Contents lists available at SciVerse ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Integration of logistic regression, Markov chain and cellular automata models to simulate urban expansion

Jamal Jokar Arsanjani[a,*], Marco Helbich[b], Wolfgang Kainz[a], Ali Darvishi Boloorani[c]

[a] Department of Geography and Regional Research, University of Vienna, Vienna, Austria
[b] Department of Geography, University of Heidelberg, Heidelberg, Germany
[c] Department of Geography and Cartography, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 18 April 2011
Accepted 22 December 2011

Keywords:
Land use change
Logistic regression
Markov chain
Cellular automata
Tehran

ABSTRACT

This research analyses the suburban expansion in the metropolitan area of Tehran, Iran. A hybrid model consisting of logistic regression model, Markov chain (MC), and cellular automata (CA) was designed to improve the performance of the standard logistic regression model. Environmental and socio-economic variables dealing with urban sprawl were operationalised to create a probability surface of spatiotemporal states of built-up land use for the years 2006, 2016, and 2026. For validation, the model was evaluated by means of relative operating characteristic values for different sets of variables. The approach was calibrated for 2006 by cross comparing of actual and simulated land use maps. The achieved outcomes represent a match of 89% between simulated and actual maps of 2006, which was satisfactory to approve the calibration process. Thereafter, the calibrated hybrid approach was implemented for forthcoming years. Finally, future land use maps for 2016 and 2026 were predicted by means of this hybrid approach. The simulated maps illustrate a new wave of suburban development in the vicinity of Tehran at the western border of the metropolis during the next decades.

1. Introduction

Over recent decades the suburbs of the metropolitan areas are shaped by urban sprawl (e.g., Glaeser and Kahn, 2004; Helbich and Leitner, 2010). Gillham (2002) characterises sprawl, as scattered commercial strip development with low density, and large expanses of single-use development with poor accessibility, as well as a lack of public open space. Beside an increasing induced traffic volume caused by this urban structure, the consumption of natural resources is advocated; biodiversity is lost, among other negative impacts. To counteract such development tendencies and to ensure sustainability, decision makers and urban planners need precise information on urban growth boundaries (Jiang and Yao, 2010). Therefore, land use/cover change (LUCC) analysis has already received considerable attention. Because the speed of development in the emerging countries is higher than those in the developed world (WTO, 1996), the structural changes facing these countries have a greater impact on the urban fabric. Nevertheless, empirical research pertaining to such developing metropolitan areas is quite few (e.g., Jokar Arsanjani, 2011).

Previous studies (e.g., Clarke et al., 1997; Dietzel and Clarke, 2006; Hu and Lo, 2007; Poelmans and Van Rompaey, 2009; Huang et al., 2010; Dubovyk et al., 2011) clearly emphasise the importance of spatiotemporal analysis of urban expansion and why the changes in landscape have recently received more attention, especially because of a number of major reasons: firstly, decision makers and urban planners need precise and detailed information on urban growth and landscape conversion, in order to assess the amount of development, its location, characteristics, and consequences of prior and subsequent urban development (Jiang and Yao, 2010). Secondly, urban landscape structure has been a vital subject for urban investigators in order to develop and modify theories of urban morphology (Batty and Longley, 1994; Longley, 2002). Thirdly, by urban development and landscape change it can be taken into consideration that the boundary between urban areas and environment is important for some environmental models; for instance, urban climate models (Lo et al., 1997; Jiang and Yao, 2010). Finally, urbanisation is one of the primary consequences of globalisation, especially when considering that more than half of the world's population is already settled in urban areas (Kaplan et al., 2008; Jiang and Yao, 2010). Hence, the research of urban growth and landscape change has been incorporated into a variety of programmes within the framework of global change research (Auch et al., 2004).

Whereas urban patterns have been a vital subject in academia to develop and modify theories of urban morphology (Batty and

* Corresponding author.
E-mail addresses: jamaljokar@gmail.com, jokarj3@univie.ac.at (J. Jokar Arsanjani).

0303-2434/$ – see front matter. Crown Copyright © 2011 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.jag.2011.12.014

Longley, 1994; Longley, 2002; Jiang and Yao, 2010), there is still a lack of incorporation between most dynamic simulation models and socioeconomic and demographic variables. Several statistical and geospatial models have been developed and implemented in different case studies to predict urban expansion at different scales, such as regression models (e.g., Hu and Lo, 2007), CA (e.g., Torrens, 2006), MC (e.g., Mousivand et al., 2007), CA-Markov models (e.g., Jokar Arsanjani et al., 2011; Mitsova et al., 2011), and machine learning algorithms (Huang et al., 2010) among others.

Logistic regression (McCullagh and Nelder, 1989) analysis has been one of the most frequently utilised approaches during the past two decades for predictive land use modelling by means of variation of inductive modelling (Verhagen, 2007). Thereby, it is crucial to consider spatial effects, namely spatial autocorrelation and spatial heterogeneity, to challenge regression assumptions (Anselin, 1988; Fotheringham et al., 2000). However, the logistic regression model suffers from the quantification of change and temporal analysis (Hu and Lo, 2007). Thus, empirical estimation and dynamic simulation models have been used to simulate land use change. Various types of rule-based modelling, for instance CA, are most appropriate for incorporating spatial interaction effects and the treatment of temporal dynamics (Hu and Lo, 2007). Whereas CA models focus on the simulation of spatial patterning rather than on the interpretation of spatiotemporal processes of urban sprawl, there is a deficiency of incorporation among dynamic simulation models and socio-economic and demographic variables (Hu and Lo, 2007). Due to limitations of each individual modelling technique, Poelmans and Van Rompaey (2009) proposed a hybrid approach based on logistic regression coupled with CA transition rules, which results in an improved model quality, nevertheless, their model was not able to quantify the amount of land use change.

Thus, the major and innovative objective of this paper is to integrate CA, logistic regression, and MC models in order to produce temporal outputs from the logistic regression model. For this purpose the case study of Tehran, Iran, is investigated. To the best of our knowledge, this is the first time such comprehensive modelling approach is conducted. Through a temporal mapping of land change between 1986, 1996 and 2006 future urban growth in Tehran is simulated.

The paper is structured as follows. Section 2 presents the study site in Tehran, the database, and the data preparation process. The third section briefly introduces the methodology of logistic regression modelling to determine essential driving forces of sprawl, CA, and the MC model. Section 4 discusses the outcomes of the implemented approach, and finally the paper concludes with a brief summary and some suggestions for future works.

2. Materials

2.1. Study area

The study area is the metropolitan area of Tehran (Iran), the largest city in the Middle East (Fig. 1). The official recorded population in 2006 of Tehran city was approximately 8.5 million, and when taking into account encroaching cities (e.g., Shahriar, Islamshahr, Shahre-Qods, Shahre-Andisheh) in the metropolitan area, that number exceeded 13.5 million (Census Information, 2006). Tehran metropolis is the most heavily populated and largest city in Iran, marked by significant differences in class and lifestyle of its citizens (Shahshahani, 2003). For the purpose of this study, an area covering Tehran city and some surrounding cities, covering approximately 1900 km^2, was selected. The geodatabase consists of a variety of environmental and socio-economic datasets; land use maps of the study area were gathered through national data providers. Furthermore, a temporal coverage of Landsat TM and ETM$^+$ images (USGS Global Visualization Viewer) from 1986 to 2006 was collected.

Tehran's elevation varies by 800 m from south to north and has a significant impact on the settlement development, covering an area of around 780 km^2. The metropolis and its surroundings benefits from a massive network of highways (around 285 km) and intersections, over-ramps, and flyovers (about 180 km). However, since 2007, around 130 km of highways and 120 km of over-ramps and intersections have been under construction, which will enhance future suburbanisation processes (Jokar Arsanjani, 2011).

2.2. Datasets

According to a review of common factors involved in land use change modelling summarised in Poelmans and Van Rompaey (2009) and Dubovyk et al. (2011), two categories of driving forces are expected to explain land use change, namely (a) environmental and (b) socio-economic factors. The utilised geospatial and attribute data in this investigation are listed in Table 1 including the sources and production dates.

3. Methods

This section discusses the essential characteristics of the utilised models, which are integrated in this approach. An overview is given in Fig. 2. First, land use maps of 1986, 1996, and 2006 were produced by the processing of Landsat images of the aforementioned years; additionally, temporal land change mapping was implemented. Second, the main driving forces determining land use change, using logistic regression, were investigated (Section 3.1). The resulting probability surface of future land change was used in the third step to estimate the quantity of change based on the MC model (Section 3.2). Fourthly, whereas the MC model is not able to allocate the estimated amount of change and has to be integrated with other geospatial models, a customised CA model was designed in order to achieve the desired objective (Section 3.3). In order to verify the results, the land use map of 2006 was estimated and compared against actual land use maps in the fifth step. Finally, the model was used to simulate future land use maps of 2016 and 2026. In effect, the shortcomings of each particular model (i.e., CA, MC, logistic regression) are eliminated by this implemented approach, which will be discussed further in Section 5.

3.1. Logistic regression

Regression is a method to discover the empirical relationships between a binary dependent and several independent categorical and continuous variables (McCullagh and Nelder, 1989). There are two basic approaches to assess spatial dependency within a regression framework: firstly, building a more complex model incorporating, e.g., an autoregressive structure (Anselin, 1988) and, secondly, designing a spatial sampling plot to enlarge the distance interval between sampled points. Spatial sampling leads to 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. In general, systematic sampling and stratified random sampling are two approved sampling methods in logistic regression. Systematic sampling reduces spatial dependency, whereas random sampling is capable of representing population, but does not efficiently reduce spatial dependency, local spatial dependency in particular (Huang et al., 2009). Nonetheless, it is a reasonable approach to eliminate spatial autocorrelation, and a reasonable design of a spatial sampling scheme will provide an ideal balance between the two sides (Xie et al., 2005). Hence, the stratified random sampling technique was chosen.

The predicted dependent variable in a logistic regression model is a function of the probability that a particular theme will be in one

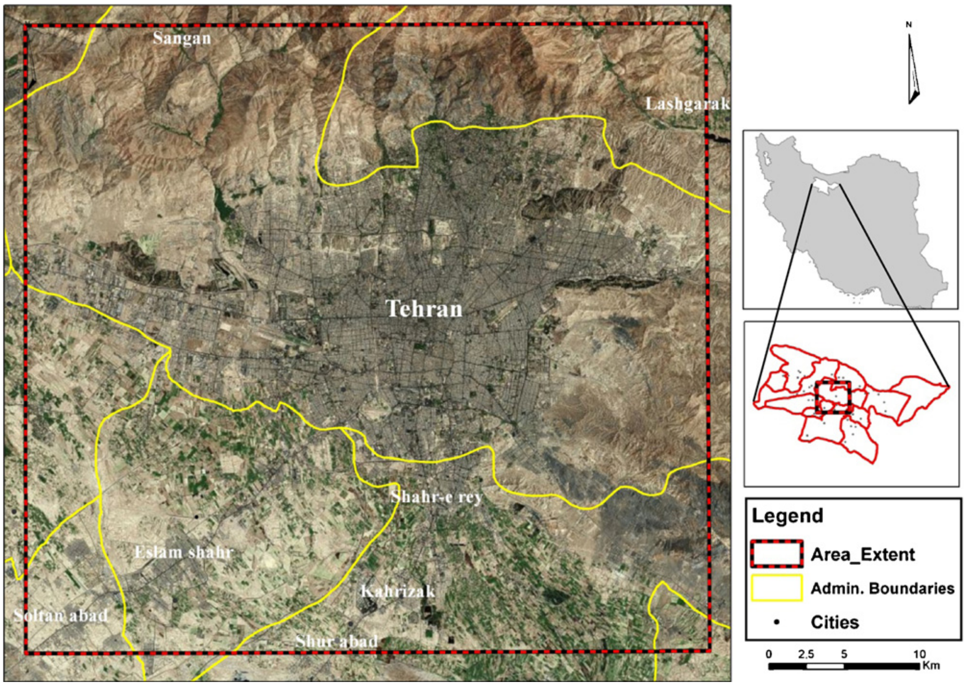


Fig. 1. Study area of the metropolitan area of Tehran.

Source: Census Information, 2006 and Google Earth.

of the categories; for instance, the probability of change of a specific land use class, based on a set of scores on the predictor variables, such as proximity to interchange network, etc. (Huang et al., 2009). The basic assumption is that the probability of a dependent variable takes the value of 1 (positive response), follows the logistic curve, and its value can be calculated with the following formula (Mahiny and Turner, 2003):

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

where, P is the probability of the dependent variable; X represents the independent variables, $X=(x_0, x_1, x_{02}, \ldots, x_k)$, $x_0=1$; B represents the estimated parameters, $B=(b_0, b_1, b_2, \ldots, b_k)$

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

$$P'=\ln\left(\frac{P}{(1-P)}\right) \tag{2}$$

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

$$\ln\left(\frac{P}{(1-P)}\right)=b_0+b_1x_1+b_2x_2+\cdots+b_kx_k+\varepsilon \tag{3}$$

3.2. Markov chain model

The MC model is a stochastic process model that describes how likely one state is to change to another state. It has a key-descriptive tool, which is the transition probability matrix (Mousivand et al., 2007). The MC model is defined as a set of states where a process begins in one of the states and moves consecutively from one state to another; each move is defined as a step (Zhang et al., 2010). In the MC model, two distinct land use maps at different time points should exist, and then it is possible to calculate the probabilities of transition between these time steps. In fact, the simplest Markov model pays no attention to the influence of neighbour cells and only

Table 1
A sketch of the utilised geospatial and attribute data.

Data	Source of data	Date	Resolution
Single building features	Tehran GIS Center	1986, 1996, 2006	30 m
Nearby cities	Tehran GIS Center	1986, 1996, 2006	30 m
Interchange network	Tehran GIS Center	1986, 1996, 2006	30 m
Building blocks	Tehran GIS Center	1986, 1996, 2006	30 m
Roads network	Tehran GIS Center	1986, 1996, 2006	30 m
Digital Elevation Model	U.S. Geological Survey	2006	30 m
Park features	Tehran GIS Center	1986, 1996, 2006	30 m
River streams	Tehran GIS Center	1986, 1996, 2006	30 m
Census	Tehran Statistics Center	1986, 1996, 2006	30 m
Residential Districts	Tehran GIS Center	1986, 1996, 2006	30 m
Land use maps of the study area	Tehran GIS Center	1986, 1996, 2006	30 m
Landsat images of the study area	U.S. Geological Survey(USGS)	1986, 1996, 2006	30 m

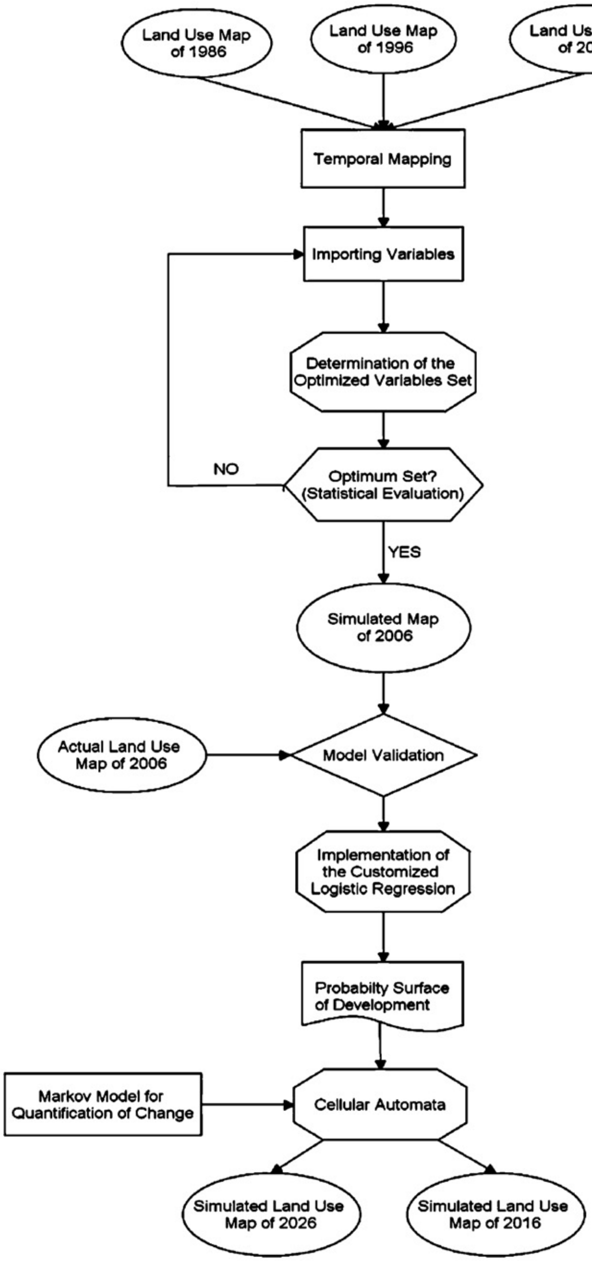

Land Use Map of 1986
Land Use Map of 1996
Land Use Map of 2006
Temporal Mapping
Importing Variables
Determination of the Optimized Variables Set
NO
Optimum Set? (Statistical Evaluation)
YES
Simulated Map of 2006
Actual Land Use Map of 2006
Model Validation
Implementation of the Customized Logistic Regression
Probabilty Surface of Development
Markov Model for Quantification of Change
Cellular Automata
Simulated Land Use Map of 2026
Simulated Land Use Map of 2016

Fig. 2. Flowchart of the Logistic–Markov–CA approach.

considers cell states at t_1 and t_2 (Eastman, 2006). The MC model has been rarely used to study urban sprawl and land use change (e.g., Lopez et al., 2001; Mousivand et al., 2007; Jokar Arsanjani et al., 2011); however, the MC model will be integrated in this approach in order to examine its performance.

3.3. Cellular automata

Proximity is one of the essential geospatial elements that emphasise the dynamics of various change events. Areas expose a higher tendency to change to a class when they are located near existing areas of the same class (i.e., sprawl phenomena). These events can be efficiently simulated by means of CA models. In effect, in cellular automata, there are cellular entities that independently vary their states, as well as their immediate neighbours, according to predefined transition rules. Various ways of defining transition rules make CA models function differently and, consequently, produce dissimilar outputs (Eastman, 2006; Lambin and Geist, 2006).

According to Wolfram (1984, page 419–424), "*cellular automata are constructed from many identical components, each simple, but together capable of complex behaviour".* Cellular automata model is a well-known simulation model where space and time are discrete and interactions are only local. In fact, in CA, space is separated into regular cells and the state of each particular cell is determined by the state of the cell itself, as well as its surrounding cells, at a previous time period through a set of predefined transition rules. However, the states of each cell can be updated synchronously. The overall performance of the system will be concluded from the combined actions of all the locally defined transition rules, therefore, the state of the system moves forward in discrete time steps. In general, this model is also known as "*What If, Then*" experiment technique (Liu, 2008). CA models have been widely implemented in the simulation of urban systems, such as sprawl (e.g., Wu, 1996; Torrens, 2006).

4. Results

4.1. Temporal urban sprawl mapping

For temporal land use mapping of the study area a set of Landsat TM and ETM$^+$ images for the period 1986–2006 was chosen to extract land use maps. The availability of rich Landsat archive imagery, with reasonable revisit time, can provide enough datasets to cover the study area temporally. The temporal mapping needed regular time steps between images. To this end, images of 1986, 1996, and 2006, constituting regular 10-year cycles were chosen to be synchronised with the environmental and socioeconomic data. Accuracy assessment by means of cross-tabulation analysis of the map classifications was conducted in order to ensure the accuracy of the maps at 91, 88, and 90%, respectively. Five land use categories were retrieved, i.e., agricultural lands, water bodies, public parks, open lands, and built up areas. Fig. 3 illustrates the produced land use maps.

Land use change over the time periods 1986–1996, 1996–2006, and 1986–2006 was considered to quantify the amount and location of change. According to Table 3, most land conversion replaces agricultural lands and open lands with urban areas. Overall, 14,333 ha have been changed to built-up areas in the total time period. Within the period 1986–2006, approximately 5172 ha of agricultural fields, as well as 10,533 ha of open lands have been turned into built-up areas. Around 8281 ha have been changed to built-up areas within 1986–1996, and 6052 ha within the period 1996–2006. Descriptive analysis shows a "big wave" of change in the north-west, west, and south-west parts of the study area between 1986 and 1996. For the period between 1996 and 2006, observably the most dramatic built-up development has occurred in the west part of the study area; however some speckle spots in the entire study site can be observed. In addition, some areas in the main core of the metropolis have also been developed.

4.2. Driving-forces of land use change

4.2.1. Regression model specification

The prior produced land use maps of 1986, 1996, and 2006 were used to specify the change maps over built up areas for the periods 1986–1996, 1996–2006, and 1986–2006, respectively. The following input dataset was designed at 30 m resolution due to compatibility with other accessible data. The dependent variable in this implementation is the developed cells (i.e., change from no built-up area to built-up area) presented as a binary raster where a value 1 indicates change on the specific pixels and zero indicates no change within a time period (e.g., 1986–1996). Fig. 4 represents the pattern of each dependent variable.

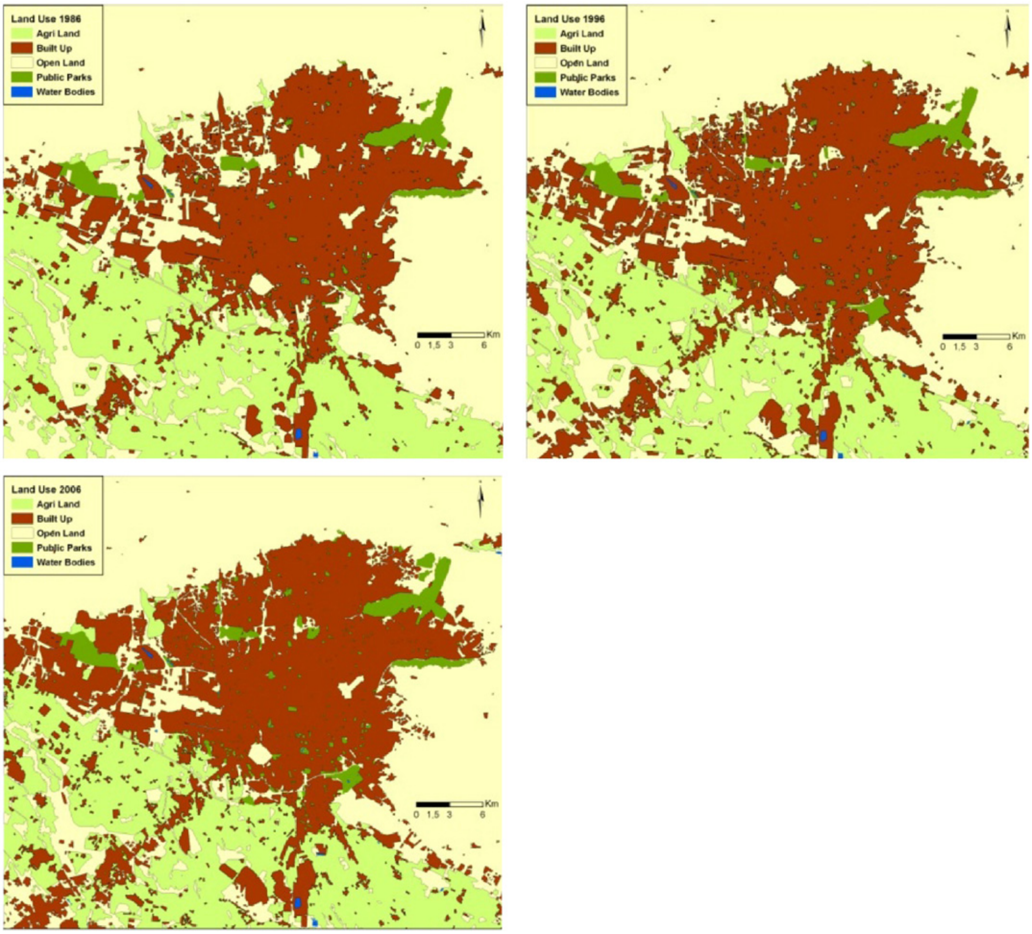


Fig. 3. Extracted land use maps of 1986 (upper left), 1996 (upper right), and 2006 (lower left).

Table 2
Quantity of land use change over time in terms of hectare and percent of each category.

Year	1986		1996		2006		1986–1996	1996–2006	1986–2006
Category	ha	%	ha	%	ha	%	(ha)	(ha)	(ha)
Agricultural Field	45,175	24.2	41,179	22.1	40,003	21.5	−3996	−1176	−5172
Built-up Area	44,783	24.0	53,064	28.5	59,116	31.7	8281	6052	14,333
Open Land	92,302	49.5	87,330	46.8	81,769	43.9	−4972	−5561	−10,533
Public Park	4080	2.2	4746	2.5	5393	2.9	666	647	1313
Water Body	104	0.1	125	0.1	163	0.1	21	38	59

A set of predictor variables was chosen based on preliminary research (e.g., Poelmans and Van Rompaey 2009; Dubovyk et al., 2011) over the study area as well as expert knowledge. It was assumed that some social variables, such as population density, categorical census, single building features, and farming land would correspond to the shape of Tehran's urban patterns, e.g., physical development might take place in non-densely populated areas, close to building blocks and near to single building features,

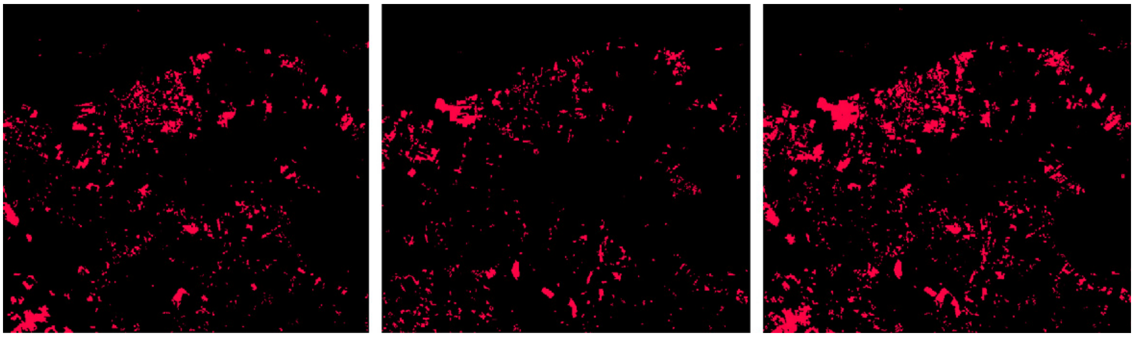


Fig. 4. Dependent variable *Y*; change to built-up area within 1986–1996, 1996–2006, and 1986–2006, respectively (left to right) (no change: black cells; change: red cells). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 3
Relative operating characteristic (ROC) and adjusted odds ratio values for 18 sets of variables.

	ROC	Adjusted odd ratio
Variables Set 1	0.8441	20.102
Variables Set 2	0.7831	76.964
Variables Set 3	0.844	217.224
Variables Set 4	0.7766	52.355
Variables Set 5	0.6635	30.513
Variables Set 6	0.9223	262.327
Variables Set 7	0.9352	503.255
Variables Set 8	0.9218	260.128
Variables Set 9	0.7167	32.804
Variables Set 10	0.7187	34.114
Variables Set 11	0.8906	16.052
Variables Set 12	0.8915	165.333
Variables Set 13	0.8945	15.942
Variables Set 14	0.7531	47.522
Variables Set 15	0.8031	11.586
Variables Set 16	0.8053	120.991
Variables Set 17	0.8039	114.385
Variables Set 18	0.7392	57.809

Table 4
Effective variables in urban growth between 1986 and 1996 and their coefficients.

	Variable	Coefficient
Urban growth between 1986 and 1996	Intercept	−23.1033
	Distance to CBD	0.000165
	Census	0.597356
	Distance to nearby Cities	−0.000001
	Northing Coordinates	−0.000072
	Population Density	0.000236
	Distance to Residential Areas	−7.428991
	Distance to Single Buildings	1.367012
	Easting Coordinates	−0.000061
	Farming	19.776172
	Distance to Building Blocks	−0.003773
	DEM	−0.001391
	Distance to Interchange	−0.000044
	Open lands	20.618511
	Distance to Parks	18.393214
	Distance to Roads	0.000026
	Slope	−0.047149
	Distance to Streams	−0.000013

and farming lands. Thus, the aim is to discover the positive or negative influences of each particular variable. However, other social variables were not accessible to be utilised in this approach. Population density is defined as a social variable, which determines per capita population per area unit and is expressed as persons per hectare. It was also taken into account that some econometric and environmental variables, such as distance to central business district (CBD), distance to the nearby cities, distance to interchange, open land features, distance to road network, altitude, distance to stream, and slope, might alter land use patterns in the study area (see Table 2). For instance, since it was presumed that the exact location influences the amount of change in Tehran, northing and easting coordinates were taken into consideration: e.g., the negative coefficient of the easting and northing coordinates variables verify that land changes are occurring towards west and south. Moreover, the coefficients of these variables also determine the severity of the variables, that is, land conversion occurs more intensively in the south rather than in the west of Tehran.

Logistic regression was estimated for 18 different sets of variable combinations in order to reach the highest possible relative operating characteristic (ROC) and adjusted odd ratio values. Essentially, the odds ratio depicts the degree of correlation or non-dependence between two binary data, which is known as the measure of effect size. It is a significant index of logistic regression and is utilised as a descriptive statistics. Despite other measures of relationship for paired binary data, such as the relative risk, the odds ratio compares the two respective variables symmetrically, and can be calculated using some types of non-random samples (Edwards, 1963; Mosteller, 1968). In fact, probability values vary in the range between 0 and 1, which state the possibility of an event as a proportion of both incidence and non-incidence. Odds depict the likelihood of an occurrence relative to the likelihood of a non-occurrence (Pampel, 2000).

Model calibration was done as a two-step procedure including initial calibration and refining, respectively. The approach was carried out for each set of variables to assess their reliability (Table 3). In order to select the optimum set of variables, it had to reach the highest ROC value, which verifies the validity of the model.

In this study, the ROC method was performed to validate the performance of the approach. The ROC method has been lately introduced to the land use/cover change modellers to compute the correlation between simulated changes and actual changes, which is a reliable technique to assess the validity of an approach (e.g., Pontius and Schneider, 2001). Essentially ROC assesses how well a pair of maps fit in terms of the location of developed pixels (Pontius, 2000; Hu and Lo, 2007). Necessarily 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 (Eastman, 2006). The highest value of 0.953 was obtained, which verifies the accuracy of this model. A descriptive table of appropriate variables, as well as their levels of measurement, are shown.

This set of variables comprises the raster variables as shown in Fig. 5 and Table 4 demonstrates the impact degree of each particular variable in the approach.

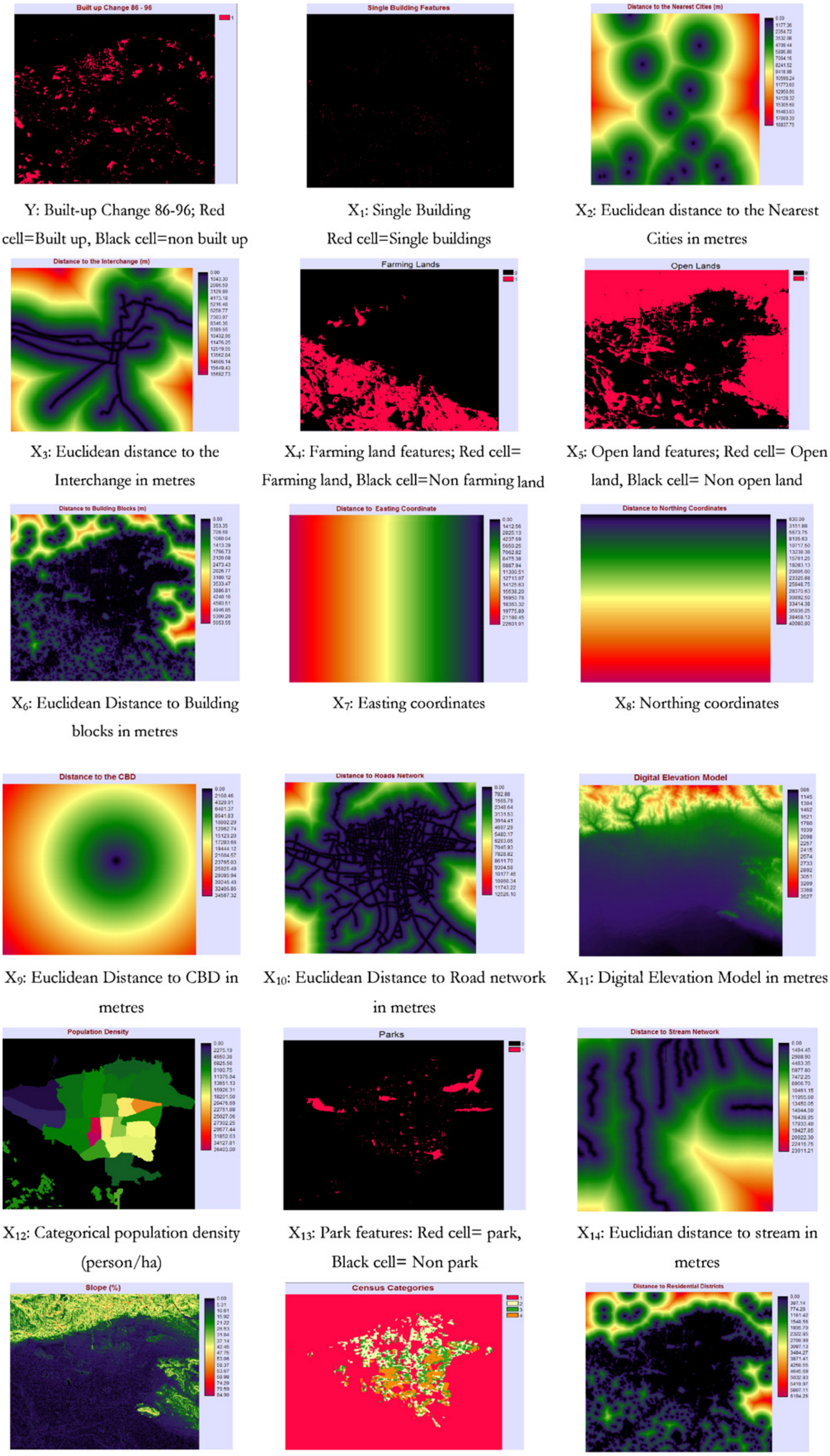
According to Table 4, some variables which have positive values have a positive impact and thus advocate suburban developments (e.g., proximity to the CBD, categorical demography, population density, proximity to single buildings, farming lands, open lands, proximity to parks, and proximity to roads). 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, whereas cells which are in steep slopes have less probability of change. Importantly, the coefficients explain the intensity of influence in the occurrence of development, for example, proximity to parks is a significant factor in such development. Although, a set of other input data, such as distance to educational institutions, administration areas and factories had been evaluated but result in low ROC values (Table 3) and consequently were rejected. Hence, seventeen predictors (Fig. 5) were eventually designed for this run.

The output product of the logistic regression model is a probability surface of dependent variable occurrence, indicating urban development (see Fig. 6). The probability surface shows that each single cell will be developed with a particular amount of probability (1 = high probability, 0 = low 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 change. Hence, this probability map will be integrated with the MC model to quantify the extent of the changes, and CA to allocate the predicted 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.

4.2.2. *Validation of logistic regression model*

By means of the prepared probability surface, the computed quantity of change can be allocated. In this approach, the amount


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

Y: Built-up Change 86-96; Red cell=Built up, Black cell=non built up

X_1: Single Building Red cell=Single buildings

X_2: Euclidean distance to the Nearest Cities in metres

X_3: Euclidean distance to the Interchange in metres

X_4: Farming land features; Red cell= Farming land, Black cell=Non farming land

X_5: Open land features; Red cell= Open land, Black cell= Non open land

X_6: Euclidean Distance to Building blocks in metres

X_7: Easting coordinates

X_8: Northing coordinates

X_9: Euclidean Distance to CBD in metres

X_{10}: Euclidean Distance to Road network in metres

X_{11}: Digital Elevation Model in metres

X_{12}: Categorical population density (person/ha)

X_{13}: Park features: Red cell= park, Black cell= Non park

X_{14}: Euclidian distance to stream in metres

X_{15}: Slope in percent

X_{16}: Categorical Census

X_{17}: Euclidian distance to Residential Area in metres

Fig. 5. Raster layers of independent variables for the optimum model represented as binary and continuous values.

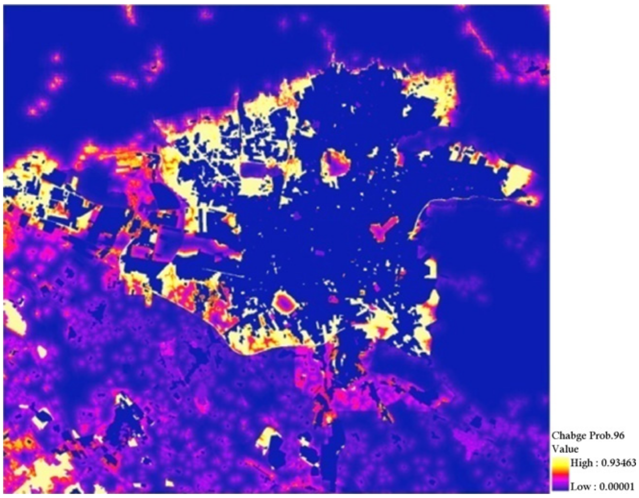


Fig. 6. Transition surface map of study area from 1996 onward.

of change was determined based on the transition matrix of the MC model to quantify the changes, then the obtained amount was input to the allocation phase. A customised cellular automata function was coded in the Python programming engine in order to subtract the existing built-up areas before beginning the allocation of change from the highest probable cell to the lowest probable cell. This function exerts a 3 × 3 size kernel over the study area and substitutes non-built up cells with built up cells in the event that the probability of change within the kernel is high. This function continues the allocation process until it distributes the estimated quantity of change (see Jokar Arsanjani, 2011).

Hence, after executing the designed logistic regression, a predicted transition probability surface map was created for 1996 onward (Fig. 7), which can be used for change specification for subsequent periods (2016, 2026).

4.3. Quantification of future changes

The MC model was run to quantify changes, with a pair of land use images as input and a transition probability matrix, a matrix of transition areas, as well as a set of conditional change probability images as output. The produced results record the probability matrix that shows the probability for each land cover category to change to any other categories (see Table 5). The transition probability matrix was calculated by the contingency matrix displaying the relative frequencies of land change at a certain time period (Cabral and Zamyatin, 2009). The transition areas matrix is a table which records the amount of pixels that are anticipated to change from one particular land use category to other categories according to a number of time units (see Table 6). The results (i.e., matrices) were used for further change analysis and determine the estimated quantity of change that is assumed to be an input for the logistic regression model.

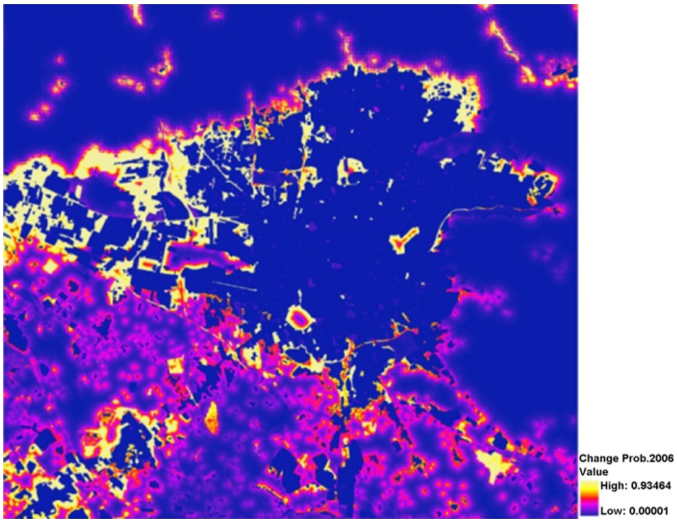


Fig. 7. Transition surface map of development from 2006 onward.

MC is not a spatially explicit model; therefore, it is not an appropriate model to estimate the location of change, which needs to be integrated with other spatial models and in this investigation, logistic regression and CA models were chosen to spatialize the estimated change quantity. Nevertheless, it is a good quantity estimator for its outcomes to be allocated (Kamusoko et al., 2009). As shown in Table 5, the probability of converting each land category to others can be determined by the MC model, i.e., based on the Markov transition probabilities matrix of 1986–1996, the agricultural field category remains constant at 88.35% probability, and will change to built-up category at 4.87% probability, etc. This matrix will be further used to quantify the amount of change for each category. Furthermore, the estimated amount of change from each land type to the other existing land types for 2016 and 2026 has been demonstrated in Table 6.

In order to validate the proposed approach, the probability map of change for 1996 was utilised to allocate the attained quantity of change (Table 5) through the customised cellular automata function. It was aimed to simulate the land use map of 2006 and compare it with the actual map of 2006. Being aware of Pontius and Millones's (2011) critics, kappa statistic for map classification comparison (actual map vs. simulated maps) was applied and revealed a Kappa index of 0.89. According to Landis and Koch (1977), this task verifies and approves the simulation process; accordingly, the approach can also be executed for future years (i.e., 2016, 2026).

4.4. Implementation of the hybrid model and final outcomes

A set of independent variables satisfied the preliminary designed approach after statistical assessment. The predefined independent variables were imported again to the logistic regression. Then the prior produced land use maps for 1986, 1996, and 2006 were used to specify the change maps over built-up areas within the time ranges 1986–1996, 1996–2006, and 1986–2006. The statistical evaluation and achieved ROC values and adjusted odds ratios for each set of combined variables (see Table 3) were suitable tools to identify the best predictor variables.

The highest ROC value determines the optimal set of input variables; moreover, higher adjusted odds ratios verify the validity and fitness of the chosen set. The highest obtained ROC value, at 0.953, allowed us to pick the appropriate set of variables as input files as shown in Table 3. A predicted change probability surface map and a residual map indicating the difference between the predicted and the observed probability were generated and its inspection shows no violations of the model assumptions. Therefore, in order to allocate the proper quantity of change on the probability surface, the transition matrix produced by the MC model was preferred to quantify the amount of change.

Based on the probability surface, as well as the change demand quantity, the CA function was applied on the transition surface map (Fig. 7) to produce the land cover maps of 2016 and 2026 (see Fig. 8). This probability surface enables us to predict upcoming changes in any proper year (i.e., 2016, 2026). In fact, once the model validation process and the qualification of this model were assured, land use maps were predicted for 2016 and 2026. Logistic regression requires updated data for the specific times to be accurate for

Table 5
Markov transition probabilities matrix for the periods 1986–1996, 1996–2006 and 1986–2006.

		Agriculture 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	0.0000	0.0335	0.0232	0.9428	0.0005
	Water Body	0.0105	0.0000	0.0000	0.0000	0.9895
Probability value of 2016 based on transition matrix of 1996–2006	Agriculture Field	0.9361	0.0402	0.0218	0.0017	0.0002
	Built-up	0.0036	0.9873	0.0055	0.0035	0.0000
	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	0.0000	0.0066	0.9621
Probability value of 2026 based on transition matrix of 1986–2006	Agriculture 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	0.0000	0.0434	0.0198	0.9362	0.0005
	Water Body	0.0105	0.0009	0.0000	0.0000	0.9886

Table 6
Calculation of quantity of change (transition area matrix) through Markov chain model for 2016 and 2026 in hectare.

		Agri lands	Built-up	Open lands	Public parks	Water body
2016	Agri lands	37,424	1605	871	69	9
	Built-up	214	58,326	324	206	2
	Open lands	1175	4702	75,355	448	26
	Public Parks	10	47	35	5290	7
	Water Body	4	2	0	1	158
2026	Agri lands	33,857	3744	1972	385	21
	Built-up	17	58,398	347	305	7
	Open lands	1537	9239	70,096	806	28
	Public Parks	0	234	107	5046	3
	Water Body	2	0	0	0	162

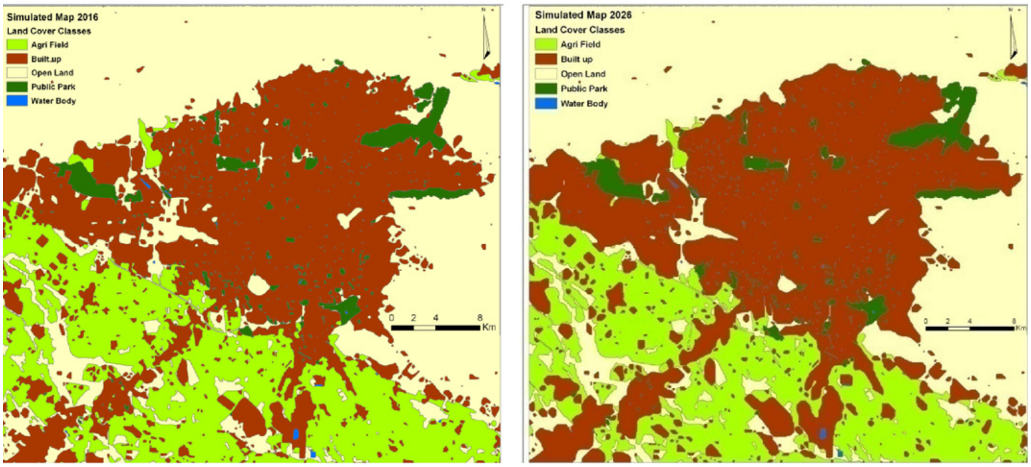


Fig. 8. Simulated land use maps of 2016 (left) and 2026 (right) through the Logistic–Markov–CA approach.

prediction, i.e., the map of the proposed road network for 2016 is needed for an enhanced forecast.

5. Discussion and conclusions

In this investigation, a logistic regression model was combined with the MC and CA models in order to develop an efficient hybrid geospatial explicit approach. The logistic regression model has the advantage of exploring relationships between land conversion and causative factors quantitatively, which enables us to distinguish between the effective variables (Park et al., 2011). However, a simple logistic regression model suffers from several limitations, such as temporal determination of change, change quantification, as well as change allocation (Hu and Lo, 2007). Therefore, the present approach was designed and performed to rectify the aforementioned constraints and to discover the interaction of various environmental and socio-economic variables which cause urban sprawl (Fig. 5).

Various factors, which were assumed to cause urban growth in the study area, were taken into account (see Table 4). Less effective variables were excluded to promote the quality of the model after consideration of possible combinations of variables. The optimum set of variables was chosen according to the computed ROC values (Table 3) as input into the developed approach, to subsequently simulate the urban expansion in the study area, for 2016 and 2026 (Fig. 8).

These three techniques were combined for the following purposes: firstly, the logistic regression model was utilised to create a probability surface and to determine the most probable sites for development; secondly, the MC model was used to retrieve the quantity of change. Because land development policy has been inconsistent in recent years (where it has allowed property developers to multiply high-rise constructions for maximum profit) population growth and land development rates are impossible to synchronise. For example, if a plot of land has been allocated for the purpose of constructing a one-family home, the developer may instead choose to turn the building into a multi occupancy apartment block. Thirdly, the CA model is a significant tool to allocate probable changes under predefined conditional rules. This CA model allocated the amount of change, beginning with the cells of highest probability. Therefore, the approach is capable of predicting the most probable sites for development, estimating the likely amount of change as well as allocating the estimated quantity within the study area. An integration of CA and MC models (CA–Markov model) has been already implemented and carried out on the same study area by Jokar Arsanjani et al. (2011) and its results have verified the validity of this approach.

This paper has attempted to demonstrate that this hybrid technique (i.e., Logistic–Markov–CA) offers certain advantages compared with traditional techniques. Firstly, this approach is capable of considering and integrating environmental and socio-economic factors, which are not considered in current CA models, e.g., SLUETH (Clarke et al., 1997; Yang and Lo, 2002; Dietzel and Clarke, 2006). Secondly, any spatial factor can be imported to this approach in order to measure its influence on urban sprawl and, accordingly, can be rejected after statistical assessment. Finally, the mentioned approach was tested and verified in two steps: (i) while the approach was being developed (i.e., the model calibration process) and (ii) through the comparison of the actual map and the simulated map of 2006, which was generated to verify the outcome of the approach. Whereas the validation of the current LUCC models is still weak (Pontius et al., 2004), it is not feasible to validate the certainty of the simulated maps for the future. Thus, the only possible way to verify the model was to validate it at the most recent time, and following the assurance of the model's performance, future land use maps could be simulated more confidently.

Although logistic regression models suffer from a lack of allocation process, in this investigation the CA module has covered this weakness. Logistic regression uses data at different scales (raster data, census data, etc.), but the model per se ignores this fact, which can result in model bias. Due to the hierarchical structure of the data (cells are nested within municipalities) multi-level modelling (Goldstein, 2010) may have high potential to model the relationship between land use change and its driving forces, by explicitly and systematically taking different spatial scales, spatial autocorrelation and heterogeneity into account. Despite its appealing properties for land use modelling, empirical applications are rare. Notable exceptions promoting its statistical appropriateness are Vance and Iovanna (2006) as well as Overmars and Verburg (2006). This enhanced model specification underpins the need for much additional work in future researches before a complete understanding of urban expansion is achieved.

Yet, despite the strengths of this approach, our investigation has also highlighted the considerable limitations of the approach. Though the method can incorporate various driving forces, it does incorporate certain limitations in parallel models, such as the non factoring of individuals' behaviour, personal preferences and governmental actions in land use conversions which agent-based modelling (ABM) performs. Neither simple logistic regression models nor hybrid logistic regression models consider any individual-related factors. Therefore, Geosimulation (Benenson and Torrens, 2004) and ABM (e.g., Crooks, 2007) take the human-related variables into consideration more effectively and, consequently, dissimilar outcomes will almost certainly occur. It is our recommendation, therefore, that agent-based modelling should be implemented on the same study area in order to compare the current and traditional models and their outcomes against the agent based modelling approach.

Acknowledgement

The authors would like to thank the anonymous reviewers for their valuable and remarkable suggestions to improve the quality of this paper.

References

Auch, R., Taylor, J., Acevedo, W., 2004. Urban Growth in American Cities: Glimpses of U.S. Urbanization. U.S. Geological Survey Circular, vol. 1252.

Anselin, L., 1988. Spatial Econometrics. Methods and Models. Kluwer Academic Publishers, Dordrecht.

Batty, M., Longley, P., 1994. Fractal Cities: A Geometry of Form and Function, 1st ed. Academic Press, San Diego.

Benenson, I., Torrens, P.M., 2004. Geosimulation: Automata-Based Modeling of Urban Phenomena. John Wiley and Sons, London.

Cabral, P., Zamyatin, A., 2009. Markov Processes in Modeling Land Use and Land Cover Changes in Sintra-Cascais, No. 158, vol. 76. DYNA-COLOMBIA, Portugal, pp. 191–198.

Census Information, 2006. Census Information, Tehran: The Statistical Centre of Iran. Available at: http://amar.sci.org.ir/index_e.aspx.

Clarke, K.C., Hoppen, S., Gaydos, L., 1997. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. Environment and Planning B: Planning and Design 24, 247–261.

Crooks, A.T., 2007. Experimenting with Cities: Utilizing Agent-Based Models and GIS to Explore Urban Dynamics. University College London, London.

Dietzel, C., Clarke, K.C., 2006. The effect of disaggregating land use categories in cellular automata during model calibration and forecasting. Computers, Environment and Urban Systems 30, 78–101.

Dubovyk, O., Sliuzas, R., Flacke, J., 2011. Spatio-temporal modelling of informal settlements development in Sancaktepe district, Istanbul, Turkey. ISPRS Journal of Photogrammetry and Remote Sensing 66 (2), 235–246.

Eastman, J.R., 2006. IDRISI 15: The Andes edition. Clark University, Worcester, MA.

Edwards, A.W.F., 1963. The measure of association in a 2 × 2 table. Journal of the Royal Statistical Society, Series A 126 (1), 109–114 (Blackwell Publishing).

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2000. Quantitative Geography: Perspectives on Spatial Data Analysis. Sage Publications, London; Thousand Oaks, California.

Gillham, O., 2002. The Limitless City: A Primer on the Urban Sprawl Debate. Island Press, Washington, D.C., USA.

Glaeser, E.L., Kahn, M.E., 2004. Sprawl and Urban Growth, Discussion Paper at the Institute of Economic Research, Harvard.

Goldstein, H., 2010. Multilevel Event History and Survival Models, in Multilevel Statistical Models, 4th ed. John Wiley & Sons, Ltd, Chichester, UK, doi:10.1002/9780470973394.ch11.

Helbich, M., Leitner, M., 2010. Postsuburban spatial evolution of Vienna's urban fringe: evidence from point process modeling. Urban Geography 31 (8), 1100–1117.

Hill, T., Lewicki, P., 2007. STATISTICS Methods and Applications. StatSoft, Washington, D.C, Available at: http://www.statsoft.com/textbook/neural-networks/linear.

Hu, Z., Lo, C., 2007. Modeling urban growth in Atlanta using logistic regression. Computers, Environment and Urban Systems 31 (6), 667–688.

Huang, B., Zhang, L., Wu, B., 2009. Spatiotemporal analysis of rural–urban land conversion. International Journal of Geographical Information Science 23 (3), 379–398.

Huang, B., Xie, C., Tay, R., 2010. Support vector machines for urban growth modelling. Geoinformatica 14 (1), 83–99.

Jiang, B., Yao, X. (Eds.), 2010. Geospatial Analysis and Modelling of Urban Structure and Dynamics. Springer, Dordrecht, The Netherlands, ISBN 9048185718.

Jokar Arsanjani, J., 2011. Dynamic Land Use/Cover Change Simulation. Springer Publication, ISBN 978-3-642-23704-1.

Jokar Arsanjani, J., Kainz, W., Mousivand, A., 2011. Tracking dynamic land-use change using spatially explicit Markov Chain based on cellular automata: the case of Tehran. International Journal of Image and Data Fusion 2 (4), 329–345, doi:10.1080/19479832.2011.605397.

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. Applied Geography 29 (3), 435–447.

Kaplan, D.H., Wheeler, J.O., Holloway, S.R., 2008. Urban Geography, 1st ed. John Wiley.

Lambin, E.F., Geist, H.J., 2006. Land-Use and Land-Cover Change: Local Processes and Global Impacts. Springer-Verlag, Berlin.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics, vol. 33, pp. 159–174.

Liu, Y., 2008. Modelling Urban Development with Geographical Information Systems and Cellular Automata. CRC Press (Taylor & Francis Group), New York.

Lo, C.P., Quattrochi, D.A., Luvall, J.C., 1997. Application of high-resolution thermal infrared remote sensing and GIS – to assess the urban heat island effect. International Journal of Remote Sensing 18 (2), 287–304.

Longley, P.A., 2002. Geographical Information Systems: will developments in urban remote sensing and GIS lead to 'better' urban geography? Progress in Human Geography 26 (2), 231–239.

Lopez, E., Bocco, G., Mendoza, M., Duhau, E., 2001. Predicting land-cover and land-use change in the urban fringe: a case in Morelia city, Mexico. Landscape and Urban Planning 55 (4), 271–285.

Mahiny, A.S., Turner, B.J., 2003. Modeling past vegetation change through remote sensing and G.I.S: a comparison of neural networks and logistic regression methods. In: Proceedings of the 7th International Conference on GeoComputation. The 7th International Conference on GeoComputation, University of Southampton, United Kingdom.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models. CRC Press, Boca Raton.

Mitsova, D., Shuster, W., Wang, X., 2011. A cellular automata model of land cover change to integrate urban growth with open space conservation. Landscape and Urban Planning 99, 141–153.

Mosteller, F., 1968. Association and estimation in contingency tables. Journal of the American Statistical Association 63 (321), 1–28.

Mousivand, A.J., Alimohammadi Sarab, A., Shayan, S., 2007. A new approach of predicting land use and land cover changes by satellite imagery and Markov chain model (Case study: Tehran). MSc Thesis. Tarbiat Modares University, Tehran, Iran.

Overmars, Verburg, 2006. Multilevel modeling of land use from field to village level in the Philippines. Agricultural Systems 80, 435–456.

Pampel, F.C., 2000. Logistic Regression: A Primer. Sage quantitative applications in the Social Sciences Series, vol. 132. Sage Publications, Thousand Oaks, CA.

Park, S.Y., Jeon, S.W., Kim, S.Y., Choi, C.U., 2011. Prediction and Comparison of Urban Growth by Land Suitability Index Mapping Using GIS & RS in South Korea. Landscape and Urban Planning 99 (2), 104–114.

Poelmans, L., Van Rompaey, A., 2009. Complexity and performance of urban expansion models. Computers, Environment and Urban Systems 34 (1), 17–27.

Pontius Jr., R.G., 2000. Quantification error versus location error in the comparison of categorical maps. Photogrammetric Engineering and Remote Sensing 66 (8), 1011–1016.

Pontius Jr., R.G., Schneider, L.C., 2001. Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. Agriculture, Ecosystems & Environment 85 (1–3), 239–248.

Pontius Jr., R.G., Huffaker, D., Denman, K., 2004. Useful techniques of validation for spatially explicit land-change models. Ecological Modelling 179 (4), 445–461.

Pontius Jr., R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing 32 (15), 4407–4429.

Shahshahani, S., 2003. Tehran: Paradox City, IIAS, Newsletter, N31, pp. 15–16.

Torrens, P.M., 2006. Geosimulation and its Application to Urban Growth Modeling. Springer-Verlag, London, pp.119–134.

Vance, C., Iovanna, R., 2006. Analyzing spatial hierarchies in remotely sensed data: insights from a multilevel model of tropical deforestation. Land Use Policy 23, 226–236.

Verhagen, P., 2007. Case Studies in Archaeological Predictive Modeling. Amsterdam University Press, ISBN 9087280076.

Xie, C., Huang, B., Claramunt, C., Chandramouli, C., 2005. Spatial logistic regression and GIS to model rural-urban land conversion. In: Proceedings of PROCESSUS Second International Colloquium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications, June 12–15, 2005, University of Toronto, Canada.

Wolfram, S., 1984. Cellular automata as models of complexity. Nature 311, 419–424.

WTO (World Trade Center) 1996. Participation of developing countries in World Trade: overview of major trends and underlying factors, WTO Internal Report, Available at: http://www.wto.org/english/tratop_e/devel_e/w15.htm.

Wu, F., 1996. A linguistic cellular automata simulation approach for sustainable land development in a fast growing region. Computers, Environment and Urban Systems 20 (6), 367–387.

Yang, X., Lo, C.P., 2002. Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area. International Journal of Remote Sensing 23 (9), 1775–1798.

Zhang, R., Tang, C., Ma, S., Yuan, H., Gao, L., Fan, W., 2010. Using Markov chains to analyze changes in wetland trends in arid Yinchuan Plain, China. Mathematical and Computer Modelling.