OPTIMUM PREDICTIVE MODEL FOR URBAN GROWTH PREDICTION

Suwit Ongsomwang[1*] and Apiradee Saravisutra[2]

Received: Dec 13, 2010; Revised: Feb 8, 2011; Accepted: Feb 8, 2011

Abstract

This study aims to establish a framework to identify an optimum predictive model for simulation of urban dynamics in Nakhon Ratchasima province, Thailand. This study focuses on 2 different stochastic algorithms including the Cellular Automata Markov and logistic regression models. The core input data from interpreted land use in 1986 and 1994 were used to predict land use in 2002. The results are then compared with interpreted land use in 2002 to identify an optimum predictive model for urban growth. Results showed that the CA-Markov model provided higher overall accuracy and kappa hat coefficient of agreement for urban growth prediction in 2002 than the logistic regression model. Therefore, the CA-Markov model was chosen as an optimum predictive model for urban growth in 2010 and 2018. It was found that the urban and built-up area increased by about 36.32 sq. km (4.8% of the study area) between 2002 and 2010 while it increased by about 70.87 sq. km (9.4 % of the study area) between 2002 and 2018.

Keywords: **Urban growth prediction, CA-Markov model, logistic regression model, Nakhon Ratchasima province**

Introduction

According to area, Nakhon Ratchasima is the biggest province in Thailand and by the number of population, it is the second province after Bangkok, with a population of 2565117 in 2008 (Department of Provincial Administration, 2009). The township municipality of Nakhon Ratchasima was upgraded to a city municipality in 1995 as its population exceeded 50000 (Nakhon Ratchasima City Municipality, 2009). At present, Mueang Nakhon Ratchasima district has become a fast-growing urban area with rapid population growth. According to the statistics, in 2002 about 7.51% of Mueang Nakhon Ratchasima district's population lived in the municipality area and this proportion increased to 9.44% in 2008 (Department of Provincial Administration, 2009).

The population increase indicates that this district became much more populous and this should be the main factor of urbanization

[1] *School of Remote Sensing, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima, Thailand, 30000; E-mail: suwit@sut.ac.th*

[2] *Department of Geography, Faculty of Humanities and Social Science, Prince of Songkla University (Pattani Campus), Thailand, 94000; E-mail: sapirade@bunga.pn.psu.ac.th*

* *Corresponding author*

and land use change (Knox, 1994; Seto *et al.*, 2002). There are many studies on land use/land cover change related to global change or global warming, as these human activities have affected the climate and ecosystem (López *et al.*, 2001). During recent decades, many researches have focused on urban land use changes (Shenghe *et al.*, 2002; Xiao *et al.*, 2006) because urban ecosystems are strongly affected by human activities and they have a close relationship with the life of almost half of the world's population (Stow and Chen, 2002).

Effective analysis and monitoring of land cover changes require a substantial amount of data about the Earth's surface. This is most widely achieved by using remote sensing tools (Araya and Cabral, 2010). Basically, developed models can be roughly divided into 2 types; process-based and data-based models. Among the process-based models, Cellular Automata (CA) is widely used in simulation of urban sprawl (Batty *et al.*, 1999). CA constitutes a possible approach to urban growth modelling by simulating spatial processes as a discrete and dynamic system in space and time (Alkheder, 2005). Logistic regression is data-based and is a common method for empirically predicting the probabilities of events (Fang *et al.*, 2005). Both urban models have no consensus on the selection method of appropriate urban growth parameters and transition rules, since they depend mainly on the examined urban structures and dynamics. Therefore, 2 predictive models: CA-Markov and logistic regression were examined to identify an optimum predictive model for urban growth in the future.

Materials and Methods

Study Area

Mueang Nakhon Ratchasima district of Nakhon Ratchasima province, located in the northeast of Thailand, was selected as the study area (Figure 1). It is located between latitudes 14° 47′ 11″ to 15° 8′ 30″ North and

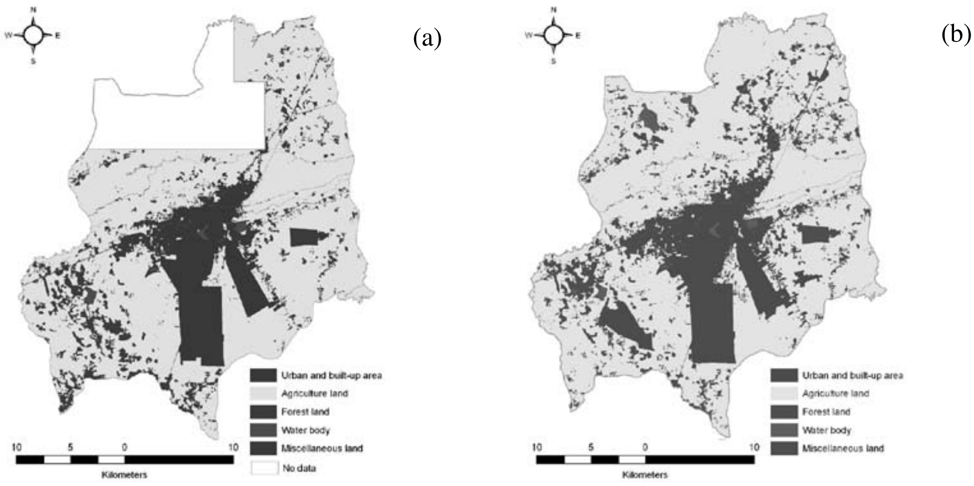
Figure 1. Study area

longitudes 101° 56′ 4″ to 102° 14′ 3″ East with a total area of 773.49 sq. km.

Data

Land use data of Mueang Nakhon Ratchasima district, Nakhon Ratchasima province in 1986, 1994, and 2002, were extracted from visual interpretation of aerial photographs based on the element of image interpretation (e.g., size, shape, tone color, texture, pattern, and site/situation and association) by the screen digitizing method at the scale of 1:10000. Owing to the land use classification system of the Land Development Department, there are 5 main land use types: urban and built-up area, agriculture land, forest land, water body, and miscellaneous land (Figure 2). Table 1 and Figure 3 summarized and compared the land use statistics in 1986, 1994, and 2002, respectively.

The accuracy of land use in 2002 was verified by calculating the overall accuracy and Kappa hat coefficient of agreement of the interpreted land use. Herein, the interpreted land use types in 2002 were evaluated with the reference data derived from the field survey in 2010. There were 127 randomly stratified sampling points based on the multinomial distribution theory with desired level of confident 90 percent and a precision of 10 percent. The error matrix between land use in 2002 and the reference land use data from



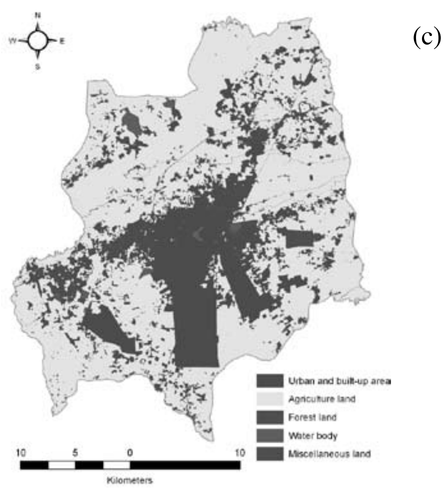


Figure 2. Distribution of land use in 1986, 1994, and 2002
(a) Land use in 1986
(b) Land use in 1994
(c) Land use in 2002

field survey in 2010 was shown in Table 2. It was found that the overall accuracy and Kappa hat coefficient of agreement for the 5 main land use types was 100% and 1.0, respectively.

In addition, selected variables as driving forces for urban growth based on relevant research papers, (including Allen and Lu, 2003; Hu and Lo, 2007; Luo and Wei, 2009), were collected to create an urban growth equation using stepwise regression analysis as shown in Table 3. These variables included population density, urban growth, urban and built-up area, agricultural land, forest land, water body, miscellaneous land, distance to existing urban area, distance to railway station, distance to main roads, and road density.

Methods

In this study, the CA-Markov and logistic regression models are firstly selected to predict urban growth and then their results are compared with the interpreted land use in 2002. After that the model which provides higher accuracy will be used as the optimum predictive model for urban growth in 2010

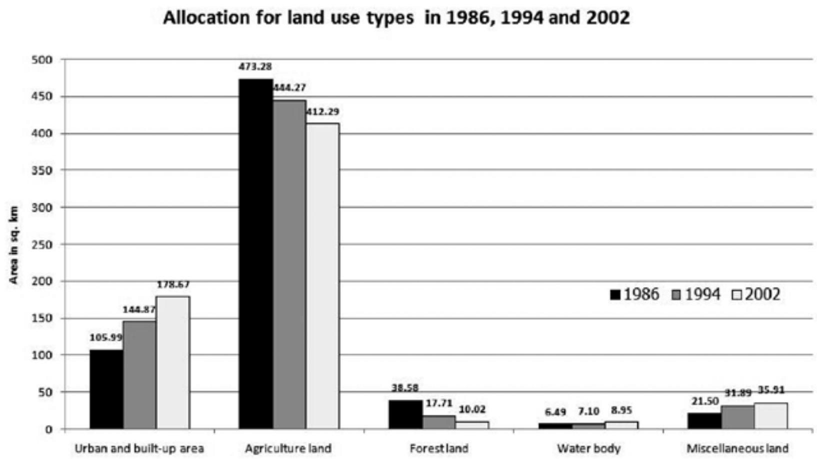


Figure 3. Comparison of land use types in 1986, 1994, and 2002

Table 1. Allocation for land use categories in 1986, 1994, and 2002

Land use types	1986		1994		2002	
	sq. km	%	sq. km	%	sq. km	%
Urban and built-up area	105.99	16.41	144.87	22.43	178.67	27.67
Agriculture land	473.28	73.28	444.27	68.79	412.29	63.84
Forest land	38.58	5.97	17.71	2.74	10.02	1.55
Water body	6.49	1.01	7.10	1.10	8.95	1.38
Miscellaneous land	21.50	3.33	31.89	4.94	35.91	5.56
Total area	645.84	100.00	645.84	100.00	645.84	100.00

Note: Total area was adapted according to available aerial photographs in 1986 and 2018. Major tasks of the methodology included urban growth prediction in 2002 using the CA-Markov and logistic regression models, identification of an optimum predictive model for urban growth, and prediction of urban growth in 2010 and 2018 and were summarized as follows:

(1) Urban growth prediction in 2002 using the CA-Markov model:-

(a) Cellular Automata

Automata are a useful abstraction of "behaving objects" for many reasons, as they can provide principally an efficient formal mechanism for representing their fundamental

Table 2. Error matrix for accuracy assessment of land use types in 2002

Land use types in 2002	Reference Data in 2010					
	U	A	F	W	M	Total
Urban and built-up area (U)	38	0	0	0	0	38
Agriculture land (A)	0	58	0	0	0	58
Forest land (F)	0	0	2	0	0	2
Water body (W)	0	0	0	14	0	14
Miscellaneous land (M)	0	0	0	0	15	15
Total	38	58	2	14	15	127

Note: 1. Overall accuracy = 100 %

2. Kappa hat coefficient of agreement = 1.0

Table 3. List of variables for regression model

Variable	Meaning	Nature of variable
Dependent		
UG	1 – urban growth; 0 – not urban growth	Dichotomous
Independent		
URBAN	1 – urban and built-up area; 0 – not urban and built-up area	Design
AGRI	1- agriculture land; 0- not agriculture land	Design
FOREST	1- forest land ; 0 – not forest land	Design
WATER	1- water body ; 0 – not water body	Design
MISC	1- miscellaneous land ; 0- not miscellaneous land	Design
DIST_URBAN	Distance to existing urban	Continuous
DIST_MRD	Distance to main roads	Continuous
DIST_RAIL	Distance to railway station	Continuous
POP_DEN	Population density (person/km^2)	Continuous
RD_DEN	Road density (m/ km^2)	Continuous

properties: attributes, behaviours, relationships, environments, and time. Formally, a finite automaton (A) can be represented by means of a finite set of states S = {S1, S2, …, SN} and a set of transition rules (T):

$$A \sim (S, T) \tag{1}$$

Two popular automata types that provide the basis for geographic automata are cellular automata and multiagent systems (Benenson and Torrens, 2004).

Cellular automata are dynamic models being discrete in time, space, and state. A simple cellular automaton A is defined by a lattice (L), a state space (Q), a neighbourhood template (δ), and a local transition function (f):

$$A = (L, Q, \delta, f) \tag{2}$$

Each cell of L can be in a discrete state out of Q. The cells can be linked in different ways. Cells can change their states in discrete time-steps. Usually cellular automata are synchronous, i.e. all cells change their states simultaneously. The fate of a cell is dependent on its neighbourhood and the corresponding transition function f (Adamatzky, 1994).

(b) Markov process

Formal definition of the Markov process is very close to that of CA. The Markov process is considered in discrete time and characterized by variables that can be in 1 of N states from S = {S1, S2,…,SN}. The set T of transition rules is substituted by a matrix of transition probabilities (P) and this is reflective of the stochastic nature of the process:

$$P = \left\| p_{ij} \right\| = \left\| \begin{matrix} p_{1,1} & p_{1,2} & \cdots & p_{1,N} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,N} \\ \cdots & \cdots & \cdots & \cdots \\ p_{N,1} & p_{N,2} & \cdots & p_{N,N} \end{matrix} \right\| \tag{3}$$

where p_{ij} is the conditional probability that the state of a cell at moment t+1 will be S_j, given it is S_i at moment t:

$$\text{Prob}\,(S_i \rightarrow S_j) = p_{ij} \tag{4}$$

The Markov process as a whole is given by a set of status S and a transition matrix P. By definition, in order to always be "in one of the state" for each i, the condition $\sum_j p_{ij} = 1$ should hold (Benenson and Torrens, 2004).

(2) Urban growth prediction in 2002 using the logistic regression model

The nature of the land use and land cover change of a cell is dichotomous: either the presence of urban growth or absence of urban growth. The binary values 1 and 0 are used to represent urban growth and no urban growth, respectively (Hu and Lo, 2007). The general form of logistic regression is described as follows:

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m \tag{5}$$

$$y = \log_e \left(\frac{P}{1-P} \right) = \text{logit}(P) \tag{6}$$

$$P = \frac{e^y}{1+e^y} \tag{7}$$

where x_1, x_2, … x_m are explanatory variables, and y is a linear combination function of the explanatory variables representing a linear relationship (Equation 5). The parameters b_1, b_2, …b_m are the regression coefficients to be estimated. P means the probability of occurrence of a new unit. Function y is represented as logit(P) (Equation 6). In logistic regression, the probability value can be a non-linear function of the explanatory variables (Equation 7) (Cheng and Masser, 2003).

(3) Identification of an optimum predictive model for urban growth

The optimum predictive model for urban growth between the CA-Markov and logistic regression models is justified based on the overall accuracy and kappa hat coefficient of agreement. The model which generates the higher accuracy will be used for urban growth prediction in 2010 and 2018.

(4) Urban growth prediction in 2010 and 2018

An optimum predictive model which was derived from the previous task will be used for urban growth prediction in 2010 and 2018. Herein, land use data in 1994 and 2002 were used for urban growth prediction in 2010 while land use in 2002 and predicted land use in 2010 were used to predict urban growth in 2018.

Results and Discussion

Urban Growth Prediction in 2002 using the CA-Markov Model

Urban growth prediction using the CA-Markov was performed under the Markov module of IDRISI software. In this study, land use types in 1986 and 1994 were used to generate a transition area matrix and a transition probability matrix between 1986 and 1994 as shown in Table 4 and Table 5, respectively. Then, a transition probability matrix will be applied to create a set of conditional probability data for the 5 major land use types between 1986 and 1994

After deriving the outputs of the Markov model, a transition area matrix, a set of conditional probability data between 1986 and 1994, and the original land use types in 1994 were exported into the CA-Markov module of IDRISI software to create predicted land use types in 2002 based on the Markov chain analysis and multi-criteria evaluation/multi-objective land allocation routines (Figure 4). Table 6 summarized the area of predicted land use types in 2002. The predicted urban

Table 4. Transition area matrix for land use change between 1986 and 1994

Land use in 1994	Land use types (sq. km)					
	U	A	F	W	M	Total
Urban and built-up area (U)	144.87	0.00	0.00	0.00	0.00	144.87
Agriculture land (A)	26.24	403.33	0.00	1.25	13.46	444.27
Forest land (F)	1.81	6.09	8.13	0.03	1.65	17.71
Water body (W)	0.03	0.56	0.00	6.05	0.46	7.10
Miscellaneous land (M)	10.34	1.22	0.00	0.24	20.09	31.89
Total	183.29	411.20	8.13	7.57	35.65	645.84

Table 5. Transition probability matrix for land use change between 1986 and 1994

Land use in 1994	Land use types					
	U	A	F	W	M	Total
Urban and built-up area (U)	1.000	0.000	0.000	0.000	0.000	1.000
Agriculture land (A)	0.059	0.908	0.000	0.003	0.030	1.000
Forest land (F)	0.102	0.344	0.459	0.002	0.093	1.000
Water body (W)	0.004	0.079	0.000	0.852	0.065	1.000
Miscellaneous land (M)	0.324	0.038	0.000	0.008	0.630	1.000

and built-up area in 2002 covered an area of 183.29 sq. km.

Urban Growth Prediction in 2002 using the Logistic Regression Model

A logistic regression model is used to associate urban growth with driving force factors and to generate urban growth probability data. Herein, the urban growth equation with 7 driving force factors based on stepwise regression analysis of data in 1986 and 1994 was used as the initial equation for the logistic regression model to predict urban growth in 2002 as shown in the following equation.

$$\begin{aligned} UG_{86\_94} = {} & 0.454 + 0.781\,AGRI_{86} + 0.385 MISC_{86} \\ & + 0.193\,DIST\_URBAN_{86} + 0.114\,FOREST_{86} \\ & - 0.093\,URBAN_{86} + 0.027\,POP\_DEN_{94} \\ & - 0.026\,DIST\_MRD_{94} \end{aligned} \tag{8}$$

where:

$AGRI_{86}$ is agriculture land in 1986;
$MISC_{86}$ is miscellaneous land in 1986;
$DIST\_URBAN_{86}$ is distance to existing urban area in 1986;
$FOREST_{86}$ is forest land in 1986;
$URBAN_{86}$ is urban and built up area in 1986;
POP_DEN_{94} is population density in 1994;
$DIST\_MRD_{94}$ is distance to main road in 1994.

This result indicates that the urban growth between 1986 and1994 was driven spontaneously and naturally by socio economic development such as land use change.

In addition, predicted urban growth in 2002 based on spatial data in 1986 and 1994 was an output from the LOGISTICREG module of IDRISI software, having probability values between 0.11 and 0.82 (Figure 5). These probability values for urban growth (0.11-0.82) were firstly used to extract the predictive urban and built-up area in 2002. After that we compared this predictive result with the interpreted urban and built-up area in 2002 using overall accuracy and kappa analysis assessment. These 2 accuracy assessment methods and the probability values for urban growth are calculated to derive a threshold value setting as suggested by Yang *et al.* (2006) as shown in Figure 6. The best threshold value

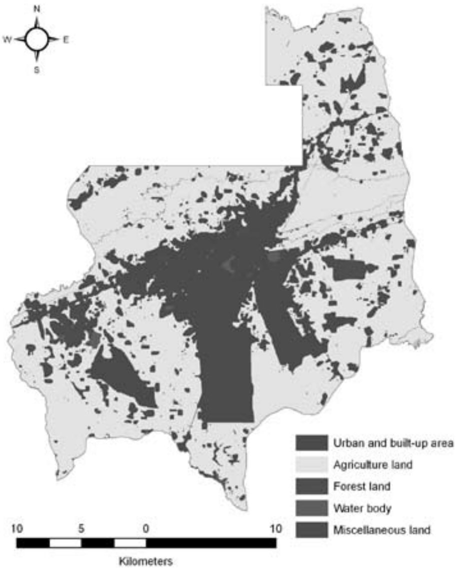


Figure 4. Predicted land use types in 2002 by CA-Markov model

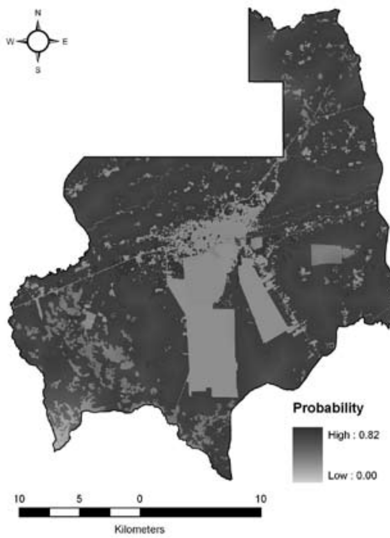


Figure 5. Probability value of predictive urban growth by logistic regression model

with high overall accuracy and kappa hat coefficient of agreement is 0.79 which is used to generate the predictive urban and built-up area in 2002. The urban and built-up area and non-urban and built-up area in 2002 were 121.21 sq. km and 524.63 sq. km, respectively. It was found that this result was underestimated by about 30% because this model used only the urban and built-up area in 1986 for the prediction of probability data.

An Optimum Predictive Model for Urban Growth

The predicted urban and built-up areas resulting from the CA-Markov and logistic regression models were here justified based on the accuracy assessment with an interpreted urban and built-up area in 2002. It was found that the overall accuracy and kappa hat coefficient of agreement for predictive urban and built-up area in 2002 using the CA-Markov model was 93.41% and 0.84, respectively. In the meantime, the overall accuracy and kappa hat coefficient of agreement for predictive urban and built-up area in 2002 using the logistic regression model was 89.41% and 0.71, respectively. Therefore, the CA-Markov model that provides higher overall accuracy and kappa hat coefficient of agreement will be here chosen as the optimum predictive model for urban growth in 2010 and 2018.

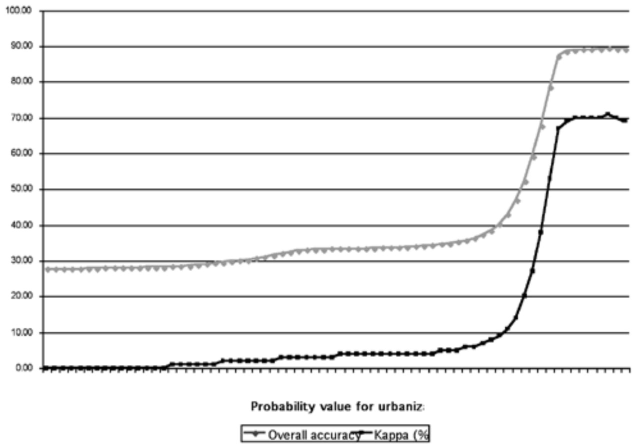


Figure 6. Relationship between overall accuracy and kappa hat coefficient (%) of predictive urban and built-up area in 2002 and probability values for urban growth

Table 6. Area and percentage of predicted land use types in 2002 using CA-Markov

Land use types	Area in sq. km	Percentage
Urban and built-up area	183.29	28.38
Agriculture land	411.20	63.67
Forest land	8.13	1.26
Water body	7.57	1.17
Miscellaneous land	35.65	5.52
Total	645.84	100.00

This result was consistent with the study of Cetin and Demirel (2010). After simulating urban dynamics of the Istanbul Metropolitan area with the logistic regression and CA based Markov models, they found that the CA-Markov model was better than the logistic regression model. In addition, Araya and Cabral (2010) mentioned that the CA-Markov analysis not only considers a change from non-built-up to built-up areas, but it could also include any kind of transition among any feature classes.

Prediction of Urban Growth in 2010 and 2018

From the previous section, the CA-Markov model provided higher overall accuracy and kappa hat coefficient of agreement than the logistic regression model, and thus the CA-Markov model is here selected to predict the urban growth in 2010 2018. The area of predicted land use in 2010 and 2018 was summarized and compared as shown in Table 7 and Figure 7, respectively. Distribution

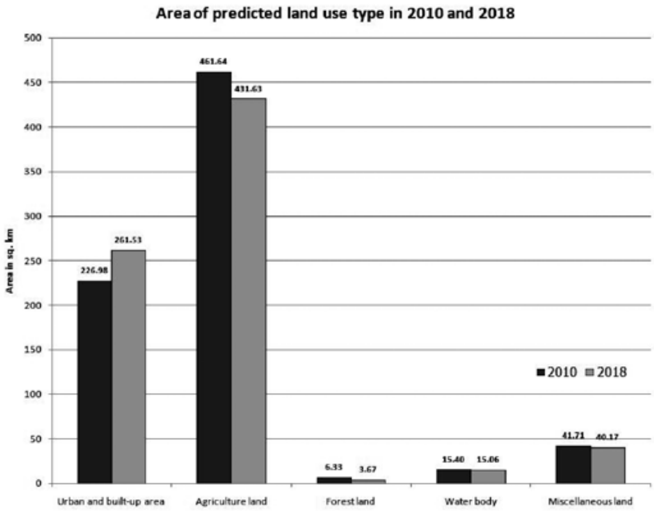


Figure 7. Comparison of predicted land use types in 2010 and 2002

Table 7. Area and percentage of predicted land use types in 2010 and 2018

Land use types	2010		2018	
	Area in sq. km	**Percentage**	**Area in sq. km**	**Percentage**
Urban and built-up area	226.98	30.18	261.53	34.78
Agriculture land	461.64	61.38	431.63	57.39
Forest land	6.33	0.84	3.67	0.49
Water body	15.40	2.05	15.06	2.00
Miscellaneous land	41.71	5.55	40.17	5.34
Total	752.06	100.00	752.06	100.00

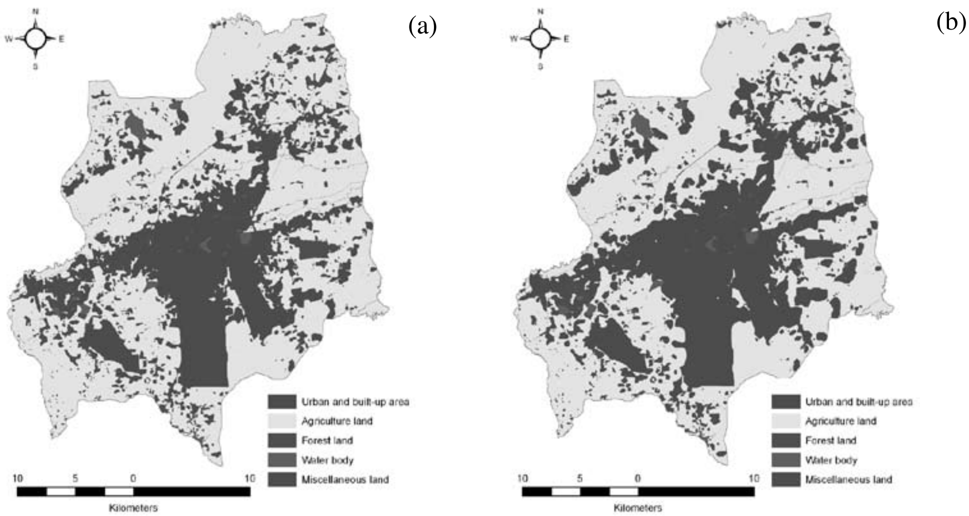


Figure 8. Distribution of predicted land use in 2010 and 2018 by CA-Markov model

of predicted land use in 2010 and 2018 was displayed in Figure 8.

Conclusions

Predictive models have shown a great potential to support planning and management decisions. With the help of the models, knowledge and understanding of the dynamics of the systems are gained, future trends are forecast, impacts can be analyzed, and various policies can be simulated and optimized. In this study, as the CA-Markov model provided higher overall accuracy and kappa hat coefficient of agreement for urban growth prediction in 2002 than the logistic regression model, the CA-Markov model was selected to be the optimum predictive model for urban growth in 2010 and 2018.

For prediction of urban growth in 2010 using the CA-Markov model based on land use types in 1994 and 2002, it was found that the predicted urban and built-up area in 2010 was about 226.98 sq. km. For prediction of urban growth in 2018 using the CA-Markov model based on the interpreted land use types in 2002 and predicted land use types area in 2010, it was found that the predicted urban and built-up area in 2018 was about 261.53 sq. km.

References

Adamatzky, A. (1994). Identification of Cellular Automata. Taylor & Francis Ltd, London, 384p.

Alkheder, S. and Shan, J. (2005). Cellular automata urban growth simulation and evaluation-A case study of Indianapolis. Geomatics Engineering, School of Civil Engineering, Purdue University, Indiana. Available from: http://www.geocomputation.org/2005/Alkheder.pdf. Accessed data: Oct 29, 2010.

Allen, J. and Lu, K. (2003). Modeling and prediction of future urban growth in the Charleston region of South Carolina. Conserv. Ecol., 8(2):2.

Araya, Y.H. and Cabral, P. (2010). Analysis and modeling of urban land cover change in Setúbal and Sesimbra, Portugal. Remote Sens., 2:1549-1563.

Batty, M., Xie, Y., and Sun, Z. (1999). Modeling urban dynamics through GIS-based cellular automata. Comput. Environ. Urban, 23:205-233.

Benenson, I. and Torrens, P.M. (2004). Introduction to urban geosimulation. In: Geosimulation: Automata-based Modeling of Urban Phenomena. Benenson, I. and

Torrens, P.M. (eds). John Wiley & Sons, Chichester, UK, p. 1-19.

Cetin, M. and Demirel, H. (2010). Modelling and simulation of urban dynamics [On-line]. Available from: http://www.ins.itu.edu.tr/jeodezi/fotog/hdemirel/duyuru/MESAEP-demirel_paper23022010.pdf. Accessed data: May 29, 2010.

Cheng, J. and Masser, I. (2003). Urban growth pattern modeling: a case study of Wuhan city, PR China. Landscape Urban Plan., 62:199-217.

Department of Provincial Administration. (2009). Statistics of population [On-line]. Available from: http://www.dopa.go.th/xstat/popyear.html. Accessed date: Jan 10, 2010.

Fang, S., Gertner, G.Z., Sun, Z., and Anderson, A.A. (2005). The impact of interactions in spatial simulation of the dynamics of urban sprawl. Landscape Urban Plan., 73:294-306.

Hu, Z. and Lo, C.P. (2007). Modeling urban growth in Atlanta using logistic regression. Comput. Environ. Urban, 31:667-688.

Knox, P.L. (1994). Urbanization: An introduction to urban geography. Prentice-Hall, Englewood Cliffs, NJ, USA.

López, E., Bocco, G., Mendoza, M., and Duhau, E. (2001). Predicting land-cover and land-use change in the urban fringe: A case in Morelia city, Mexico. Landscape Urban Plan., 55:271-285.

Luo, J. and Wei, Y. (2009). Modeling spatial variations of urban growth patterns in Chinese cities: The case of Nanjing. Landscape Urban Plan., 91:51-64.

Nakhon Ratchasima City Municipality. (2009). General information data of Nakhon Ratchasima city municipality [On-line]. Available from: http://www.koratcity.net/main_menu/general/general01.html. Accessed date: Jan 10, 2010.

Seto, K.C., Woodcock, C.E., Song, C., Huang, X., Lu, J., and Kaufmann, R.K. (2002). Monitoring land-use change in the Pearl River Delta using Landsat TM. Int. J. Remote Sens., 23(10):1985-2004.

Shenghe, L., Prieler, S., and Xiubin, L. (2002). Spatial patterns of urban land use growth in Beijing. J. Geogr. Sci., 12(3):266-274.

Stow, D.A. and Chen, D.M. (2002). Sensitivity of multitemporal NOAA AVHRR data of an urbanizing region to land-use/land-cover changes and misregistration. Remote Sens. Environ., 80:297-307.

Xiao, E., Shen, Y., Ge, J., Tateishi, R., Tang, C., Liang, Y., and Huang, Z. (2006). Evaluating urban expansion and land use change in Shijiazhuang, China, by using GIS and remote sensing. Landscape Urban Plan., 75:69-80.

Yang, X., Skidmore, A.K., Melick, D.R., Zhou, Z., and Xu, J. (2006). Mapping non-wood forest product (Matsutake mushrooms) using logistic regression and a GIS expert system. Ecol. Model., 198:208-218.