Predicting Future Land Use Change Using Support Vector Machine Based GIS Cellular Automata: A Case of Lagos, Nigeria

Onuwa Okwuashi[1], Jack McConchie[2], Peter Nwilo[3], Mfon Isong[1], Aniekan Eyoh[1], Okey Nwanekezie[4], Etim Eyo[5] & Aniekan Danny Ekpo[6]

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

[2] School of Geography, Environment, and Earth Sciences, Victoria University of Wellington, Wellington, New Zealand

[3] Department of Surveying & Geoinformatics, Faculty of Engineering, University of Lagos, Lagos, Nigeria

[4] Department of Estate Management, Faculty of Environmental Studies, University of Uyo, Uyo, Nigeria

[5] School of Civil Engineering & Geosciences, Newcastle University, Newcastle upon Tyne, UK

[6] Amana Consortium Engineers Ltd, Uyo, Akwa-Ibom State, Nigeria

Correspondence: Onuwa Okwuashi, Department of Geoinformatics & Surveying, Faculty of Environmental Studies, University of Uyo, Uyo, Nigeria. Tel: 234-3044-4355. E-mail: onuwaokwuashi@yahoo.com

Received: February 1, 2012 Accepted: March 23, 2012 Online Published: May 1, 2012

doi:10.5539/jsd.v5n5p132 URL: http://dx.doi.org/10.5539/jsd.v5n5p132

Abstract

Lagos has undergone an unprecedented urban expansion. Contemporary findings favour the integration of cellular automata and geographic information systems for modelling land use change. This research introduces the support vector machine based GIS cellular automata calibration for land use change prediction of Lagos. The support vector machine based cellular automata model is loosely coupled with the geographic information systems. Support vector machine parameters are optimised with the *k*-fold cross-validation technique, using the linear, polynomial, and RBF kernels functions. The land use change prediction is based on three land use epochs: 1963-1978, 1978-1984, and 1984-2000. The performance of the model was evaluated using the Kappa statistic and receiver operating characteristic. The order of performance of the three kernels is: RBF, polynomial, and linear. The results indicate substantial agreement between the actual and predicted maps. The urban forms in 2015 and 2030 are predicted based on the three land use epochs.

Keywords: GIS, cellular automata, support vector machine, land use change

1. Introduction

Urban sprawl in Lagos has put profound pressure on housing, infrastructure, and the environment (Braimoh & Onishi, 2007). Technological methods, such as Geographic Information Systems (GIS) and other predictive models necessary for sustainable physical planning are rarely utilised by urban planners in Lagos (Oduwaye, 2009). Modelling an unregulated complex urban environment like Lagos may be unyielding without employing robust predictive tools that can realistically model their complexity, dynamism, and growth (Barredo et al., 2004). In this research, a loose coupling of the GIS and the Cellular Automata (CA) model has been adopted as the most appropriate tool for modelling land use change in Lagos. This is because CA models present the necessary structure for modelling complex adaptive systems like land use change; another merit of CA models is their compatibility with the GIS (Torrens & O'Sullivan, 2001).

Common parametric and non-parametric CA applications are based on logistic regression and artificial neural network respectively (Okwuashi, 2011). The objective of this research is to present a loosely coupled GIS-CA model based on the novel non-parametric Support Vector Machine (SVM) (Cortes & Vapnik, 1995) model for predicting future land use change of Lagos, Nigeria in 2015 and 2030.

2. Support Vector Machine

SVM is intrinsically a binary classifier. For the linear case, let us classify a binary problem that belongs to classes -1 and +1 respectively using a linear hyperplane. To separate these two sets of objects, we need to choose a few training samples. Let us assume that our training set has *n*-training samples, that is,

$(x_1, y_1),(x_2, y_2),...,(x_n, y_n)$, where $x_i \in \Re^N$ is an N dimensional vector that belongs to one of classes $y_i \in \{-1,+1\}$. The stated binary classification problem can be separated using a linear decision function (Vapnik, 2000),

$$f(x) = w \cdot x + b \tag{1}$$

where $w \in \Re^N$ is a vector that determines the orientation of the desired hyperplane required for the separation, and $b \in \Re$ is called the "bias."

The optimal hyperplane needed to separate the two objects is,

$$y_i(w \cdot x + b) \geq 1 \tag{2}$$

The solution to this problem can be found by solving the following constrained optimization problem (or primal problem) (Vapnik, 2000), Minimise:

$$\frac{1}{2} w \cdot w + C \sum_{i=1}^{n} \xi_i \tag{3}$$

subject to: $y_i(w \cdot x + b) \geq 1 - \xi_i$, $\xi_i > 0$, and for $\forall i = 1,...,n$; where C, $0 < C < \infty$, is called the penalty value or regularization parameter; while ξ_i are the slack variables.

The optimisation problem or dual form derived by solving equation 3 can be expressed as, maximise:

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{4}$$

subject to: $\sum_{i=1}^{n} \alpha_i y_i = 0$, and, $0 \leq \alpha_i \leq C$, for $i = 1,...,n$.

The resulting decision function for the linear case can be given as,

$$f(x) = sign\left[\sum_{i=1}^{n} y_i \alpha_i^0 (x_i \cdot x) + b^0 \right] \tag{5}$$

where x_i are the training samples; y_i are the target labels of the training samples (such that, $y_i \in \{-1,+1\}$); α_i^0 are the Lagrangian multipliers; b^0 is known as the "bias;" while x denotes the test set.

For the nonlinear case the optimisation problem can be written as, maximise:

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{6}$$

subject to: $\sum_{i=1}^{n} \alpha_i y_i = 0$, and, $0 \leq \alpha_i \leq C$, for $i = 1,...,n$.

While the resulting decision function can be given as,

$$f(x) = sign\left[\sum_{i=1}^{n} y_i \alpha_i^0 K(x_i, x) + b^0 \right] \tag{7}$$

Given two arbitrary support vectors $x_A \in$ class *A* and $x_B \in$ class *B*, the bias can be evaluated as,

$$b^0 = -\frac{1}{2} \sum_{i=1}^{n} y_i \alpha_i^0 \left[K(x_A, x_i) + K(x_B, x_i) \right] \tag{8}$$

Equation 8 can be used to evaluate the value of b^o in equations 6 and 8. The kernel $K(x_i, x_j)$ can be any of the following common kernel functions: the linear kernel $x \cdot x_i$, polynomial kernel $(x \cdot x_i + 1)^d$, and Radial Basis Function (RBF) kernel $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\gamma^2} \right)$ (Vapnik, 2000).

3. Support Vector Machine Based Cellular Automata Calibration

SVM output $f(x) = sign\left[\sum_{i=1}^{n} y_i \alpha_i^0 K(x_i, x) + b^0\right]$ given in equation 7 can be mapped into probabilities using a sigmoid function (Platt, 1999). Therefore, SVM-based land use development probability can be expressed as (Okwuashi, 2011),

$$P = \frac{1}{1+e^{-(sign[\sum_{i=1}^{n} y_i \alpha_i^0 K(x_i, x) + b^0])}} \tag{9}$$

By introducing the Moore neighbourhood function $\Omega_{3\times3}$ (Wu, 2002); a coefficient Q; constraints contributions $cons_{ij}$; and a stochastic function $1+(-\ln\gamma)^{\alpha}$ (γ is a uniform random variable; while α controls the magnitude of the perturbation) (White & Engelen, 1993); equation 9 can be revised as the final development probability (Okwuashi, 2011),

$$P_{ij}^{t} = Q * \left(\frac{1}{1+e^{-(sign[\sum_{i=1}^{n} y_i \alpha_i^0 K(x_i, x) + b]}}\right) * \left(1+(-\ln\gamma)^{\alpha}\right) * \Omega_{3\times3}^{t-1} * \prod_{i=1}^{m} cons_{ij} \tag{10}$$

Equation 10 is the SVM-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,

$$\begin{cases} P_{ij}^{t} \geq \psi & developed \\ Otherwise & undeveloped \end{cases} \tag{11}$$

Q is introduced 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.

4. Application

4.1 Data

The study area for this experiment is Lagos, Nigeria (see Figure 1). 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 2). The land use data of Lagos consist remotely sensed Landsat Thematic Mapper images, acquired in 1978, 1984, and 2000 respectively; and an analogue base map acquired in 1963. The Landsat images were classified with the *k*-means algorithm using the MATLAB software. The analogue map was processed with ArcGIS. The analogue and remote sensing data were geo-referenced to ensure both data were in the same coordinate system. Twelve land use independent variables were used for the experiment. They were grouped into two categories: (i) proximity variables: distance to water, distance to residential structures, distance to industrial and commercial centres, distance to major roads, distance to railway, distance to Lagos Island, distance to international airport (1984-2000 only), distance to international seaport, distance to University of Lagos, distance to Lagos State University (1984-2000 only); and (ii) weighted variables: income potential and population potential. The proximity variables were extracted with the GIS while the weighted variables were extracted in MATLAB.

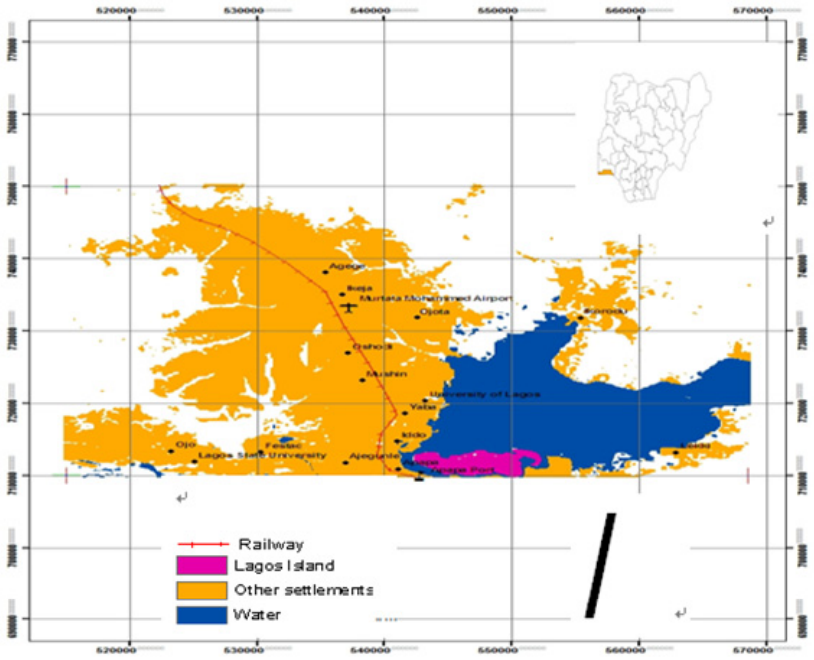

Figure 1. Lagos in relation to Nigeria

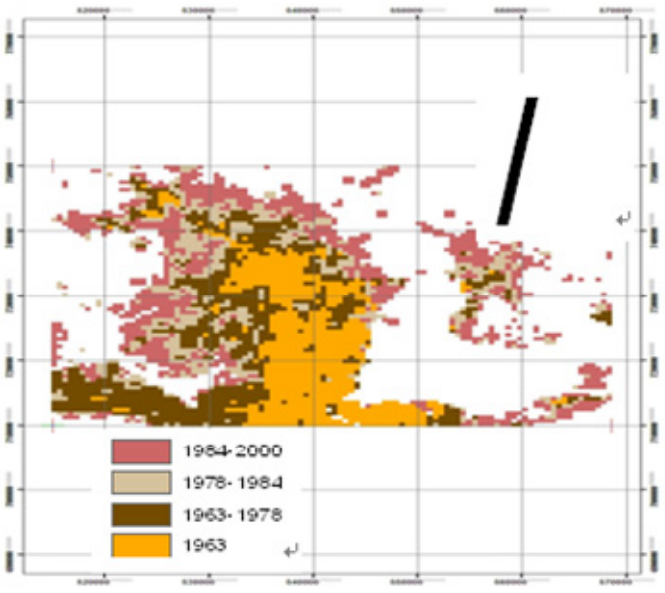

Figure 2. Land use of Lagos between 1963 and 2000

4.2 Modelling

The modelling was implemented and MATLAB and visualised in ArcGIS. The training data were extracted using the stratified random sampling. The training data consist of developed and undeveloped cells. Developed cells were labelled +1 while undeveloped cells were labelled -1. The SVM 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. The training data must not be extracted from the regions where change occurred among the three periods, since those change regions are not common to both maps. The polynomial, RBF, and linear kernels were used for all the experiments. Water and developed cells are considered immutable in the modelling.

The SVM parameters (regularisation parameter C, polynomial kernel degree d, and RBF kernel gamma γ) were optimised using a k-fold cross-validation procedure (where k=10). The optimisation equations were solved using Quadratic Programming (QP) (Gunn, 1998; Vapnik, 2000). Ten designated C values, *ln10e0, ln10e1, ln10e2, ln10e3, ln10e4, ln10e5, ln10e6, ln10e7, ln10e8, and ln10e9* were used to perform the k-fold cross-validation. The training data were split into 10 equal datasets. Nine datasets out of the 10 datasets were put together to train the model, while the remaining one dataset was used to test the model. The process was repeated until all the 10 datasets were used as both training and test sets. The designated values for the determination of optimal values for γ and d were 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. The cross-validation results for C with respect to the RBF, polynomial, and linear kernels are given in Figure 3. The cross-validation results for determining optimal values for γ and d are depicted in Figure 4.

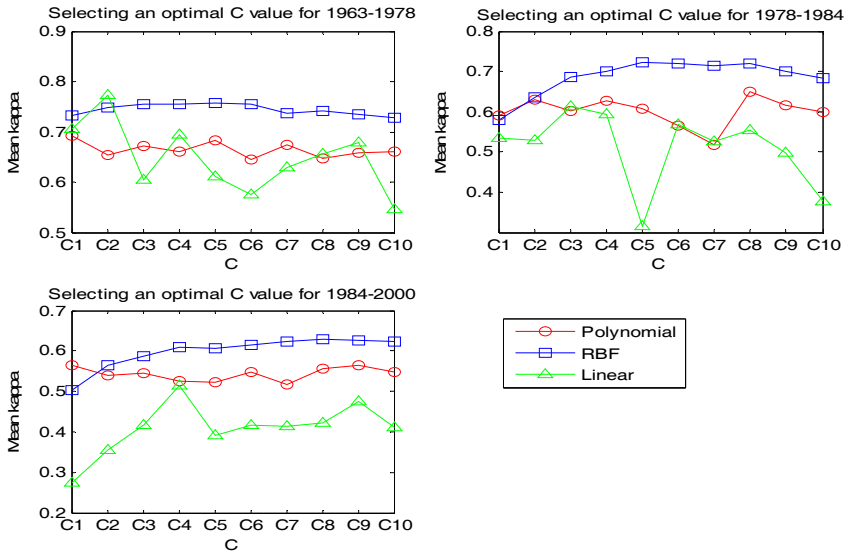

Figure 3. Selecting an optimal C value for periods 1963-1978, 1978-1984, and 1984-2000 (C1=ln10e0, C2=ln10e1, C3=ln10e2, C4= ln10e3, C5= ln10e4, C6= ln10e5, C7= ln10e6, C8= ln10e7, C9= ln10e8, and C10= ln10e9)

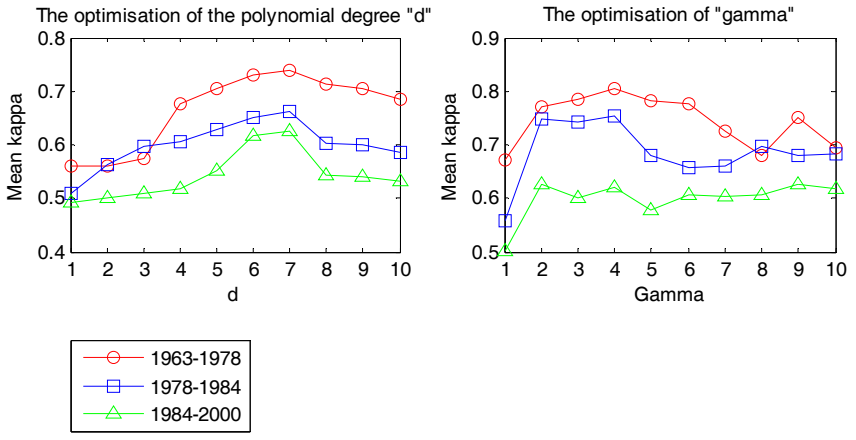

Figure 4. Cross-validation results for obtaining optimal values for d and gamma

Two hundred iterations were run to determine the best predictions for periods 1963-1978, 1978-1984, and 1984-2000. Figure 5 shows the plotted mean kappa coefficients for periods 1963-1978, 1978-1984, and 1984-2000; obtained by running the CA model at each 20 designated iteration thresholds. The calculated mean Kappa statistic for linear, polynomial, and RBF kernels were 1963-1978: 0.4475, 0.4869, and 0.5216; 1978-1984:

0.5324, 0.6497, and 0.6283; and 1984-2000: 0.5518, 0.6224, and 0.6386. The respective designated number of iterations for the calculated mean kappa results for linear, polynomial, and RBF kernels were 1963-1978: 140th, 90th, and 100th; 1978-1984: 120th, 90th, and 70th; and 1984-2000: 120th, 140th, and 90th.

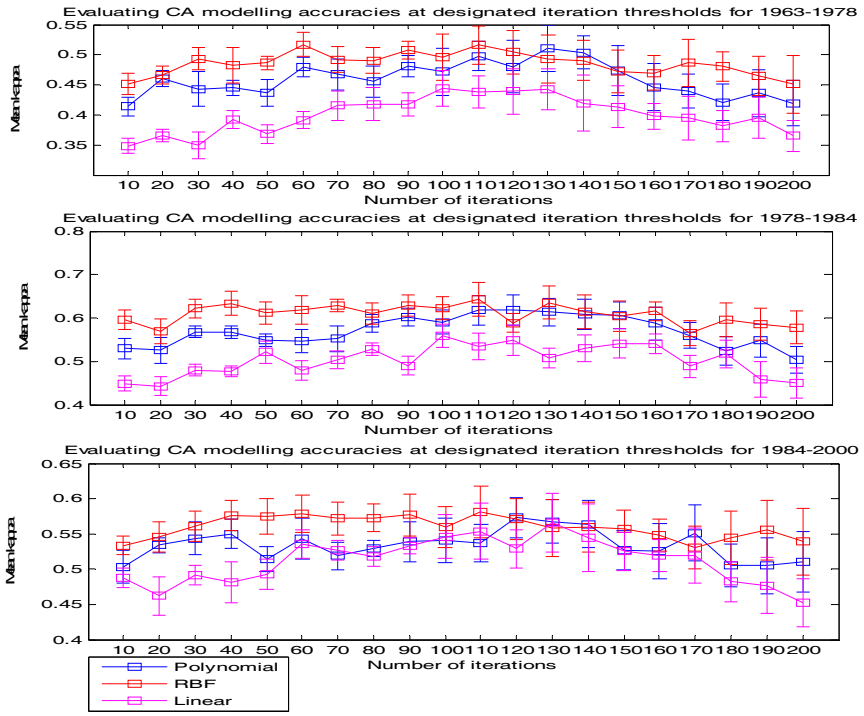

Figure 5. Computed mean kappa statistic and standard deviations for 200 designated iteration thresholds

The Receiver Operating Characteristics (ROC) was also used to assess the performance of the SVM-based CA model. The ROC is the plot of sensitivity against 1-specificity. The Area Under Curve (AUC) determines the result of the plot. Experiments that yield AUC indices <0.5 are usually regarded as worthless. Figures 6 depicts the plots 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. The mean sensitivity and mean 1-specificity were calculated by comparing the simulated maps with the actual maps. The computed AUC resulting from ROC for polynomial, RBF, and linear kernels respectively for 1963-1978: 0.7629 $\pm$ 0.0270, 0.7863 $\pm$ 0.0273, 0.7464 $\pm$ 0.0287; for 1978-1984: 0.8003 $\pm$ 0.0300, 0.8139 $\pm$ 0.0248, 0.7673 $\pm$ 0.0214; and for 1984-2000: 0.7804 $\pm$ 0.0316, 0.7939 $\pm$ 0.0290, 0.7714 $\pm$ 0.0255.

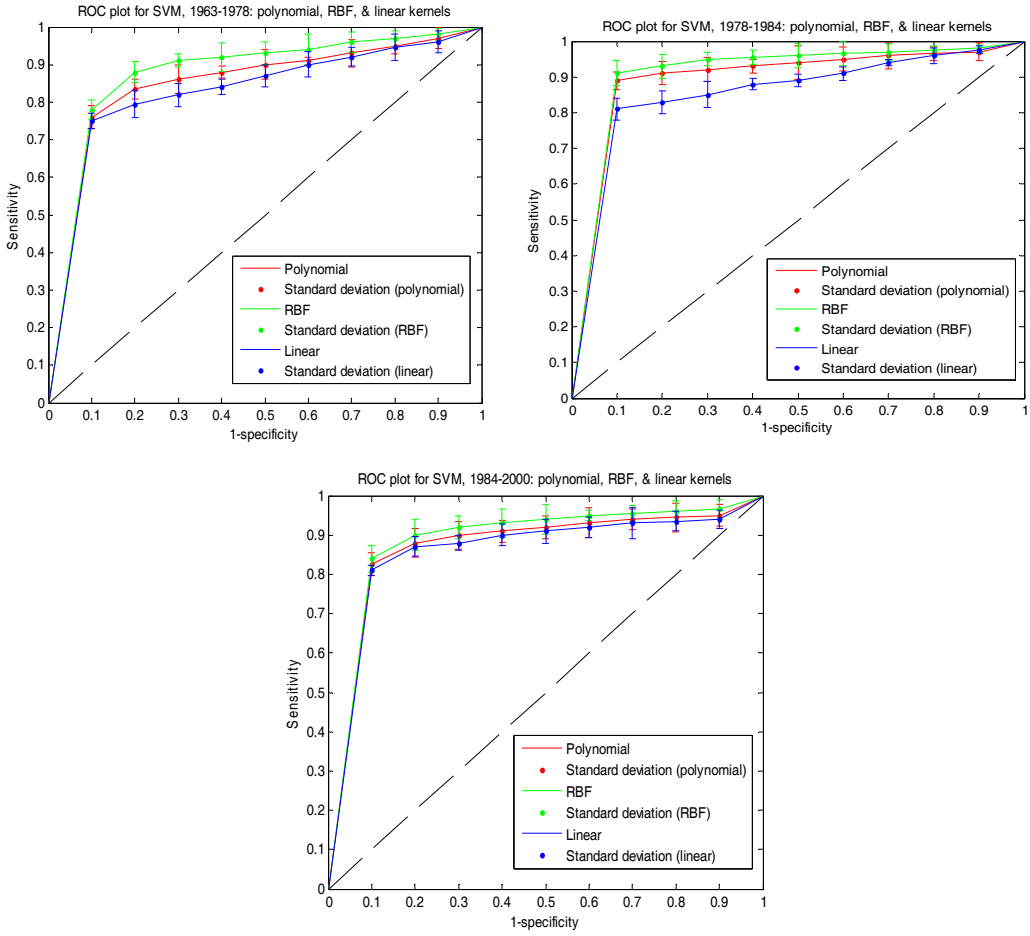

Figure 6. ROC plot for periods 1963-1978, 1978-1984, and 1984-2000

The historical land use change from 1963-1978, 1978-1984, and 1984-2000 was used to forecast the most probable land use maps in 2015 and 2030. The future land use maps in 2015 and 2030 were derived by running the SVM based CA model iteratively. The predicted maps in 2015 and 2030 using the polynomial, RBF, and linear kernels functions are depicted in Figure 7.

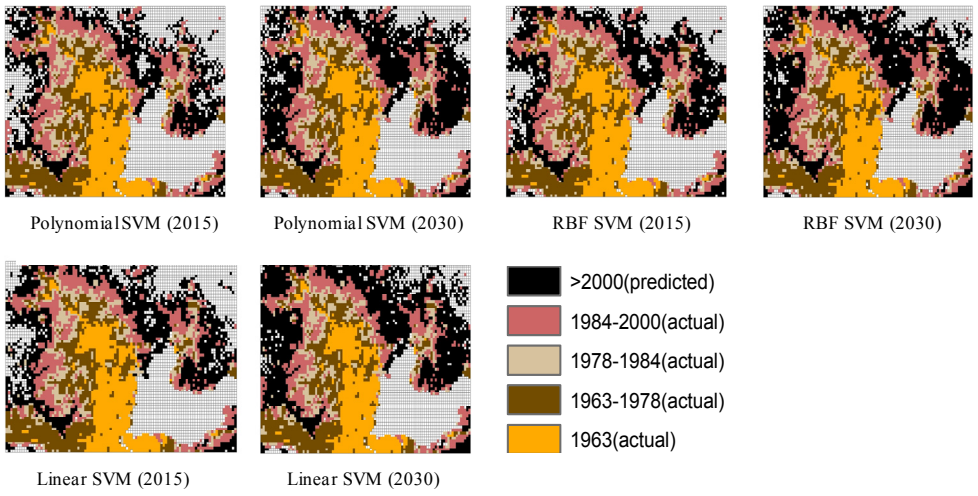

Figure 7. Predicted land use maps in 2015 and 2030

5. Conclusion

From Figure 3, 1963-1978, the optimal C values using polynomial, RBF, and linear kernels were ln10e0, ln10e5, and ln10e1 respectively. For 1978-1984, the optimal C values for the polynomial, RBF, and linear kernels were ln10e7, ln10e4, and log1n10e2 respectively. For 1984-2000, the optimal C values for polynomial, RBF, and linear kernels were ln10e0, ln10e7, and ln10e3 respectively. From Figure 4, the optimal values of d was found to be 7 for the three periods 1963-1978, 1978-1984, and 1984-2000; while the optimal value for γ was found to be 4 for periods 1963-1978 and 1978-1984. The optimal gamma γ value for period 1984-2000 was 2. From Figure 6, The RBF kernel yielded the highest AUC estimate, followed by the polynomial and linear kernels respectively. The computed ROC results corroborated the kappa statistic results. The order of performance of the three kernel functions based on kappa and AUC estimates was: RBF, polynomial, and linear. The computed AUC and Kappa statistic results from this experiment indicate substantial agreement between the actual and the predicted maps of Lagos. The satisfactory results from this experiment imply that the support vector machine based GIS cellular automata model is a promising tool for predicting land use change.

References

Barredo, J. I., Demicheli, L., Lavelle, C., Kasanko, M., & McCormick, N. (2004). Modelling future urban scenarios in developing countries: An application case study in Lagos, Nigeria. *Environment and Planning B: Planning and Design, 32*, 65-84. http://dx.doi.org/10.1068/b29103

Braimoh, A. K., & Onishi, T. (2007). Spatial determinants of urban land use change in Lagos, Nigeria. *Land Use Policy, 24*(2), 502-515. http://dx.doi.org/10.1016/j.landusepol.2006.09.001

Cortes, C. & Vapnik, V. (1995). Support vector networks. *Machine Learning 20*(3), 273-297.

Gunn, S. (1998). *Support vector machines for classification and regression* (Technical Report, ISIS). Southampton, England: Department of Electronics and Computer Science, University of Southampton.

Oduwaye, L. (2009). Challenges of sustainable physical planning and development in metropolitan Lagos. *Journal of Sustainable Development, 2*(1), 159-171.

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.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61-74). Cambridge, MA: MIT Press.

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. http://dx.doi.org/10.1068/b2802ed

Vapnik, V. N. (2000). *The nature of statistical learning theory*. New York, NY: Springer-Verlag.

White, R., & Englelen, G. (1993). Fractal urban land use patterns: A cellular automata approach. *Environment and Planning A. 25*, 1175-1199. http://dx.doi.org/10.1068/a251175

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.