (ii) evaluate the accuracy of the LR model in predicting land use change for the three periods. The selected 1000 points were split into 10 equal datasets. Each dataset had a sample size of 100. Each dataset consists of 50 developed cells/points and 50 undeveloped cells/points. As usual, the developed cells were labelled +1 while the undeveloped cells were labelled 0. The *k*-fold cross-validation technique (where *k*=10) was used to evaluate the accuracy of the LR model in the prediction. The model was trained by putting together 9 subsets (*k*-1) out of the 10 datasets while 1 subset was used to test the accuracy of the prediction. The experiment was repeated in 10 folds by eventually using all the 10 datasets for both training and testing. The LR model invokes the land use change between periods 1963 and 1978, 1978 and 1984, and 1984 and 2000, based on training samples only selected from the regions/points common to 1963 and 1978, 1978 and 1984, and 1984 and 2000. Using equation 3, development probabilities >0.5 were classified as developed cells, while probabilities <0.5 were classified as undeveloped cells. The best maps yielded by the prediction for periods 1963-1978, 1978-1984, and 1984-2000 are given in Figure 2; and their computed confusion matrices are presented in Tables 2-4. The calculated kappa coefficients for the predicted maps given in Figure 3 were determined by comparing/validating the predicted maps with the actual maps in 1978, 1984, and 2000. The calculated kappa coefficients were 0.5057, 0.5525, and 0.5080 for periods 1963-1978, 1978-1984, and 1984-2000 respectively.

Now the next procedure presents the bottom-up simulation results from the LR-based CA simulation. The basic difference between the CA model and the non-CA model is simply the incorporation of the Moore neighbourhood function $\Omega_{3\times3}$ into the LR model (see equation 4). CA models simply use the neighbourhood influence of the initial state of the observed object and the independent variables to predict the future state of the target object. The bottom up approach of CA models makes their predictions highly accurate (Torrens & O'Sullivan, 2001). There is no training of data in the CA modelling; since the LR model has already been trained. Only the neighbourhood function is updated to determine the conversion of undeveloped cells to develop. CA are iterative systems; for example, to predict from 1963-1978, the starting point of the simulation will be the 1963 data/map (that is iteration=0 is 1963), while the target map 1978 is used to validate the prediction. At iteration>0, the neighbourhood function $\Omega_{3\times3}^{t-1}$ calculates the number of developed cells surrounding each undeveloped cell in the 1963 map. This process is repeated in each iteration.

Two hundred iterations were run to simulate the maps from periods 1963-1978, 1978-1984, and 1984-2000 (see Figure 4). There are no rules that guide the number of iterations needed for a simulation, but 100 to 200 iterations are common in most applications (Wu, 2002; Li & Yeh, 2004). Figure 3 depicts the mean kappa coefficients and standard deviations for 10,20,30,...,200 designated iteration thresholds, calculated by running each threshold ten times and comparing the simulated maps with the actual maps for periods 1963-1978, 1978-1984, and 1984-2000.

From Figure 3, at iteration=10 the mean kappa coefficients for 1963-1978, 1978-1984, and 1984-2000 were low, but gradually increased as the number of iterations increased. The highest mean kappa coefficients were found from the 90th to 140th iterations. The mean kappa coefficients decreased beyond the 140th iterations until the 200th iteration. The final maps for periods 1963-1978, 1978-1984, and 1984-2000 shown in Figure 4 were obtained from 90th to 170th iterations. The calculated kappa coefficients from the confusion matrices given in Tables 5-7 for the predicted maps for periods 1963-1978, 1978-1984, and 1984-2000 were 0.5847, 0.7543, and 0.7101. The highest kappa statistic estimate was obtained with 1978-1984 while the lowest was obtained with 1963-1978. Evaluating the performance of the CA model can be intricate. The Receiver Operating Characteristics (ROC) was used to assess the performance of the LR-based CA model. The ROC is the plot of sensitivity against 1-specificity. Sensitivity is calculated by dividing the number of the true positive matches by the sum of the true positive and false negative matches; while specificity is calculated by dividing the number of the true negative matches by the sum of the true negative and false positive matches. The Area Under Curve (AUC) determines the result of the plot. Experiments that yield AUC indices <0.5 are usually regarded as worthless. Figure 5 depicts the plot of mean sensitivity against mean 1-specificity, and their respective standard deviations calculated from 10 ROC curves sampled at fixed 1-specificity points: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 (see Fawcett, 2004). The mean sensitivity and mean 1-specificity was calculated by comparing the simulated maps with the actual maps. The calculated AUC for 1963-1978, 1978-1984, and 1984-2000 were 0.7354 ± 0.0295, 0.7549 ± 0.0267, 0.7451 ± 0.0298. The calculated ROC results corroborated the results from the CA predicted maps because the order of best fit of the target maps was still: 1978-1984, 1984-2000, and 1963-1978 respectively.

Using test of hypothesis, we can statistically determine whether the results from the non-CA and CA models are significantly different or not. The McNemar's test is used to assess the statistical significance of two related samples (Bradley, 1968; Agesti, 1996; Foody, 2004; Huang, Xie, & Tay, 2010). The McNemar's test is based on the elements in a confusion matrix. Therefore, McNemar's test in this test is based on the confusion matrices that yielded the kappa coefficients given in Table 8.

According to Foody (2004), the McNemar's test evaluates the z-score from a standardised normal test statistic,

$$z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}}$$

(11) where f_{12} and f_{21} are the sum of incorrectly classified pixels resulting from a 2-class problem (see Table 9).

In this test, let the non-CA predicted maps be map 1 while CA predicted maps be map 2. The analysis was done by comparing the non-CA and CA predicted maps. The respective values of f_{12} and f_{21} were used to compute the respective z-scores using equation 11. The statement of hypothesis can be written as,

H_0: there is no significant difference between the two predicted maps at 95% CL

H_1: there is a significant difference between the two predicted maps (reject H_0 if p-value<0.05).

Using a two-tailed test, the null hypothesis is rejected at 5% significant level that is when $|z| > 1.96$. The computed confusion matrices, calculated *z*-scores, and p-values for the non-CA and CA models are presented in Table 10.

From Table 10, the computed *p*-values for the three periods were < 0.05. We therefore reject the null hypothesis, and accept the alternative hypothesis that the non-CA and CA predictions were significantly different. This implies that based on the McNemar's test, the CA predictions were better than those of the non-CA for three periods. The McNemar's test corroborated the kappa estimates for non-CA and CA models that showed that the CA kappa statistic estimates were higher than those of the non-CA.

4. Conclusion

The CA based technique performed better than the non-CA based technique. The significant agreement between the actual and the predicted maps showed that the logistic regression model is a promising tool for modelling land use change; despite the preference for non-parametric methods of CA calibration, such as the artificial neural networks.

REFERENCES

Agresti, A. (1996). *An introduction to categorical data analysis. New York, NY: Wiley.*

Aguiar, A. P. D. (2006). *Modelling land use change in the Brazilian Amazon: Exploring the intra-regional heterogeneity* (Unpublished doctoral dissertation). INPE, São José dos Campos, SP, Brazil.

Bradley, J. V. (1968). *Distribution-free statistical tests.* Upper Saddle River, NJ: Prentice-Hall.

Braimoh, A. K. & Onishi, T. (2007). Spatial determinants of urban land use change in Lagos, Nigeria. *Land Use Policy, 24*(2), 502-515.

Clarke, K. C. & Gaydos, L. (1998). Long term urban growth prediction using a cellular automaton model and GIS. *International Journal of Geographical Information Science, 12*(7), 699-714.

Couclelis, H. (2002). Modelling frameworks, paradigms and approaches. in K. C. Clarke, B. E. Parks, & M. P. Crane(Eds). *Geographic information systems and environmental modelling* (pp. 36-50). Upper saddle River, NJ: Prentice Hall.

Fawcett, T. (2004). *ROC graphs: Notes and practical considerations for data mining researchers* (Technical report HPL-2003-4). Palo Alto, CA: HP Laboratories.

Foody, G. M. (2004). Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogramm. Eng. Remote Sensing, 70*, 627–633.

Gong, P. (1996). Integrated analysis of spatial data from multiple sources: Using evidential reasoning and artificial neural network techniques for geological mapping. *Photogrammetric Engineering & Remote Sensing, 62*(5), 513–523.

Huang, B., Xie, C., & Tay, R. (2010). Support vector machines for urban growth modelling. *Geoinformatica, 14*, 83–99.

Li, X. & Yeh, A. G. (2002). Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science, 16*(4), 323-343.

Li, X. & Yeh, A. G. (2004). Data mining of cellular automata transition rules. *International Journal of Geographical Information Science, 18*(8), 723-744.

Longley, P. & Batty, M. (Eds.). (2003). *Advanced spatial analysis: The CASA book of GIS*. Redlands, CA: ESRI Press.

Okwuashi, O. (2011). *Application of geographic information systems cellular automata based models to land use change modelling of Lagos, Nigeria* (Unpublished doctoral dissertation). Victoria University of Wellington, Wellington, New Zealand.

Overmars, K. P., Koning, G. H., & Veldkamp, A. (2003). Spatial autocorrelation in multiscale land use models. *Ecological Modelling, 164*, 257-270.

Park, S. & Wagner, D. F. (1997). Incorporating cellular automata simulators as analytical engines in GIS. *Transactions in GIS, 2*(3), 213-231.

Pohlmann, J. T. & Dennis, W. L. (2003). A comparison of ordinary least squares and logistic regression. *Ohio Journal of Science, 103* (5), 118-125.

Torrens, P. M. & O'Sullivan, D. (2001). Cellular automata and urban simulation: Where do we go from here? *Environment and Planning B., 28*, 163-168.

Wagner, D. F. (1997). Cellular automata and geographic information systems. *Environment and Planning B*, *24*, 219-234.

White, R. & Englelen, G. (1993). Fractal urban land use patterns: A cellular automata approach. *Environment and Planning A*. *25*, 1175-1199.

Wu, F. (2002). Calibration of stochastic cellular automata: The application to rural urban land conversions. *International Journal of Geographical Information Science*, *16*(8), 795-818.

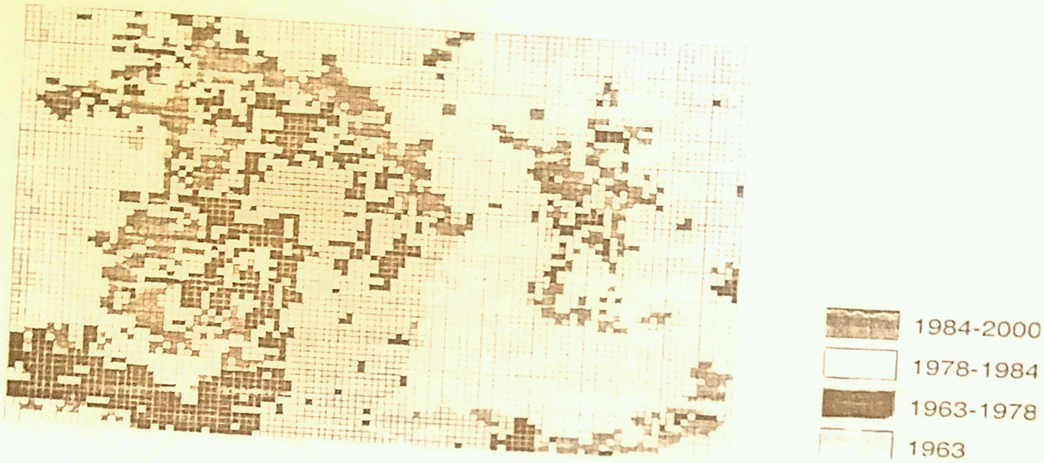

Figure 1 *Land use of Lagos between 1963 and 2000*

Table 1 *Extracted land use variables*

Land use variables	
Proximity variables	x_1 : distance to water
,,	x_2 : distance to residential structures
,,	x_3 : distance to industrial and commercial centres
,,	x_4 : distance to major roads
,,	x_5 : distance to railway
,,	x_6 : distance to Lagos Island
,,	x_7 : distance to international airport (1984-2000 only)
,,	x_8 : distance to international seaport
,,	x_9 : distance to University of Lagos
,,	x_{10} : distance to Lagos State University (1984-2000 only)
Weighted variables	x_{11} : income potential
,,	x_{12} : population potential

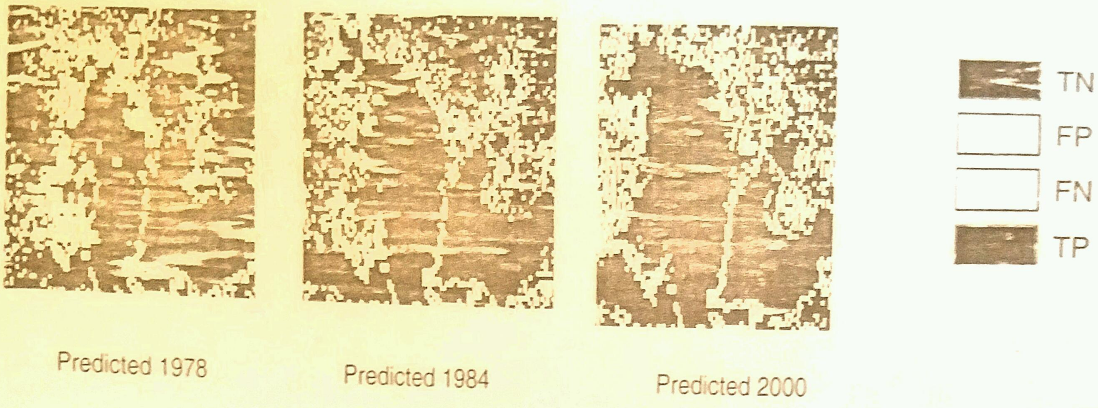

Figure 2 *LR non-CA predicted maps for periods 1963-1978, 1978-1984, and 1984-2000 (TN=True negative; FP=False Positive; FN=False Negative; TP=True Positive)*

Table 2 *LR-based non-CA: Confusion matrix for period 1963-1978*

	Reference data 1978	
	Developed	Undeveloped
Predicted data 1978		
Developed	1578	1387
Undeveloped	214	3821

Table 3 *LR-based non-CA: Confusion matrix for period 1978-1984*

	Reference data 1984	
	Developed	Undeveloped
Predicted data 1984		
Developed	2224	1283
Undeveloped	284	3209

Table 4 *LR-based non-CA: Confusion matrix for period 1984-2000*

	Reference data 2000	
	Developed	Undeveloped
Predicted data 2000		
Developed	3019	1317
Undeveloped	413	2251

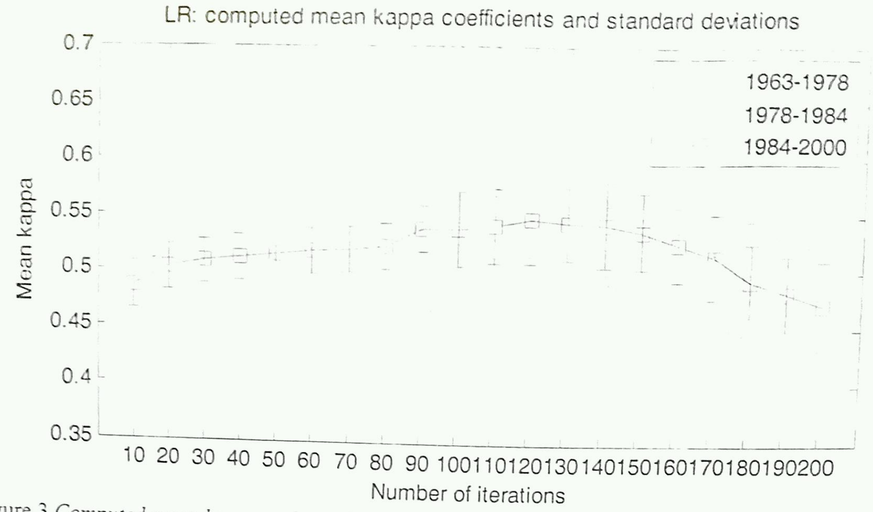

Figure 3 *Computed mean kappa and standard deviations for 200 designated iteration thresholds*

Figure 4 *LR-based CA predicted maps for 1963-1978, 1978-1984, and 1984-2000*

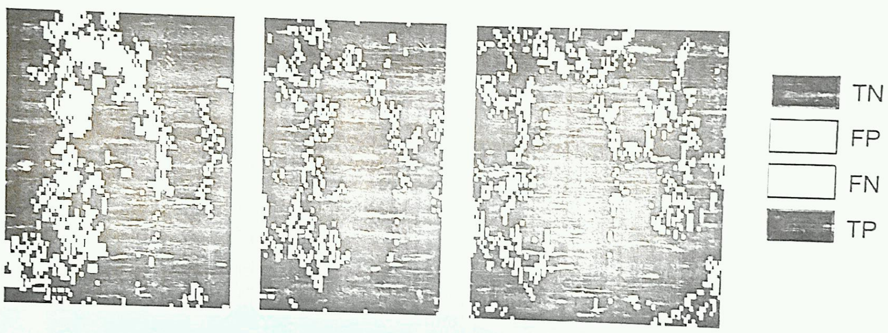

Table 5 *LR-based CA: confusion matrix for period 1963-1978*

	Reference data 1978	
	Developed	Undeveloped
Predicted data 1978		
Developed	1566	1048
Undeveloped	226	4160

Table 6 *LR-based CA: confusion matrix for period 1978-1984*

	Reference data 1984	
	Developed	Undeveloped
Predicted data 1984		
Developed	2280	588
Undeveloped	228	3904

Table 7 *LR-based CA: confusion matrix for period 1984-2000*

	Reference data 2000	
	Developed	Undeveloped
Predicted data 2000		
Developed	2889	471
Undeveloped	543	3091

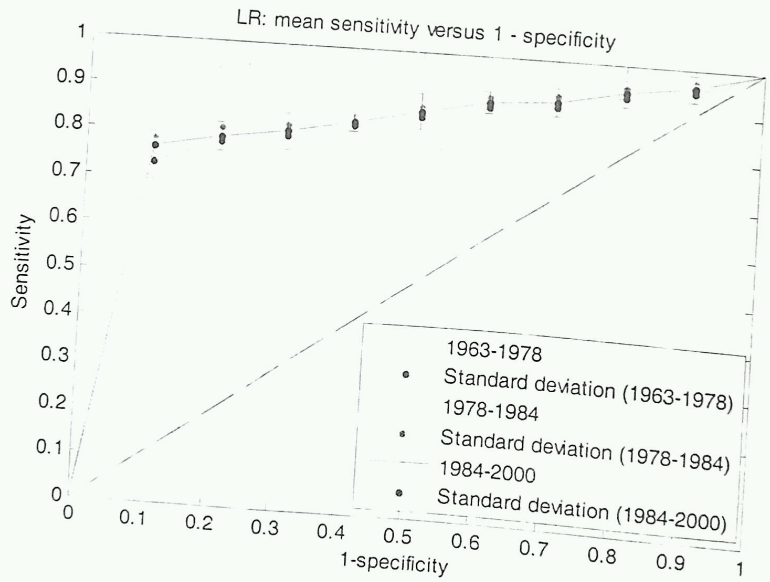

Figure 5 *Plotted mean sensitivity versus 1-specificity and standard deviations for periods 1963-1978, 1978-1984, and 1984-2000*