Chapter 5
Implementation of Traditional Techniques

5.1 Introduction

The inherent aim here is to implement the traditional and common methodologies which have been employed to illustrate land use change, and thereafter to simulate its forthcoming status. In this chapter, the cellular automata model, the Markov chain model, the CA-Markov model and the logistic regression model will be designed and executed. Each single model will be evaluated to verify its outcomes. This will allow us to validate their results and acquire enough assurance of their performance. Thus, verified models will be chosen in order to integrate in the ABM.

5.2 Selected Techniques for Implementation

In this part of the chapter, it is intended to review and also execute preferable and useful methodologies such as cellular automata, Markov chain, cellular automata Markov, and logistic regression models. The outcomes of these models will be evaluated and the different results will enable us to compare them to each other. The strategy of creating different results by means of different techniques will enable this research to represent various methods upon a specific area. Therefore, a general flowchart for this section can be presented as in Fig. 5.1.

In Sect. 5.3, we will describe the theoretical background of the aforementioned models, as well as their implementation, starting with the cellular automata model.

J. Jokar Arsanjani, *Dynamic Land-Use/Cover Change Simulation: Geosimulation and Multi Agent-Based Modelling*, Springer Theses,
DOI: 10.1007/978-3-642-23705-8_5, © Springer-Verlag Berlin Heidelberg 2012

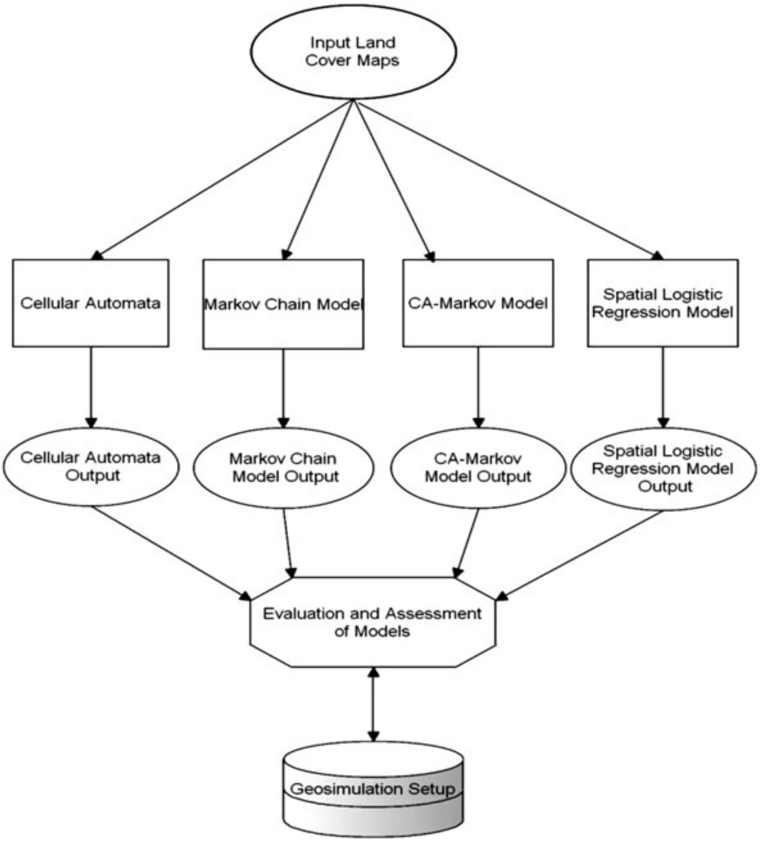


Fig. 5.1 Flowchart of the general strategy for the implementation of the models

5.3 Cellular Automata Model Scenario

In recent decades, investigations for developing geographical cellular automata in order to simulate complex systems have been raised. Cellular automata have been employed to simulate wildfire propagation (Goodchild et al. 1996), population dynamics (Couclelis 1985), and land use change (Batty and Xie 1994; White and Engelen 1993).

The cellular automata model is known as CA which is a dynamic model originally conceived by Ulam and Von Neumann in the 1940s to afford a formal framework for investigating the behaviour of complex systems (Moreno et al. 2009). CA is also the main framework of agent-based modelling scenarios. Land use changes simulation using CA is a complicated process, whereas various spatial variables and factors have to be employed (Li 2008). A critical matter in CA modelling is defining appropriate transition rules based on training data. In fact, these transition rules conduct this model. Linear boundaries have been used to

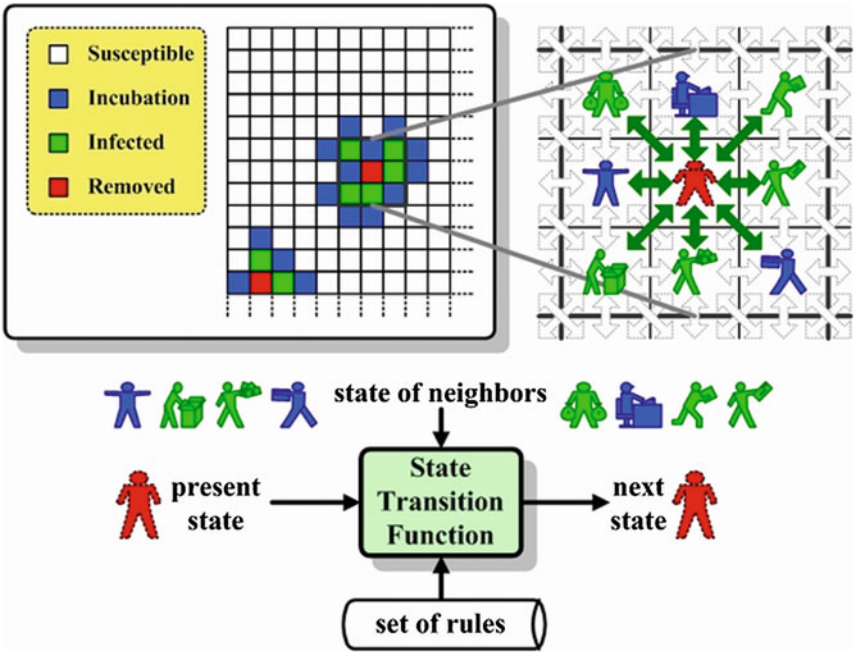


Fig. 5.2 Cellular Automata, state transition rules and the Moore neighbourhood notion (Huang et al. 2004)

define the rules (see Fig. 5.2). However, land use dynamics or changes, and many other geographical phenomena, are vastly complex and require nonlinear boundaries for the rules definition (Moreno et al. 2009). Figure 5.3 demonstrates the flowchart of implementing the CA Model.

5.3.1 CA Transition Rules

Land use changes on the fringe of cities (i.e. urban sprawl) is the consequence of both internal and external forces; the internal impact means an area tends to continue its development if it has begun to develop from a rural to an urban status, particularly if this natural tendency is supported by development from within the neighbourhood. The external impact means factors such as the geographical conditions of the area, socio-economic circumstances and institutional controls, also impact on the process of development. Physical constraints (e.g. water bodies and steep terrain, etc.) restrict or slow down the development of urban areas (Fig. 5.4).

Socio-economic factors, such as land availability and demands on available lands, accessibility to nodes of employment, accessibility to public services and facilities, such as schools, shops, public transport, and contiguity to existing urban areas also play key roles in urban development; therefore, they are able to define

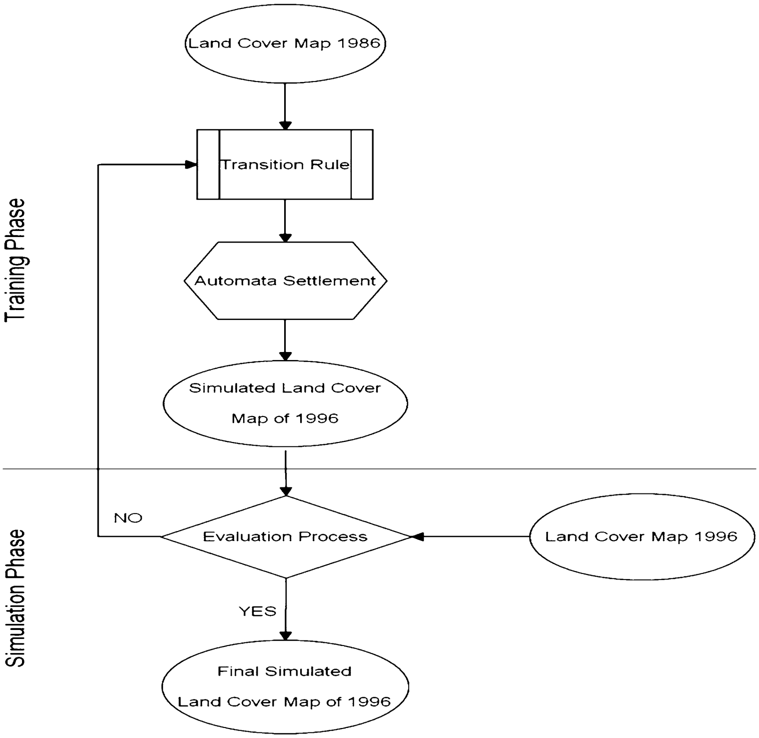


Fig. 5.3 Flowchart of the implementation procedure of the CA scenario

appropriate conditions (Liu 2008). The transition rules are the major inputs in a CA model. Basically, the aforementioned rules have been defined in linear forms, using methods such as multi-criteria evaluation (MCE) (Yang et al. 2008). Transition rules can be defined through a filter file at a variety of kernel sizes, and various decision rules can make that CA model completely different from other existing CA models. Whereas these simulated maps are on hand, a training phase can be utilised by means of these preliminary results and the map of reality. This training phase helps to realise the appropriate kernel size and transition rules.

5.3.2 Training Process and Calibration of the CA Model

The training process consists of choosing a certain time step for the simulation through the CA model. Different transition rules and neighbourhood distances can result in various outcomes; thus, a preliminary evaluation of the obtained results was carried out to pick the optimum settings. The optimum settings will lead us to implement this model by coordination of time step and results. Accordingly, after

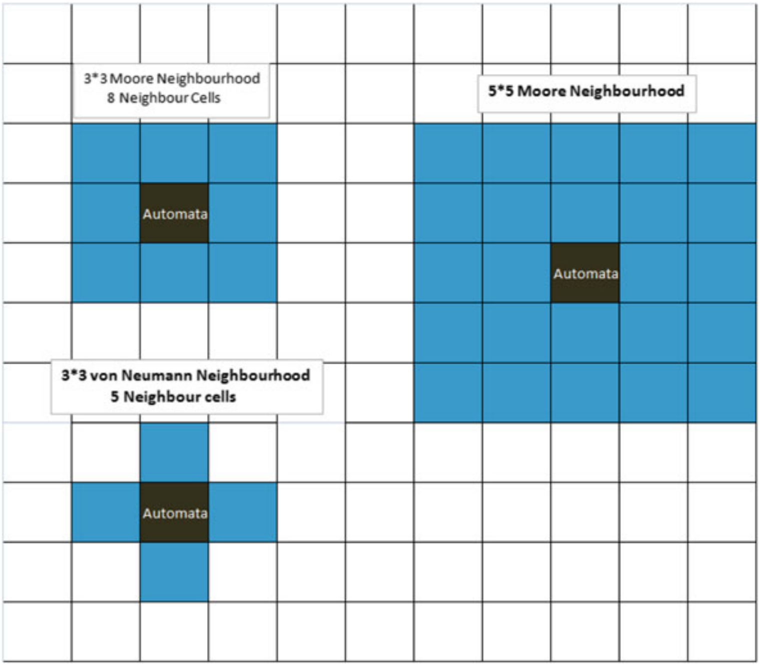
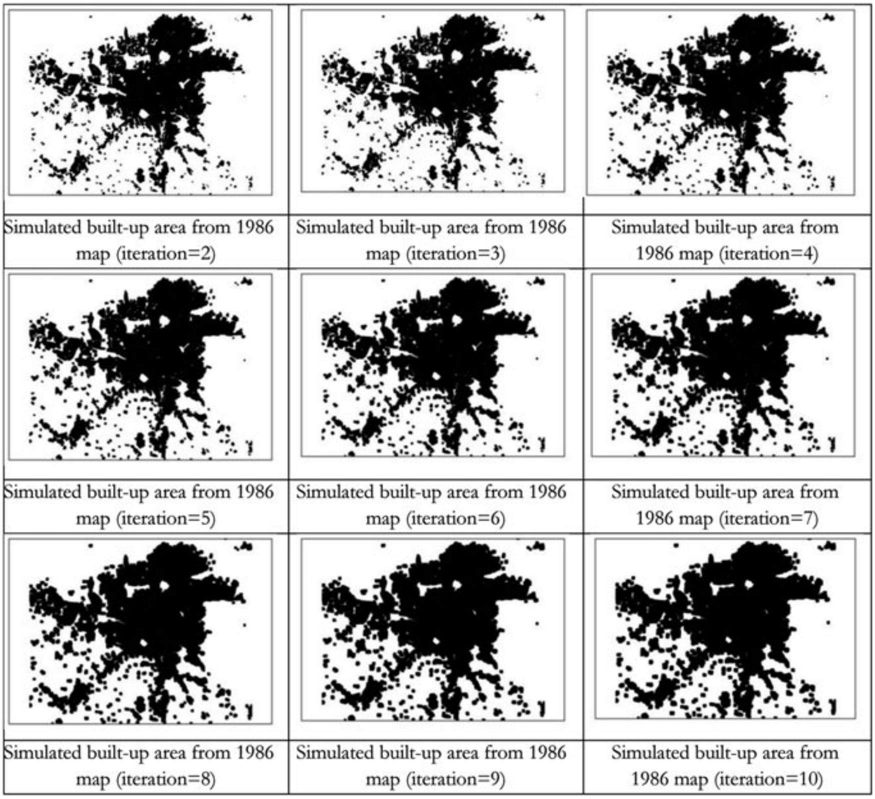


Fig. 5.4 Schematic explanation of automata and different neighbourhood layouts

implementing the training phase and retrieving calibrated factors, a simulated map of development of forthcoming years was prepared.

The other key issue to implement a CA model is to estimate an appropriate iteration number. This enables users to stop the modelling process at accurate times. Therefore, a training process is applied to the model in order to control the predefined transition rules. This helps to stop our model at a certain time and reach a certain amount of change, to better estimate the locations of changes. A code was written in the Python environment and imported into the ArcGIS Toolbox. This script has the typical characteristics of a CA model. The code comprises all cellular automata components, i.e. neighbourhood size and transition rules. This CA code performs according to a predefined iteration number, and it stops at a certain time. At each time step, a filter is applied to the entire image then the output image is reclassified according to the reclassification file, and the produced output image is then used as an input for the next iteration. The process goes on until the predefined iteration number is reached.

Results from different settings can be evaluated and compared with the actual map. This model was implemented to the land use maps of 1986 and 1996 to achieve the simulated maps of 1996 and 2006, respectively. The simulated maps of built-up areas at different iteration numbers of 1986, 1996 and 2006 are shown through Figs. 5.5 and 5.6.



Simulated built-up area from 1986 map (iteration=2)

Simulated built-up area from 1986 map (iteration=3)

Simulated built-up area from 1986 map (iteration=4)

Simulated built-up area from 1986 map (iteration=5)

Simulated built-up area from 1986 map (iteration=6)

Simulated built-up area from 1986 map (iteration=7)

Simulated built-up area from 1986 map (iteration=8)

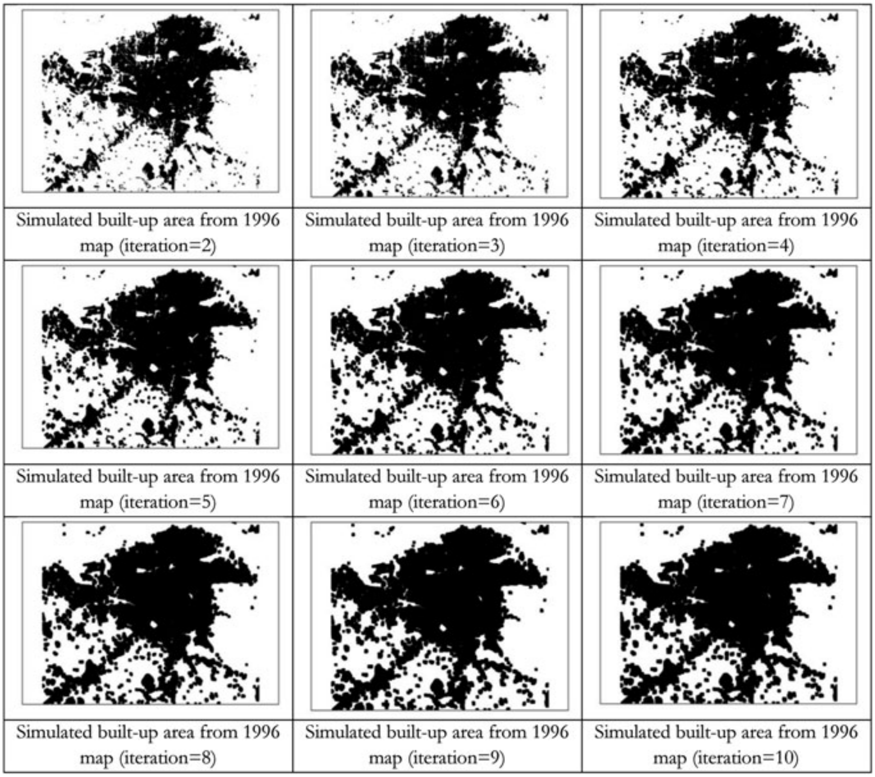
Simulated built-up area from 1986 map (iteration=9)

Simulated built-up area from 1986 map (iteration=10)

Fig. 5.5 Simulated maps of built-up areas at different iteration numbers from 1986 to 1996

The simulated maps of 1996 and 2006 were compared with the maps of reality of 1996 and 2006; therefore, the optimum transition rules and settings can be determined. In Table 5.1, the number of iterations and resulted ROC values are cross compared to pick the optimum iteration number. The determined transition rules will be chosen as the optimum designed CA model. This model will be implemented on the map of 2006 in order to simulate built-up map of 2016.

Table 5.1 shows that the maximum ROC value yielded at iteration number nine, therefore, this amount of iteration and associated transition rules were employed for the prediction process. The model validation process and resulted map will be presented in Chap. 7 (see Sect. 7.5.1).



Simulated built-up area from 1996 map (iteration=2)

Simulated built-up area from 1996 map (iteration=3)

Simulated built-up area from 1996 map (iteration=4)

Simulated built-up area from 1996 map (iteration=5)

Simulated built-up area from 1996 map (iteration=6)

Simulated built-up area from 1996 map (iteration=7)

Simulated built-up area from 1996 map (iteration=8)

Simulated built-up area from 1996 map (iteration=9)

Simulated built-up area from 1996 map (iteration=10)

Fig. 5.6 Simulated maps of built-up areas at different iteration numbers from 1996 to 2006

5.4 The Markov Chain Model Scenario

Markov chain theory is a stochastic process theory that describes how likely one state is to change to another state. The Markov chain has a key-descriptive tool which is its transition probability matrix (TPM). Markov chain theory has been used generally to study water resource systems and simulate precipitation sequences, particularly to describe and predict lithological transition, plant succession, and land utilisation change (Li et al. 1999).

Stochastic processes generate sequences of random variables$\{X_n,\ n \in T\}$ by probabilistic laws. In Eq. 5.1, index n stands for time. This process is measured discrete in time and $T = \{0, 5, 10, \ldots\}$ years approximately. This time step is a reasonable time unit for land use change studies. Therefore, if the stochastic process considered a Markov process then the sequence of random variables will be produced by the Markov property, formally (Cabral and Zamyatin 2009):

$$P[X_{n+1} = a_{in+1} | X_0 = a_{i0}, \ldots, X_{in} = a_{in}] = P[X_{in+1} = a_{in+1} | X_{in} = a_{in}] \quad (5.1)$$

Table 5.1 Comparison of different accuracy assessment indices arising from diverse CA rules

Input file	Iteration number	Kappa index for built-up cells	Overall kappa	ROC value
1986 (predicted 1996)	2	0.6684	0.6953	0.837
	3	0.6684	0.6953	0.837
	4	0.7156	0.7000	0.846
	5	0.7546	0.6984	0.851
	6	0.7877	0.6925	0.854
	7	0.8151	0.6830	0.857
	8	0.8379	0.6705	0.86
	9	0.8562	0.6553	0.861
	10	0.8709	0.6382	0.859
1996 (predicted 2006)	3	0.6977	0.6912	0.825
	4	0.7475	0.6909	0.834
	5	0.7869	0.6838	0.84
	6	0.8201	0.6729	0.843
	7	0.8488	0.6598	0.847
	8	0.8736	0.6447	0.851
	9	0.8944	0.6280	0.854
	10	0.9116	0.6099	0.852

5.4.1 Markovian Property Test

Land use change in the study area needs to be proved as a Markovian process. In fact, it must have statistical dependence between $X_{\mathrm{n}+1}$ and X_{n}; and that statistical dependence is a first-order Markov process.

$$P(X_n = a_n | X_{n-1} = a_{n-1}) \neq P(X_n = a_n) \times P(X_{n-1} = a_{n-1}) \tag{5.2}$$

$$P[X_n = a_n | X_{n-1} = a_{n-1}] = P[X_n = a_n, X_{n-1} = a_{n-1}] / P[X_{n-1} = a_{n-1}] \tag{5.3}$$

A first-order Markov process is defined as a Markov process that the transition from one category to any other categories does not necessitate intermediate transitions to other states. The statistical dependence can be tested in any contingency table demonstrating the land cover changes between X_{n} and $X_{\mathrm{n}-1}$. In this research, this test was performed for land cover changes between 1986–1996 and 1996–2006. To deduce from the association or independence between the land cover categories within the years from the contingency table, the random variable, with the chi-square distribution is defined by:

$$x^2 = \sum_i \sum_i \left(\left(N_{ij} - M_{ij} \right)^2 / M_{ij} \right) \tag{5.4}$$

Here, N is the contingency matrix showing the land cover change between two assumed time scales; for instance, either 1986–1996 or 1996–2006 or 1986–2006, and also, M the contingency matrix with the expected values of change, assuming the independence hypotheses.

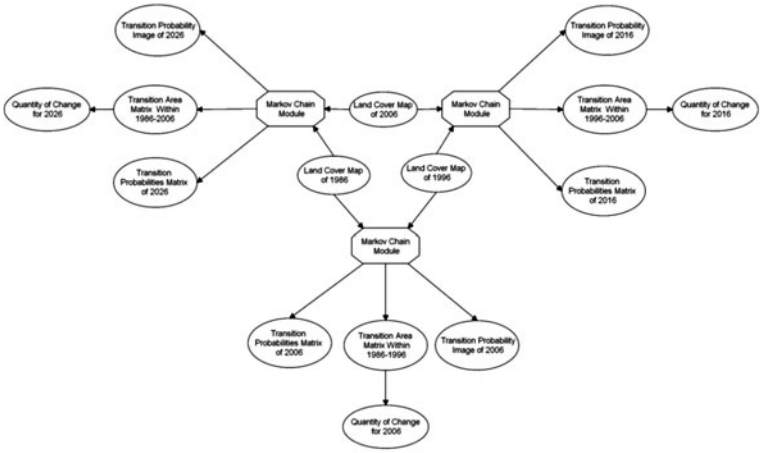

Transition Probability Image of 2026
Markov Chain Module
Land Cover Map of 2006
Markov Chain Module
Transition Probability Image of 2016
Quantity of Change for 2026
Transition Area Matrix Within 1986-2006
Transition Area Matrix Within 1996-2006
Quantity of Change for 2016
Transition Probabilities Matrix of 2026
Land Cover Map of 1986
Land Cover Map of 1996
Transition Probabilities Matrix of 2016
Markov Chain Module
Transition Probabilities Matrix of 2006
Transition Area Matrix Within 1986-1996
Transition Probability Image of 2006
Quantity of Change for 2006

Fig. 5.7 Schematic view of the Markov chain model approach

x^2 basically measures the distance between the actual values of land cover change and the projected ones, assuming independence hypothesis and accordingly must be high enough to verify. The same non-parametric test was performed to assess the Markovian property. Thus, the values have to be compared with the observed values computed with the Chapman–Kolmogorov equation, supposing that these variables are generated by a first-order Markov process:

$$P(X_n = a_n | X_m = a_m) = P(X_1 = a_1 | X_m = a_m) \times P(X_n = a_n | X_1 = a_1), \quad m \leq 1 \leq n \tag{5.5}$$

The Chapman–Kolmogorov equation expresses that the probability of transition between 1986 and 2006 can be projected by multiplying the transition probabilities matrix 1986–1996 by the transition probabilities matrix 1996–2006.

$$x^2 = \sum_i \sum_j \left(\left(N_{ij} - o_{ij} \right)^2 / o_{ij} \right) \tag{5.6}$$

5.4.2 Execution of the Markov Chain Module

The transition probabilities matrix is calculated by the contingency matrix displaying the relative frequencies of land change at a certain time period (Cabral and Zamyatin 2009). The IDRISI MARKOV module inputs a pair of land-cover images and outputs a transition probability matrix, a matrix of transition areas, as well as a set of conditional change probability images. A text file records the probability matrix that each land cover category will change to other categories under a certain probability value.

Table 5.2 Markov transition probabilities matrix between 1986–1996, 1996–2006 and 1986–2006

		Agricultural field	Built-up	Open land	Public park	Water body
Probability value of 2006 based on transition matrix of 1986–1996	Agricultural field	0.8835	0.0487	0.0615	0.0062	0.0001
	Built-up	0.0007	0.9907	0.0054	0.0031	0.0001
	Open land	0.0133	0.0689	0.9124	0.0052	0.0001
	Public park	0	0.0335	0.0232	0.9428	0.0005
	Water body	0.0105	0	0	0	0.9895
Probability value of 2016 based on transition matrix of 1996–2006	Agricultural field	0.9361	0.0402	0.0218	0.0017	0.0002
	Built-up	0.0036	0.9873	0.0055	0.0035	0
	Open land	0.0144	0.0576	0.9223	0.0055	0.0003
	Public park	0.0018	0.0088	0.0064	0.9816	0.0013
	Water body	0.0211	0.0102	0	0.0066	0.9621
Probability value of 2026 based on transition matrix of 1986–2006	Agricultural field	0.8469	0.0936	0.0493	0.0096	0.0005
	Built-up	0.0003	0.9885	0.0059	0.0052	0.0001
	Open land	0.0188	0.1131	0.8579	0.0099	0.0003
	Public park	0	0.0434	0.0198	0.9362	0.0005
	Water body	0.0105	0.0009	0	0	0.9886

The transition area matrix is a table which records the amount of pixels that are anticipated to change from one land cover category to other category according to a number of time units. The produced results (i.e. matrices) arising from this implementation were stored for use in further change analyses. This output determines the estimated quantity of change that can be used for the process of change allocation. Figure 5.7 presents a schematic view of the implementation of the Markov chain scenario.

In effect the Markov chain is not a spatially explicit model; therefore the Markov chain is not an appropriate model to estimate the location of change, which is the aim of GIS projects. Nevertheless, it is an excellent quantity estimator (Kamusoko et al. 2009) such that its outcomes can be allocated by means of other approaches. As is shown in Table 5.2, the probability of converting each land category to the others can be determined by the Markov chain model.

5.5 Cellular Automata Markov Scenario

This section of the chapter aims, in particular, to depict the cellular automata Markov model and how this module was executed. The cellular automata Markov model that has been designed into the IDRISI software (Andes Version) is an extension of multi criteria evaluation procedure which combines CA and Markov chain modules. By using the quantity of change which is calculated through the

Markov chain analysis (i.e. transition area matrix) the cellular automata Markov model applies a contiguity kernel to ‘grow out’ a land use map to a later time period; hence, this approach converts the outcomes of the Markov chain model to a spatially explicit model by integration of CA functionality. The certainty and accuracy of this module will be examined and demonstrated (see Fig. 5.8).

Some efforts were performed to construct high-resolution regional models by integration of the Markov and CA approaches (Clark 1990), and investigations in this area have been growing extensively (Wegener 2001). The Markov cellular automata model is a robust approach in terms of quantity estimation as well as spatial and temporal dynamic modelling of land use/cover changes, because GIS and remote sensing data can be capably incorporated. Biophysical and socioeconomic data could be used, firstly, to define preliminary conditions; secondly, to parameterise the Markov cellular automata model; thirdly, to analyse transition probabilities and, finally, to determine the neighbourhood rules with transition potential maps (Kamusoko et al. 2009). In the cellular automata Markov model, the Markov chain process manages temporal dynamics among the land use/cover categories based on transition probabilities, while the spatial dynamics are controlled by local rules determined either by the cellular automata spatial filter or transition potential maps (Maguire et al. 2005). In fact, the cellular automata Markov model begins allocating changes from the nearest cells to each land use type (Pontius and Malanson 2005).

In this section, the future land use/cover changes (up to 2026) in the study area were simulated based on the cellular automata Markov model, which combines Markov chain analysis and cellular automata models in order to change the essence of the Markov chain to a spatially explicit model.

The spatial resolution of output maps was defined at 30 m in accordance with Landsat imagery spatial resolution. The original cell size could avoid further uncertainty by employing reclassification functions. Hence, the quantity and percentage of each type of land use maps was calculated for the period of 1986–2006 in accordance with cross tabulation analysis.

5.5.1 Execution of the Cellular Automata Markov Model

Markov chain models have been broadly used to model land use changes including both urban and rural areas at coarse spatial scales. After preparing land use maps, transition probability matrices for both time periods were calculated as well as Markovian conditional probability images in IDRISI software (See Table 5.1).

The first record of Table 5.1 identifies the next 10-year step (i.e. 2006) as a description of transition probability matrix, where the agricultural areas category will remain at the same category at 88.35% probability and 4.87% will be converted to built-up area category. Furthermore, the value of the fourth row which identifies the probability of converting public parks category to agricultural land category is zero; in other words, it is not expected to observe any public park

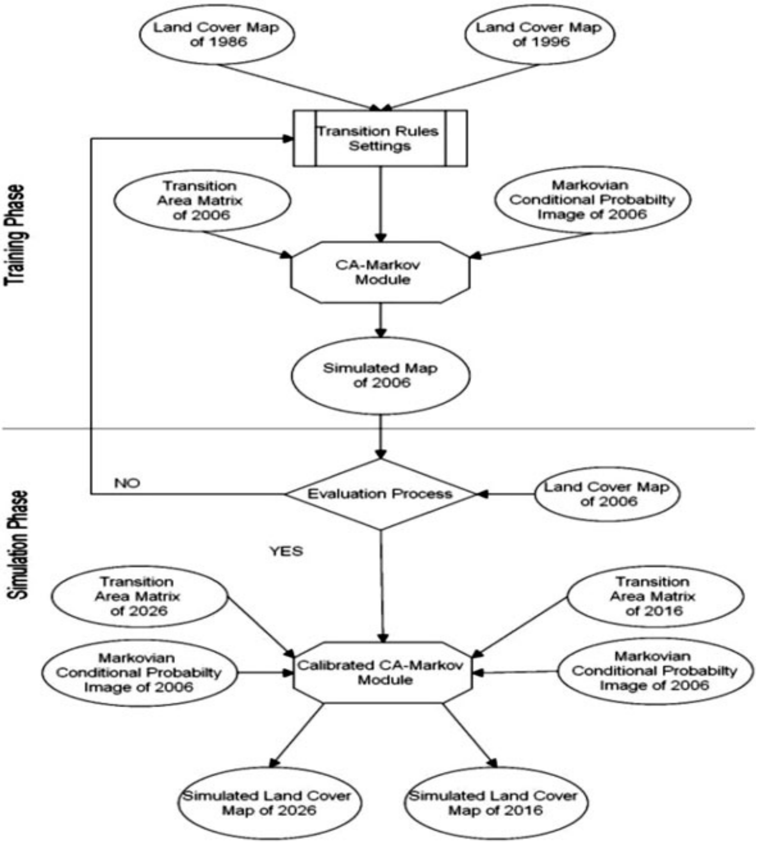
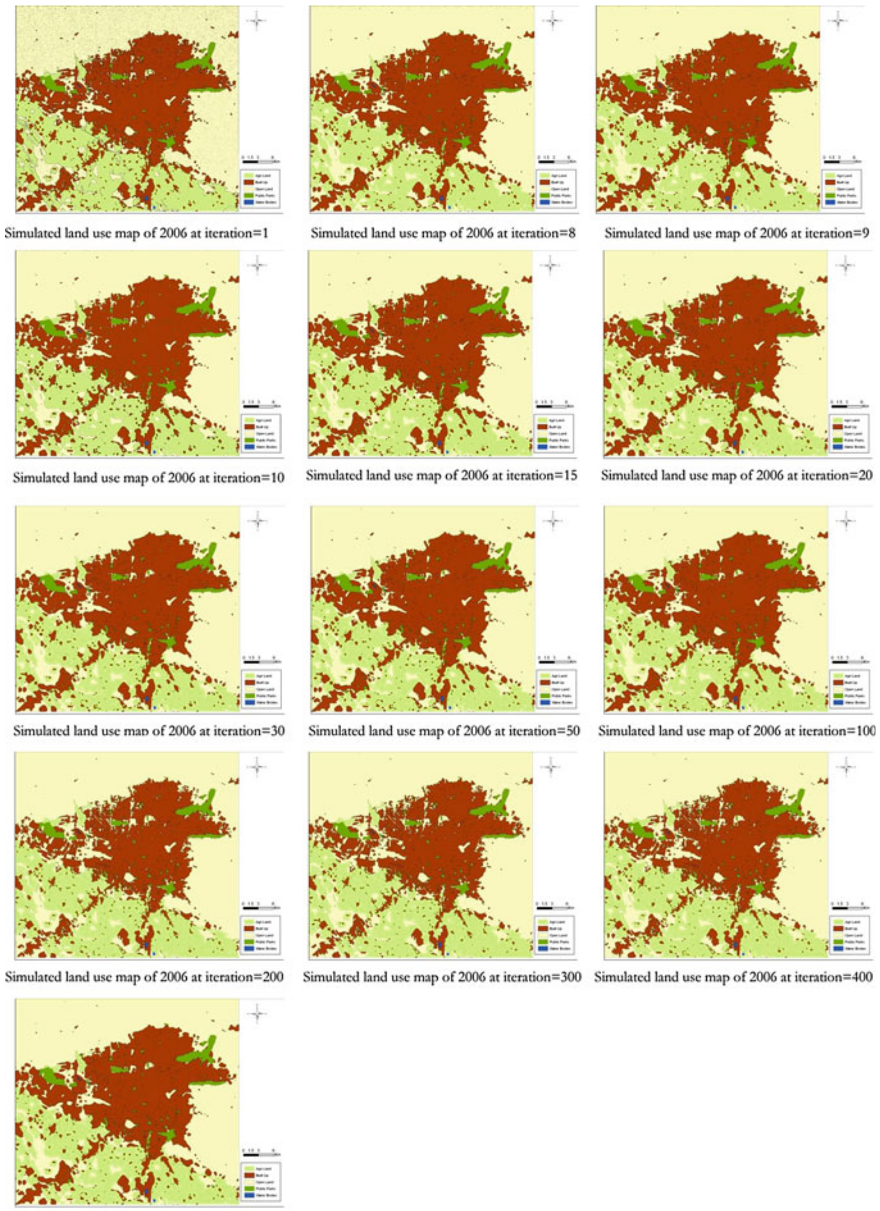

Land Cover Map of 1986
Land Cover Map of 1996
Transition Rules Settings
Training Phase
Transition Area Matrix of 2006
Markovian Conditional Probabilty Image of 2006
CA-Markov Module
Simulated Map of 2006
NO
Evaluation Process
Land Cover Map of 2006
Simulation Phase
YES
Transition Area Matrix of 2026
Transition Area Matrix of 2016
Markovian Conditional Probabilty Image of 2006
Calibrated CA-Markov Module
Markovian Conditional Probabilty Image of 2006
Simulated Land Cover Map of 2026
Simulated Land Cover Map of 2016

Fig. 5.8 Flowchart of the cellular automata Markov simulation process

cell that has been converted to agricultural field cells. Figure 5.9 demonstrates simulated maps arising from the implementation of the cellular automata Markov model at different iteration numbers.

The next step requires the need to set up the cellular automata Markov model for predicting the land use map. Since this module has Markovian property and CA behaviour, the cellular automata Markov model must be defined for both properties. Hence, by inputting the land use map of 1986, Markov transition areas parameters and transition suitability image parameters for Markovian property of the model were employed, as well as filter contiguity definition and number of iterations in support of cellular automata behaviour.

As shown in Fig. 5.8, it is aimed to simulate multiple land use maps for one-time step (e.g. 2006) by defining different transition rules. The simulated maps will be compared with the actual map, which allows us to evaluate the validity of this approach. Therefore, the verified model can be used to simulate future years.



Simulated land use map of 2006 at iteration=1

Simulated land use map of 2006 at iteration=8

Simulated land use map of 2006 at iteration=9

Simulated land use map of 2006 at iteration=10

Simulated land use map of 2006 at iteration=15

Simulated land use map of 2006 at iteration=20

Simulated land use map of 2006 at iteration=30

Simulated land use map of 2006 at iteration=50

Simulated land use map of 2006 at iteration=100

Simulated land use map of 2006 at iteration=200

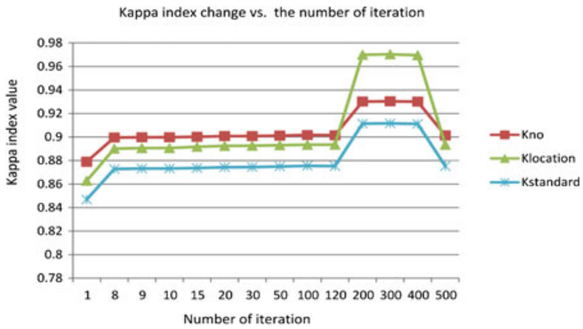
Simulated land use map of 2006 at iteration=300

Simulated land use map of 2006 at iteration=400

Simulated land use map of 2006 at iteration=500

Fig. 5.9 Simulated land use map of 2006 from land use map of 1996 at different iteration numbers

Fig. 5.10 Kappa indices at different iteration numbers



The land use maps of 1986 and 1996 were input to the cellular automata Markov model to produce a simulated map of 2006. This implementation requires a Markovian conditional probability image of 2006 and, also, a transition area matrix of 2006 to be input. Several types of filter contiguity and a number of iterations were examined to achieve the optimal kernel size and number of iterations. With the aim of reaching the optimal parameters, the simulated and actual land use maps of 2006 were crossed to validate the results. One of the setting parameters was to define the iteration number that will reproduce different maps. This model evaluation process needs to verify all the simulated maps to compare them with the actual map; consequently, the most statistically similar map will be selected. The predefined parameters will be chosen as the proper settings for the next runs.

The produced maps under different transition rules were assessed with different indices. A diagram of correlation between those maps and the number of iterations was accordingly drawn (see Fig. 5.10). The kappa indices of *location* and *quantity* for the simulated maps were calculated, and subsequently the most appropriate iteration number at iteration of 300 was determined with a Kappa standard index of 0.91. The input transition rules were considered in order to run this approach and predict future land use maps. This was done based on the transition probabilities matrices of land change (1996–2006) and land change (1986–2006). Markovian conditional probability images have to be input to derive the simulated land use maps of 2016 and 2026. Eventually, the simulation process of predicting the land use maps of 2016 and 2026 was implemented to output the respective maps. These maps are represented in Chap. 7 (Figs. 7.3, 7.4).

5.5.2 Validation of the Cellular Automata Markov Model

A cross comparison between the simulated maps at different iterations and actual maps was employed to verify the certainty of the model. The highest value of accuracy among the resultant maps was chosen, which is approximately 91% for the kappa index, and 97% for K-Location (Fig. 5.11). Investigation of this model shows that the cellular automata Markov model is a good estimator for the quantification of change and continuous-space change modelling. Based on visual analysis, this model

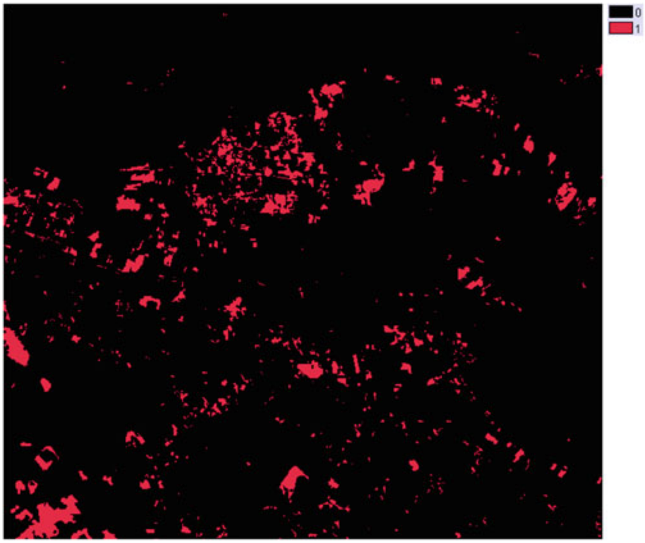


Fig. 5.11 Dependent variable Y; change to built-up area between 1986 and 1996 (no change: $Y = 0$; change $Y = 1$)

produces some diffused-speckle developed cells which do not correspond with the reality. Besides, this model needs a lot of time to run the simulation process and, also, to be replicated for a huge number of iterations (e.g. 300, 400). Although the simulated maps have high Kappa indices the edges of land categories appear wavy and circular in shape, which do not match with the reality, i.e. they seem unreal.

5.6 The Logistic Regression Model Scenario

The logistic regression analysis has been the most frequently used approach during the past two decades for predictive modelling by means of variation of inductive modelling (Verhagen 2007). Empirical estimation and dynamic simulation models have been used to simulate land use/cover changes. Various types of rule-based modelling (e.g. cellular automata model) are the most suitable models for incorporating spatial interaction effects and handling temporal dynamics. CA models, however, focus primarily on the simulation of spatial patterns rather than the interpretation of spatio-temporal processes of urban sprawl. There is a lack of incorporation between most dynamic simulation models over socioeconomic variables (Hu and Lo 2007). In this section, another approach by means of the logistic regression model on urban sprawl will be explained. The aim of executing this technique was to observe the presumed relationship and interactions between social, economic and environmental parameters which could drive urban expansion. As far as it has been realised, this technique has never been published or even employed upon the study area. Hence, this implementation and its outcomes could lead to more accurate results in this area of research, and achieve a better understanding of the interaction between those variables.

5.6.1 An Overview of the Logistic Regression Technique

Regression is a method to discover the coefficients of the empirical relationships from observations. Linear regression, log-linear regression and logistic regression are the most used regression approaches (Hu and Lo 2007). In logistic regression, the dependent variable can be either binary or categorical, and the independent variables could be a set of categorical and continuous variables. Routine assumption is not required for the logistic regression model. Hence, logistic regression is advantageous in comparison with the linear regression or log-linear regression. It is fundamental to extract the coefficients of independent variables from the observation of land use conversion, since urbanisation does not frequently follow typical supposition, and its prominent factors are usually a combination of continuous and categorical variables (Xie et al. 2005). The general form of logistic regression is as follows:

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_mx_m \tag{5.7}$$

$$y = \log_e\left(\frac{P}{1-P}\right) = \log\ it\,(p) \tag{5.8}$$

$$P = \frac{e^y}{1+e^y} \tag{5.9}$$

Where x_1, x_2, …, x_m are independent variables, y defines a linear combination function of the independent variables representing a linear relationship. Moreover, the b_1, b_2, …, b_m parameters are the regression coefficients to be retrieved. Function y is known aslog it (P) i.e. the logarithm (base-e) of the odds or likelihood ratio that the dependent variable Z is 1. Probability value (P) strictly increases while y value goes up. Regression coefficients b_1 to b_m imply the contribution of each independent variable on the probability value. A positive value implies that the independent variable helps to increase the probability of land change and a negative value implies the reverse effect. The statistical method is a multivariate estimation process which examines the relative significance and strength of the factors. While employing logistic regression to simulate rural–urban land transformation, it is crucial to consider the spatial heterogeneity of spatial data. Spatial statistics such as spatial dependence and spatial sampling also have to be taken into account to eliminate spatial autocorrelation (Hu and Lo 2007). Otherwise, unreliable factor estimation or unproductive estimates (i.e. wrong results) of the hypothesis test will be produced.

There are two basic approaches to assess spatial dependence: firstly, building a more complex model incorporating an autoregressive structure and, secondly, designing a spatial sampling plot to enlarge the distance interval between sampled points. Spatial sampling creates a smaller sample size that loses certain information and conflicts with the large sample of asymptotic normality of maximum likelihood method, upon which logistic regression is based on. Nonetheless, it is a reasonable approach to eliminate spatial auto-correlation, and a reasonable design of spatial sampling scheme will make an ideal balance between the two sides (Xie et al. 2005).

The logistic regression model is employed to predict a categorical variable from a set of predictor variables. A discriminated function analysis is generally employed if all of the predictors are continuous and properly distributed; Logit analysis is generally utilised if every predictor is categorical. In fact, logistic regression is often preferred if the predictor variables are a set of categorical and continuous variables. Besides, they should be properly distributed. The predicted dependent variable in a Logistic Regression Model is a function of the probability that a particular theme will be in one of the categories; for instance, the probability of change upon a specific land use based on a set of scores on the predictor variables such as proximity to interchange network, and so on (Huang et al. 2009).

LOGISTICREG module in IDRISI Andes performs binomial logistic regression, in which the input dependent variable must be binary in nature and can have only two possible values (0, 1). Such regression analysis is usually employed in the estimation of a model that depicts the relationship between continuous independent variables to a binary dependent variable. The basic assumption is that the probability of a dependent variable takes the value of 1 (positive response). The logistic curve and its value can be calculated with the following formula: (Mahiny and Turner 2003)

$$P(y = 1|X) = \frac{\exp(\sum BX)}{1 + \exp(\sum BX)} \tag{5.10}$$

Where:

P is the probability of the dependent variable occurrence
X is the independent variables, $X = (x_0, x_1, x_2 ... x_k), \quad x_0 = 1;$
B is the estimated parameters, $B = (b_0, b_1, b_2 ... b_k)$

In order to linearize the above model, as well as remove the 0/1 boundaries for the original dependent variable which is probability, the following transformation is usually applied:

$$\mathrm{P}' = \mathrm{In}(p/(1 - p)) \tag{5.11}$$

This transformation is referred to as the Logit or logistic transformation. Thus, after the transformation P′ can theoretically assume any value between plus and minus infinity (Hill and Lewicki 2007). By performing the Logit transformation on both sides of the above Logit regression model, we obtain the standard linear regression model:

$$\mathrm{In}(p/(1 - p)) = b_0 + b_1 \times x_1 + b_2 \times x_2 + \ldots + b_k \times x_k + \text{error\_term} \tag{5.12}$$

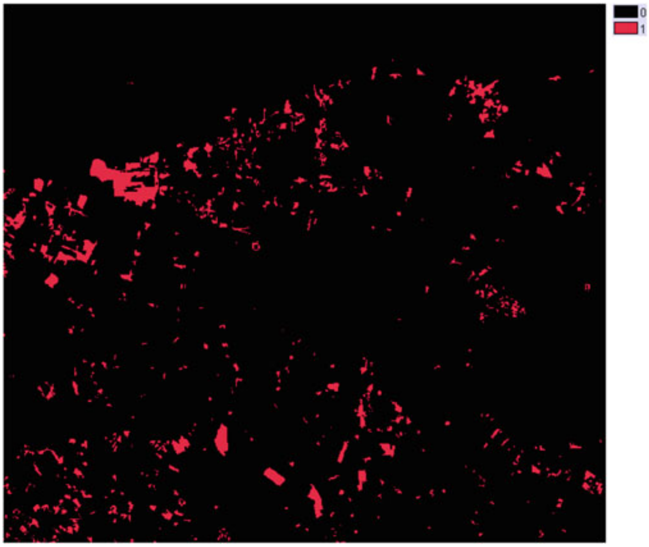


Fig. 5.12 Dependent variable Y; change to built-up area between 1996 and 2006 (no change: $Y = 0$; change $Y = 0$

In fact the Logit transformation of binary data ensures that the dependent variable will be continuous, and the new dependent variable (Logit transformation of the probability) is boundless. Furthermore, it ensures that the probability surface will be continuous within the range from 0 to 1. In general, systematic sampling and random sampling are two approved sampling methods in logistic regression. Systematic sampling reduces spatial dependence. On the other hand, random sampling is capable of representing population, but does not efficiently reduce spatial dependence, especially local spatial dependence (Huang et al. 2009).

5.6.2 Implementation of the Spatially Explicit Logistic Regression Model

In this section of this chapter, it is intended to clarify the assumed independent and dependent variables and the interactions between these variables. Also, a description over model validation and outputs will be presented and simulated maps of future years will be demonstrated. Accordingly, we start with the identification of dependent and independent variables, and then the effective factors upon the dependent variable will be depicted.

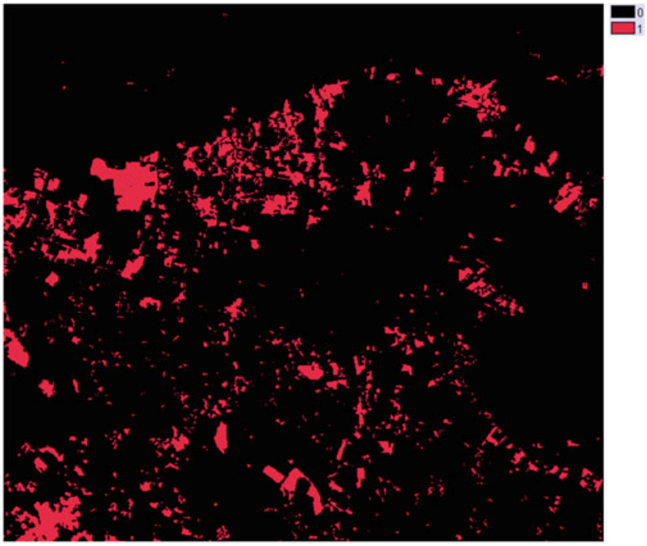


Fig. 5.13 Dependent variable *Y*; change to built-up area between 1986 and 2006 (no change: $Y = 0$; change $Y = 1$)

Table 5.3 ROC and adjusted odd ration values for 18 sets of variables

	ROC	Adjusted odd ratio
Variables set 1	0.8441	20.102
Variables set 2	0.7831	7.6964
Variables set 3	0.844	21.7224
Variables set 4	0.7766	5.2355
Variables set 5	0.6635	3.0513
Variables set 6	0.9223	26.2327
Variables set 7	0.9352	50.3255
Variables set 8	0.9218	26.0128
Variables set 9	0.7167	3.2804
Variables set 10	0.7187	3.4114
Variables set 11	0.8906	16.052
Variables set 12	0.8915	16.5333
Variables set 13	0.8945	15.942
Variables set 14	0.7531	4.7522
Variables set 15	0.8031	11.586
Variables set 16	0.8053	12.0991
Variables set 17	0.8039	11.4385
Variables set 18	0.7392	5.7809

5.6.2.1 Identification of the Dependent Variable

The dependent variable in this implementation is the quantity of change from no-built-up area to built-up area presented as a binary raster lattice where value 1 introduces change on the specific pixels and zero indicates no-change pixels. Figures 5.11, 5.12, and 5.13 represent the structure of the dependent variable files.

A set of independent variables was imported to the Logistic Regression Model in order to become self-calibrated, with the support of IDRISI Andes GIS software (see Table 5.3). A defined mask upon all input data was employed at 30 m resolution to create equal dimension raster files; however, it was an intensive computation for the computer hardware.

5.6.2.2 Predictor Variables (Independent Variables)

In this section, the prior produced land use maps for the years 1986, 1996 and 2006 were employed to specify the change over built-up areas between 1986–1996, 1996–2006 and 1986–2006. Logistic regression modelling executes a data-driven rather than a knowledge-based approach in picking the predictor variables (Hu and Lo 2007). A set of predictor variables was chosen based on preliminary investigations over the case study as well as expert knowledge. A review of effective variables, which was employed in previous similar studies, was a helpful guide. Statistical evaluation, retrieving ROC values and adjusted odd ratios for each set of variables were investigated to pick the optimum set (see Table 5.2). Thus, a calibration process needed to be utilised in order to assure the effectiveness of the assumed variables. These variables and process of data compilation will be explained in the next section.

5.6.2.3 Data Compilation

The social variables correspond to the four affordable elements shaping Tehran's urban patterns (population density, distance to building blocks, single building features, farming lands, categorical demography). Other social variables data were not accessible to be utilised in this approach. Population density is a social variable which determines per capita population per area unit and is expressed as persons per hectare. The econometric and biophysical variables correspond to the eleven affordable elements shaping Tehran's metropolitan patterns (distance to CBD; distance to nearby cities; distance to road networks and interchange; open land features; easting and northing coordinate; digital elevation model; park features; distance to stream; and slope) (Hu and Lo 2007). A set of independent variables (X_1–X_{17}) was imported to a logistic regression model, supported by IDRISI Andes software. An input dataset was designed at 30 m resolution due to compatibility with other available data. Although, a set of other input data, such as distance to education and administration areas, and distance to factories had been evaluated as input to the model, because of weak results, this input data were rejected; hence, these seventeen datasets were imported into the model (see Table 5.4).

Table 5.4 Dependent and independent variables in the logistic regression approach

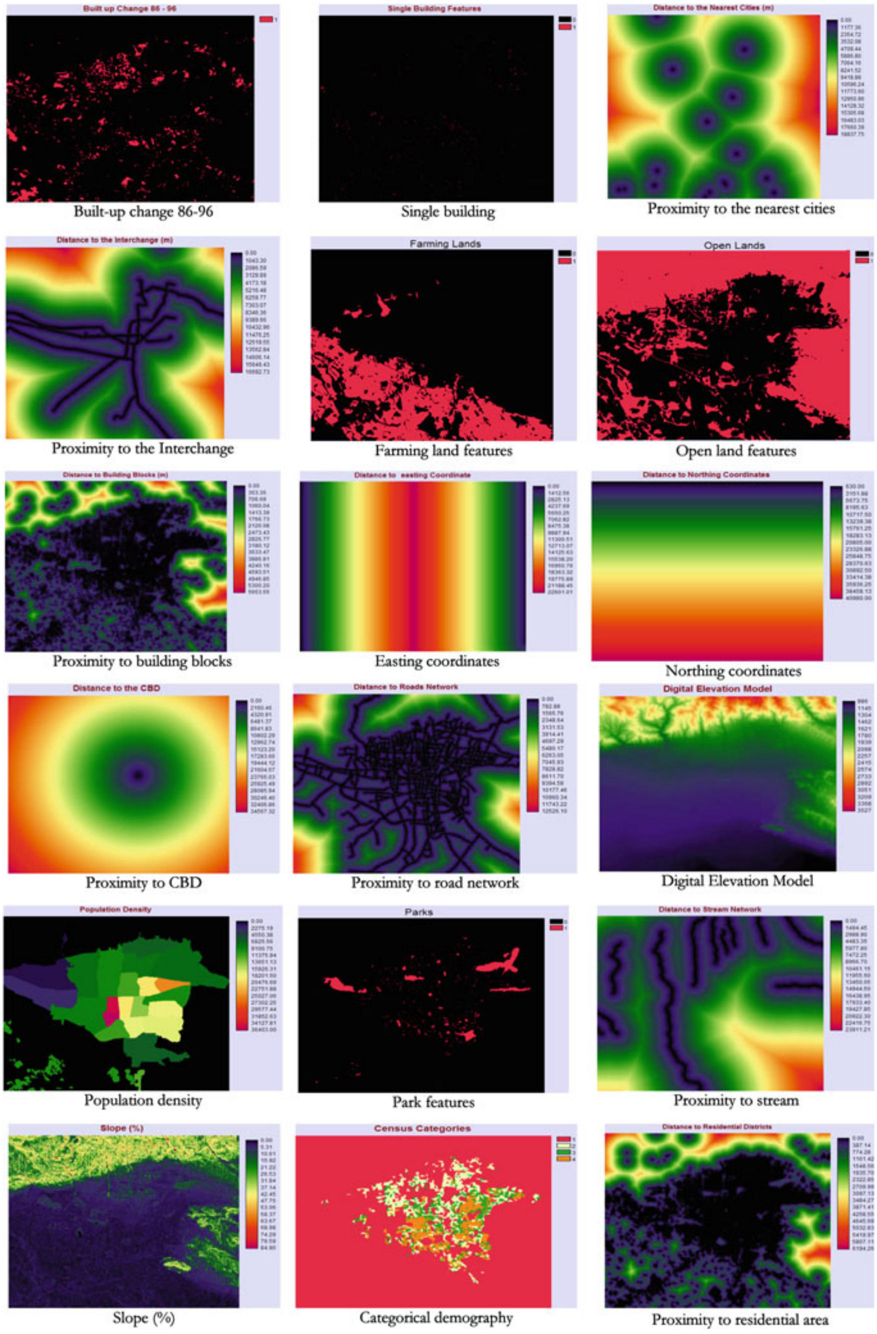
	Variable	Denotation	Structure of variable
Dependent	Y	0—No change to built-up 1—Change to built-up	Dichotomous
Independent	X_1	1—Single building features 0—Non single building features	Binary
	X_2	Proximity to nearby cities (m)	Continuous
	X_3	Proximity to interchange (m)	Continuous
	X_4	1—Farming land features 0—Non farming land features	Binary
	X_5	1—Open land features 0—Non open land features	Binary
	X_6	Proximity to building blocks (m)	Continuous
	X_7	Easting coordinates (m)	Continuous
	X_8	Northing coordinates (m)	Continuous
	X_9	Proximity to CBD (m)	Continuous
	X_{10}	Proximity to road network (m)	Continuous
	X_{11}	Digital Elevation Model (m)	Continuous
	X_{12}	Population density (person/ha)	Continuous
	X_{13}	Park features	Binary
	X_{14}	Proximity to stream (m)	Continuous
	X_{15}	Slope (%)	Continuous
	X_{16}	Categorical demography	Categorical
	X_{17}	Proximity to residential districts	Continuous

Spatial correlation may exist between each category of variables so that logistic regression is able to drop the correlated variables according to the statistical calibration. This calibration basically checks for multi co-linearity. Model calibration in this study was done in two steps, including initial calibration and refining, respectively. All required data were converted to raster format at 30 m resolution.

5.6.3 Calibration of the Logistic Regression Model

The optimum set of variables was picked based on Table 5.2. Each set of variables had different ROC and adjusted odd ratio, which verified the validity of the model, and the approach was carried out numerous times. In order to select the optimum set of variables, it had to reach the highest ROC value. In fact, ROC = 1 indicates a perfect fit and ROC = 0.5 indicates a random fit. A higher adjusted odds ratio is expected for a better fit and higher validity. Therefore, the optimum set of variables is demonstrated in Fig. 5.14.

The logistic regression module was implemented 18 times for 18 sets of variables in order to reach the highest possible ROC and adjusted odd ratio values. The highest value of 0.9532 was obtained, which verifies the accuracy of this model.


Built up Change 86 - 96
Single Building Features
Distance to the Nearest Cities (m)
Distance to the Interchange (m)
Farming Lands
Open Lands
Distance to Building Blocks (m)
Distance to easting Coordinate
Distance to Northing Coordinates
Distance to the CBD
Distance to Roads Network
Digital Elevation Model
Population Density
Parks
Distance to Stream Network
Slope (%)
Census Categories
Distance to Residential Districts

Built-up change 86-96

Single building

Proximity to the nearest cities

Proximity to the Interchange

Farming land features

Open land features

Proximity to building blocks

Easting coordinates

Northing coordinates

Proximity to CBD

Proximity to road network

Digital Elevation Model

Population density

Park features

Proximity to stream

Slope (%)

Categorical demography

Proximity to residential area

Fig. 5.14 Raster layers of independent variables represented in binary and continuous values

Furthermore, the optimum set of variables was incorporated in the model refining phase in order to correct any spatial autocorrelation that might exist. Thus, the selected combination had the minimum spatial autocorrelation. In Table 5.4, a descriptive table of appropriate variables, as well as their structure, is shown. The dependent variable (i.e. built-up change) and independent variables (X_1–X_{17}) are separated by assigned units in the mentioned table.

The employed data and the input maps are shown in Fig. 5.14. These maps are the ultimate variables which have been discussed previously.

The model produces an equation that shows the rate of effectiveness of each particular variable. This equation is presented in the following Eq. 5.13.

$$\begin{aligned}\text{Logit (Urban growth 86} - \text{96)} = & -23.1033\ \text{(intercept)} \\ & + 0.000165 \times \text{Proximity to CBD} \\ & + 0.597356 \times \text{Categorical demography} \\ & - 0.00001 \times \text{Proximity to nearby cities} \\ & - 0.000072 \times \text{Northing coordinates} \\ & + 0.000236 \times \text{Population density} \\ & - 7.428991 \times \text{Proximity to residential area} \\ & + 1.367012 \times \text{Proximity to single buildings} \\ & - 0.000061 \times \text{Easting coordinates} \\ & + 19.776172 \times \text{Farming lands} \\ & - 0.003773 \times \text{Proximity to building blocks} \\ & - 0.001391 \times \text{DEM} \\ & - 0.000044 \times \text{Proximity to interchange} \\ & + 20.618511 \times \text{Open lands} \\ & + 18.393214 \times \text{Proximity to parks} \\ & + 0.000026 \times \text{Proximityto roads} \\ & - 0.047149 \times \text{Slope} \\ & - 0.000013 \times \text{Proximity to streams}\end{aligned} \tag{5.13}$$

According to Eq. 5.13, some variables which have positive values are more favourable for development (e.g. proximity to the CBD, categorical demography, population density, proximity to single buildings, farming lands, open lands, proximity to parks, and proximity to roads). Where variables return negative values the attraction for development falls significantly (e.g. proximity to nearby cities, proximity to streams, northing coordinates, easting coordinates, proximity to residential area, proximity to building blocks, elevation, slope, and proximity to interchange). In other words, those pixels which are closer to the CBD area have more probability of development, and those cells which are in steep slopes have less probability of change. Importantly, the coefficients explain the intensity of

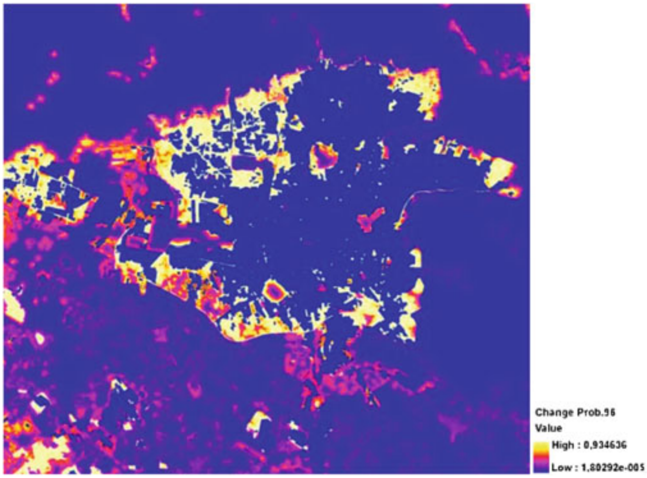


Fig. 5.15 Transition surface maps of study area for 1996

influence in the occurrence of development, for example, proximity to parks is a significant factor in such development.

The output product of the logistic regression model is a probability surface of dependent variable occurrence, which is in this approach urban development (see Figs. 5.15 and 5.16). The probability surface shows that each single cell will be developed with a particular amount of probability. However, this approach is not able to specify the amount and location of change, but can be integrated with other techniques to quantify and allocate the quantity of change. Hence, this probability map will be integrated with the Markov chain model to quantify the extent of the changes. Thereafter, the obtained quantity of change will be allocated in the entire map. The allocation process starts from the maximum value of probability working downward. This process will be explained in Chap. 7.

5.6.4 Validation of the Logistic Regression Model

By means of the prepared probability surface, the quantity of change can be specified through possible techniques, either the Markov chain model or by population growth estimation. The Markov chain model has already been explained in detail. The second method is to employ a footprint of inhabitants to reach the quantity of change (see Sect. 4.10). In this approach, the amount of change was determined based on the transition matrix of the Markov chain model to quantify the changes. The obtained amount was input to the allocation phase. A code was written in Python to subtract the existing built-up areas before beginning the allocation of change from the highest probable cell to the lowest probable cell.

Hence, after executing the designed logistic regression approach, a predicted transition probability surface map, and a residual map indicating the difference

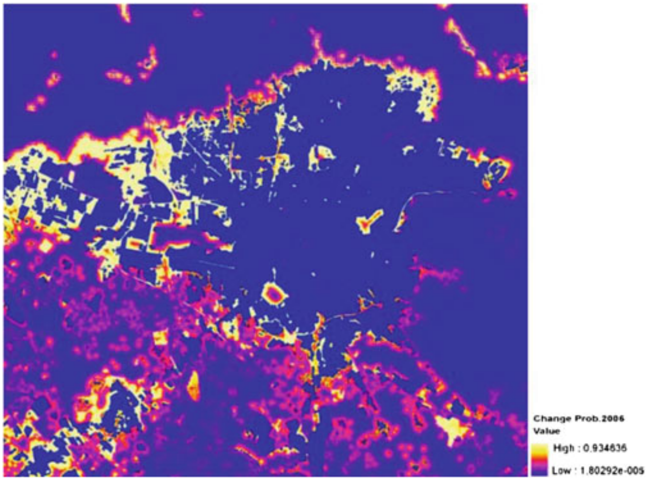


Fig. 5.16 Transition surface maps of study area for 2006

between the predicted and the observed probability, were achieved. Therefore, a transition surface map was produced for 2006 onward. The mentioned prediction surface maps are shown in Figs. 5.15 and 5.16, which can be used for change specification for upcoming periods (2016, 2026). This task was carried out and is demonstrated in Chap. 7.

5.6.5 Land Change Prediction

After the process of model validation was undertaken and the qualification of this model was ensured, land use maps were predicted for 2016 and 2026. Logistic regression requires updated data for the specific times to establish more accurate prediction. In other words, the actual road network map for 2016 is required for the creation of the probability surface at this juncture in time. Therefore, a multi temporal data set of the study area was gathered.

5.7 Summary

Several traditional techniques were demonstrated within this chapter (e.g. CA, Markov chain model, CA-Markov model, logistic regression model). Each model was firstly evaluated and validated and then, once assured of its performance, a land change map was predicted for two future time steps (i.e. 2016, 2026). Each model had some advantages and disadvantages which were investigated, and will be discussed in Chap. 7. The intention was to gather these results in order to integrate them into the assumed ABM. In the next chapter, we start designing the ABM model based on results emanating from traditional methodologies as well as ABM characteristics.

References

Batty M, Xie Y (1994) From cells to cities. Env Plann B: Plann Des 21(7):31–48

Cabral P, Zamyatin A (2009) Markov processes in modeling land use and land cover changes in Sintra-Cascais, Portugal. Dyna 76(158):191–198

Clark JD (1990) Modeling and simulating complex spatial dynamic systems: a framework for application in environmental analysis. SIGSIM Simul Dig 21(2):9–19

Couclelis H (1985) Cellular worlds: a framework for modeling micro–macro dynamics. Env Plann A 17(5):585–596

Goodchild MF, Steyaert LT, Parks BO, Johnston C, Maidment D, Crane M, Glendinning S (1996) GIS and environmental modeling: progress and research issues. Wiley, New York

Hill T, Lewicki P (2007) Statistics methods and applications. StatSoft. Tulsa, OK. http://www.statsoft.com/textbook/neural-networks/#linear

Hu Z, Lo C (2007) Modeling urban growth in Atlanta using logistic regression. Comput Env Urban Syst 31(6):667–688

Huang CY, Sun CT, Hsieh JL, Lin H (2004) Simulating SARS: small-world epidemiological modeling and public health policy assessments. J Artif Societies Soc Simul 7(4)

Huang B, Zhang L, Wu B (2009) Spatiotemporal analysis of rural-urban land conversion. Int J Geog Inf Sci 23(3):379–398

Kamusoko C, Aniya M, Adi B, Manjoro M (2009) Rural sustainability under threat in Zimbabwe—simulation of future land use/cover changes in the Bindura district based on the Markov—cellular automata model. Appl Geogr 29(3):435–447

Li X (2008) Simulating urban dynamics using cellular automata. In: Liu L, Eck J (eds) Artificial crime analysis systems: using computer simulations and geographic information systems, pp 125–139

Li W, Li B, Shi Y (1999) Markov-chain simulation of soil textural profiles. Geoderma 92(1–2): 37–53

Liu Y (2008) Modelling urban development with geographical information systems and cellular automata, 1st edn. CRC Press, Boca Raton, FL

Maguire D, Batty M, Goodchild M (2005) GIS, spatial analysis and Modeling. Esri Press, Red lands, CA

Moreno N, Wang F, Marceau DJ (2009) Implementation of a dynamic neighborhood in a land-use vector-based cellular automata model. Comput Env Urban Syst 33(1):44–54

Pontius RG Jr, Malanson J (2005) Comparison of the structure and accuracy of two land change models. Int J Geog Inf Sci 19(2):243–265

Salman Mahiny AS, Turner BJ (2003) Modeling past vegetation change through remote sensing and GIS: a comparison of neural networks and logistic regression methods. In: Proceedings of the 7th international conference on geocomputation, University of Southampton, United Kingdom

Verhagen P (2007) Case studies in archaeological predictive modeling. Leiden University Press, Leiden

Wegener M (2001) New spatial planning models. Int J Appl Earth Obs Geoinf 3(3):224–237

White R, Engelen G (1993) Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. Env Plann A 25(8):1175–1199

Xie C, Huang B, Claramunt C, Chandramouli C (2005) Spatial logistic regression and gis to model rural-urban land conversion. In processus second international colloquium on the behavioural foundations of integrated land-use and transportation models: frameworks, models and applications. University of Toronto, Canada

Yang Q, Li X, Shi X (2008) Cellular automata for simulating land use changes based on support vector machines. Comput Geosci 34(6):592–602