Article

An integrated parcel-based land use change model using cellular automata and decision tree

Florencio Ballestores Jr. [1], Zeyuan Qiu[2]

[1]Department of Chemical Engineering, University of Philippine Diliman, Quezon City, Philippines

[2]Environmental Policy Studies, New Jersey Institute of Technology, University Heights, Newark, New Jersey 07102, USA

E-mail: zeyuan.qiu@njit.edu

Received 6 March 2012; Accepted 10 April 2012; Published online 5 June 2012

IAEES

Abstract

Ecological changes are driven by changes in land use. Modeling land use change is an essential step to adaptively manage ecosystem to mitigate the negative impacts of such ecological changes. This study developed a parcel-based spatial land use change prediction model by coupling a couple of machine learning and interpretation algorithms: cellular automata and decision tree. The model was developed and validated using the historical land use data in Hunterdon County of New Jersey in the United States. Specifically, the data on historical land uses and various driving factors that affect land use changes for Hunterdon County were collected and processed using a Geographic Information System. A set of transition rules illustrating the land use change processes during the period 1986-1995 were developed using decision tree J48 Classifier. The derived transition rules were applied to the 1995 land use data in a cellular automata model Agent Analyst to predict future spatial land use pattern, which were then validated by the actual land use in 2002. The decision tree-based cellular automata model has reasonable overall accuracy of 84.46 percent in predicting land use changes and the Cohen's Kappa Index is 0.644. The model shows much higher capacity in predicting the quantitative changes than the locational changes in land use. Sensitivity analysis indicates that simply changing the size of neighborhood has slight impacts on the simulation results, but insignificant impacts on the model accuracy.

Keywords land use change; cellular automata; decision tree; parcel; geographic information system; J48 Classifier; Agent Analyst.

1 Introduction

Land use in the United States and many other parts of the world has been experiencing rampant changes over the last several decades because of the profound social and economic changes in the society. Land use changes not only in return bring about significant social and economic changes, but also have profound impacts on human health and the natural environment. Since such changes result from the interplay of complex socioeconomic and biophysical processes, they are impossible to duplicate through experiments (Walker, 2003; Verburg et al., 2004). Modeling becomes an important tool to simulate land use changes (Baker, 1989; Briassoulis, 2000). Appropriately calibrated land use change models can be used to predict future land use changes and to explore land use system response to policy interventions through "what-if" scenarios for local,

regional and/or global land use decisions (Riebsame et al., 1994; Pielke et al., 1999; Kalnay and Cai, 2003; Reid et al., 2004; Salmun and Molod, 2006).

Land use change modeling has become sophisticated with substantial advances in spatial sciences and technologies. Cellular automata (CA) emerged as one of the most effective tools to simulate local and regional land use changes (White and Engelen, 1993; Clarke et al., 1997; Batty et al., 1999; Chen et al., 2002; Cheng and Masser, 2004). CA is a collection of cells that evolves through a number of discrete time steps according to a set of rules based on the states of its neighboring cells. In a CA-based land use change model, a cell is defined as the smallest geographic unit where land use changes are being evaluated. CA has five principal elements: cell state, lattice, neighborhood, time and transition rule. Cell state represents one of finite land use types that a cell is in. Lattice refers to the space in which the CA exists and evolves over time and usually represents the geographic region under consideration. The neighborhood comprises the localized region in a CA lattice and is a group of cells surrounding the cell being assessed. Transition rules specify how cells change from one state to another based on cell's own state and its neighborhood conditions. The complication of CA-based land use change models depends on how cell, cell states, neighborhood and the transition rules are being defined.

Early CA-based land use change models tended to define the cell as regular square-shaped grids because most land use maps are prepared using the conventional per-pixel land use classification derived from the spectral signature of a regular pixel. These models using regular grids can be easily integrated with a raster Geographic Information Systems (GIS). Recent development in CA-based land use change modeling defines the cell as irregular cadastral parcel (Stevens et al., 2007). Parcels are generally recognized as the most proper unit of analysis to evaluate land use changes (Landis and Zhang, 1998a, 1998b; Irwin and Geoghegan, 2001; Allen and Lu, 2003). Many land use decisions such as purchasing, selling, and developing land are made and observed at the parcel level. It is at the parcel level that most land use policies such as zoning are crafted and implemented.

Another significant progress in CA-based land use change modeling is the development of the methods that elicit the transition rules, i.e. how various driving factors such as the initial state of the parcel being evaluated, land use conditions of the neighboring parcels, suitability, accessibility to roads and sewers, local and regional land use regulations and policies collectively dictate local land use changes. Traditional land use change models defined the transition rules using the regression models. Regression analyses are dependent on expert knowledge and as such prone to subjectivity bias as discussed by Verburg et al. (2004). Recent CA-based land use change models developed several new methods to elicit realistic and objective transition rules. Wu (1996) introduced the fuzzy logic in a land use change model to evaluate the impacts of different urban development policies. Li and Yeh (2002) and Pijanowski et al. (2002) used artificial neural networks to elicit land use change patterns based on the information on the existing land use change and their driving factors. Some CA-based land use change models incorporate stochastic transition rules to imitate stochastic land use change processes (Ward et al., 2000; de Kok et al., 2001; Guan et al., 2005). Decision Tree (DT), a machine learning and interpretation algorithm for classification, has also been used to elicit the transition rules in CA-based land use change models (Li and Yeh, 2004; McDonald and Urban, 2006; Liu et al., 2007).

This study also aims to develop a loosely coupled CA-based land use change model by applying DT to elicit the transition rules that govern land use changes for evaluating the impacts of urban growth management policies. The model will use irregular cadastral parcel as the basic unit of analysis. It also uses the more realistic and finer land use classification types instead of a simple, dichotomous, urban/non-urban classification scheme. DT will help address a greater range of complications and avoid the subjectivity bias when eliciting transition rules from numerous driving factors and neighborhood effects. The machine learning

and interpretation approach for deriving transition rules does not require extensive quantitative skills and would be better appreciated by non-technical users such as stakeholders and land use change decision makers. The coupled DT and CA-based land use change model are implemented through a GIS-based Agent Analyst model Recursive Porous Agent Simulation Toolkit (RePAST) (North et al., 2005). The coupled land use change model is applied to Hunterdon County, New Jersey, where dramatic land uses have taken place during last three decades.

2 Study Area

Hunterdon County is one of 21 counties in New Jersey in the United States and encompasses 1,094 km^2 of the western portion of the State. It ranks eighth among New Jersey's counties in terms of land area and has 26 municipalities. As shown in Fig. 1, the County is traversed from east to west by the I-78 interstate highway designed to carry traffic between regions of the state and to serve as a corridor between Port Newark/Liberty Airport and points westward. Accessibility between municipalities and adjoining counties is provided by a network of county and municipal roads that includes Routes 12, 31, 202, and 517. Hunterdon County is home to approximately 129,000 people (NJDLWD, 2006). The population in Hunterdon County grew by 87 percent between 1970 and 2004 making it the third fastest growing county in New Jersey. Hunterdon County also experienced considerable economic growth owing to its proximity to high growth areas in the state as firms like Exxon, Foster Wheeler, and Merck established their corporate offices in the county in 1980s and 1990s.

Hunterdon County is still considered to be a mostly rural and suburban county. Population growth and economic development has been shaping the land use pattern in the county. The high density residential developments are more typical during the 1970s and 1980s. However, residential development has been gradually shifted to single-family houses on large lots during the last two decades. Hunterdon is one of six counties situated in an extensive growth area known as "the wealth belt" characterized by high property values, high population, plenty of jobs and high personal income" (Hughes and Seneca, 1999). Such trend is accelerated by the "ratables chase" policy in New Jersey that encourages local governments to permit more development to maintain the low property tax rate and to finance their public service requirements such as sewer, solid waste collection (HCPB, 2007). To maintain the appearance of a rural and agricultural character, some communities in the county are implementing the large-lot zoning that requires at least 0.8-, 2- and 4-hectare (two-, five- or ten-acre) lots for residential homes. On the other hand, various land use policies have been implemented in the county to restrict the land use development toward smart growth and environmental protection. The notable examples are open space preservation, farmland preservation, and purchase of development rights. About 13 townships, towns, and boroughs in Hunterdon County fall partly or completely within the Highlands Preservation Area. Their future land use development will be subject to more stringent restrictions enforced by the New Jersey Highlands Water Protection and Planning Act (NJDEP, 2005).

3 Methods

The coupled land use change prediction model has two modules: a DT module that generates the transition rule from the driving factors and a CA module that predicts future land use change using the derived transition rules. The two modules and the methods used to evaluate the accuracy of the land use change prediction model are discussed in this section.

3.1 Decision tree: J48

DT is a data mining tool initially used as a classification method (Moore et al., 1991; Speybroeck et al., 2004, Wu et al., 2007). A DT structure entails a series of yes/no questions in which the sequence of the questions that are asked depends on the answers given in the previous question. When applied to land use/cover classification,

the specific questions assume values equivalent to land attributes, the sequence of which eventually determines the appropriate land use/cover classification (Aalders and Aitkenhead, 2006). This model uses the DT algorithm J48 developed by the Machine Learning Group of the University of Waikato, New Zealand. When applied to a training sample dataset that is comprised of a list of instances with a set of attributes, the J48 algorithm operates by recursively splitting the instances in the training sample dataset based on their attribute values to produce a tree that preferably generates just one branch. The first attribute to be chosen is designated as the root of the tree. The instances in the training sample dataset are split among branches based on their attribute values. If an attribute value is continuous, each branch takes a certain range of that value. A new attribute feature (node) is then chosen and the process is repeated for the remaining instances. The process stops at a terminal note when the classification of a branch is pure (i.e., it contains only instances in a certain class). As to what attribute to use for a given split, the choice is based on the attribute having the largest value for information gain (Quinlan, 1996; Goodman and Smyth, 1988). The final decision tree generated by J48 contains various paths from the root to the terminal node. Each path can be translated into a transition rule for its subsequent use in CA.

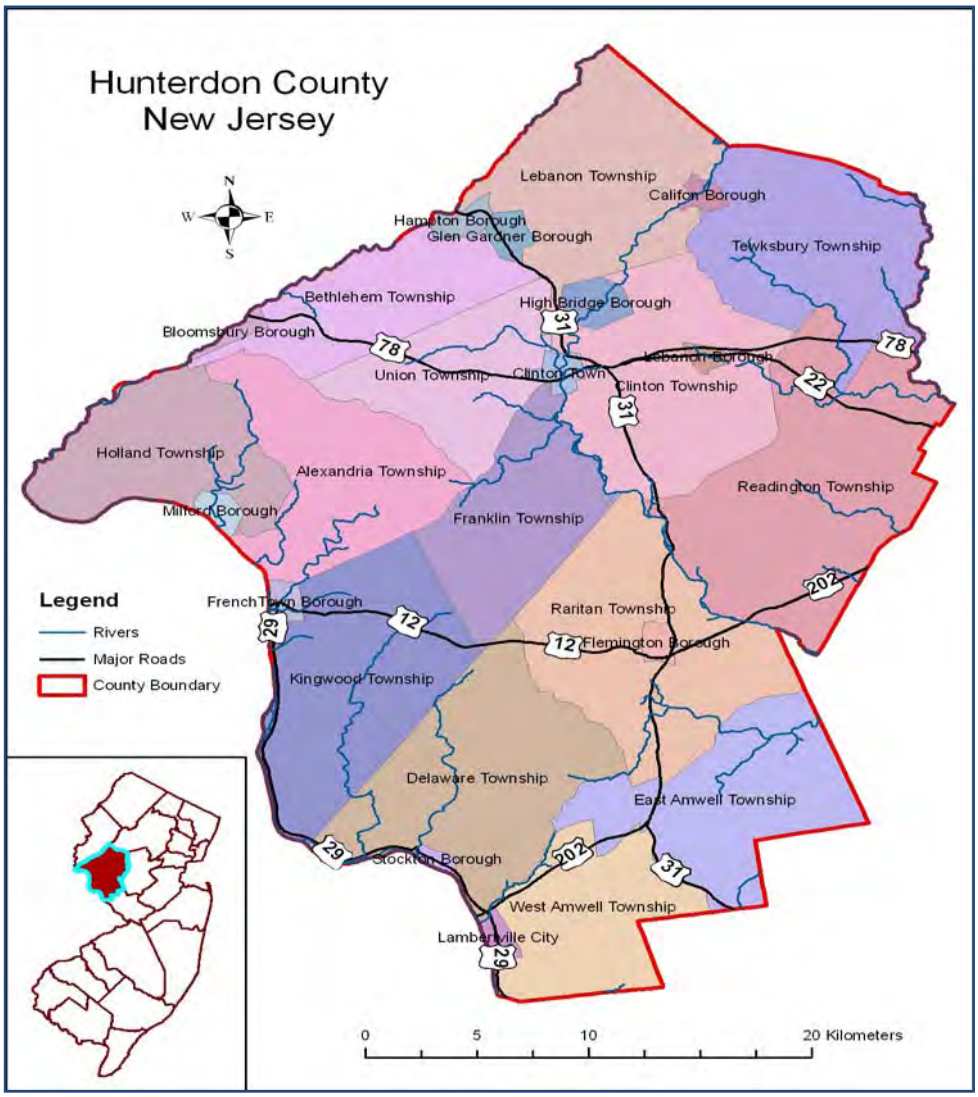

Fig. 1 An overview of Hunterdon County, New Jersey, USA

Creating transition rules through DT is superior than using statistical regressions especially when (1) there is a large number of variables to predict land use changes (Pal and Mather, 2003; Speybroeck et al., 2004); (2) there is the existence of non-linear relationships between variables in the data (Razi and Athappilly, 2005); and (3) the underlying relationship between dependent and independent variables is not known (Pal and Mather, 2003). Although artificial neural network approach has the similar advantages over statistical regression methods, it is not intuitive to policy makers and land use planners because of its black box nature. DT is a white box model and can be easily interpreted (Breiman et al., 1984; Quinlan, 1996; Li and Yeh, 2004).

3.2 Cellular automata: Recursive Porous Agent Simulation Toolkit (RePAST)

In a subsequent step, the transition rules derived using J48 is used by a CA module to evaluate how land parcels are converted from their current land uses to their future land uses. Agent Analyst/RePAST (North et al., 2005) is chosen to implement the CA process. Through Agent Analyst, users can create, edit, and run RePAST model within the ArcGIS environment (Groff, 2007). This graphical user interface allows the modeler to create agents, schedule simulations, visualize the ArcGIS layers, and specify the behavior and interactions of the agents. Aside from having the power and flexibility of ArcGIS, Agent Analyst/RePAST has two outstanding features relevant to this study. First, the model has provisions to allow the modification of agent properties, agent behavioral equations, and model properties during run time. Second, it has libraries for genetic algorithms and neural networks, including the ability to handle irregular grids and vector data as a model component.

The transition rules embedded in the CA model can also be modified to include government regulatory policies on land use changes. Such policies may include regulations or limitations on the conversion of agricultural land to developed land, the steering of urban development to where the soils are considered low value for agriculture, or the designation of zoning laws. Those policies can be specified as additional transition rules to be included in CA model. For examples, the preservation of farmland and open space was represented by specifying transition rules that designates these areas as non-developable.

3.3 Accuracy assessment

The assessment of land use change prediction accuracy relies on the use of a confusion matrix, which is simply a cross tabulation of the predicted land use types against the same land use types in reference. Suppose there are *M* types of land uses. Let the information in a row corresponds to the land use type *i* in reference and in a column shows the predicted land use type *j* by the land use change model. The element in the confusion matrix, n_{ij}, where *i, j* = 1, …, *M*, represents the number of instances in reference land use type *i* that are predicted to land use type *j*. The diagonal elements of the matrix where *i* = *j*, n_{ii} or n_{jj}, represent correct predictions. The overall accuracy is calculated as the sum of the diagonal elements divided by the total instances number of, i.e. $\sum_{i=1}^{M} n_{ii} \Big/ N$, where $N = \sum_{i=1}^{M}\sum_{j=1}^{M} n_{ij}$.

Although overall accuracy is useful, it does not give much information about the accuracy of the individual land use types, which are usually evaluated by user's accuracy, producer's accuracy, errors of commission, and errors of omission (Story and Congalton, 1986). The user's accuracy and the error of commission is a pair of complementary measures that indicate how well instances from a predicted land use type represents that land use type in reference. Let $N_{+j} = \sum_{i=1}^{M} n_{ij}$ be the total number of instances in predicted land use type *j*. The

user's accuracy is estimated as n_{jj} / N_{+j} for j = 1, …, M. The error of commission is simply equal to (100 percent – the user's accuracy). The producer's accuracy and the error of omission, on the other hand, express how well a reference land use type has been predicted correctly. Let $N_{i+} = \sum_{j=1}^{M} n_{ij}$ be the total number of instances in land use type i in reference. The producer's accuracy is n_{ii} / N_{i+} for i = 1, …, M., The error of omission is equal to (100 percent – the producer's accuracy).

The overall prediction accuracy is usually considered as an overestimation since it does not account for agreements that would have occurred by chance (Lillesand and Kiefer, 2000). Another way to assess prediction accuracy is to use the Cohen's Kappa Index, a measure that will eliminate the agreements made by pure chance (Cohen, 1960; Monserud and Leemans, 1992; Pontius and Cheuk, 2006). According to Foody (2002), the Cohen's Kappa Index, $\hat{K}$, can be estimated as $\frac{N\sum_{j=1}^{M} n_{jj} - \sum_{j=1}^{M} N_{j+} N_{+j}}{N^2 - \sum_{j=1}^{M} N_{j+} N_{+j}}$, where all variables are defined above. This index ranges between 0 and 1 and is interpreted as the proportionate reduction in error achieved by the model being evaluated as compared with the error of a completely random prediction model.

Pontius (2000) argued that Kappa Index also has limitations in assessing prediction accuracy. Specifically, the index does not give information about location and quantification errors. Quantification error occurs when the number of parcels for a given land use type predicted is different from that land use type in reference. Location error occurs when the predicted land use type of a given parcel is different from that in reference. Pontius (2000) further derived two variants of Kappa Index, namely, $K_{location}$ and $K_{quantity}$, that measure the accuracy in predicting location and quantity, respectively. The calculation of $K_{location}$ and $K_{quantity}$ involves complicated transformation of the confusion matrix and can be found in Pontius (2000). In this application, the overall accuracy, error of commission, error of omission, Kappa Index and its two variants are used to evaluate the performance of the land use change prediction model.

4 Data Analysis

Three sets of land use/cover data in 1986, 1995, and 2002 maintained by NJDEP were used to apply the land use change prediction model in Hunterdon County. The land use data were compiled from aerial photography and Landsat satellite images. The land use/cover is classified into 6 categories as agriculture, barren lands, forest, urban, wetlands and water based on a modified Anderson Classification System. The model uses all six land use types. A land parcel layer for the county is obtained from the Hunterdon County Office of GIS. Besides the land use and land use parcel data, other spatial data such as digital elevation models (DEM), soil, streams, major roads and urban centers are also obtained from the NJDEP Bureau of GIS and/or the Hunterdon County Office of GIS in digital format.

4.1 Development of parcel-based land uses

A parcel-based land use change prediction model requires a single land use type assigned to each land parcel, which is in fact a challenging task. For example, a residential parcel in a low density residential development area may look like a forest as the house is blended into its natural surroundings. Wu et al (2007) used municipal tax assessment database to evaluation land use changes. Local land assessment records for real estate tax purposes may contain the information for the intended uses of land parcels, but not necessarily

reflect their actual land uses. In this application, the land use type for a parcel was assessed from the land use/cover data.

A parcel may include multiple land user/covers after overlaying the land use/cover layers with the land parcel layer. The following classification scheme was developed to assign a single land use to a parcel to develop the parcel-based land uses in 1986, 1995, and 2002 in the county. Each parcel is initially tested for any agricultural land present. If the agricultural land is over 45 percent in a parcel, the parcel is classified as agricultural lands. This threshold of 45 percent is based on the percentage of agricultural land in all parcels in 2002, which has a mean of 19 percent with a standard deviation of 26 percent.

For a parcel with less than 45 percent of agricultural land, the urban area in the parcel is compared to a threshold value of 0.5 hectares. The threshold value represents a typical house footprint in the region including the area occupied by house, driveway, patio, pool, and etc (NJWSA, 2003). If the parcel with the urban area greater than 0.5 hectares is located in a residential, commercial or industrial zone and the parcel size is less than 4 hectares (10 acres), the parcel is classified as urban. If the urban area is less than 0.5 hectares, but represents more than 45 percent of the parcel, the parcel is also classified as urban. For a parcel that failed to be classified as agriculture and urban, its final designation will be determined by the dominant land uses in the parcel.

It would be ideal to have the re-classified land use distribution re-assembly the original land use distribution. However, the classification scheme used in this study consistently over-allocates land to agriculture parcels by approximately 26 percent in all three years. At the same time, the scheme under-allocates the land areas to forest, urban use, and wetlands parcels by 13 percent, 21 percent, and 38 percent, respectively. Many other schemes had been tried, but this one gave the best accuracy.

4.2 Definition of the neighborhood

The usual neighborhood configurations used in CA models include the von Neumann or the Moore patterns that assume a lattice composed of regularly shaped cells or grids. Since irregular shaped land parcels are used as a unit of analysis in this study, an alternative neighborhood configuration has to be defined accordingly. In their parcel-based CA model, Stevens et al. (2007) evaluated three neighborhood configurations: (A) an adjacency neighborhood that includes all parcels having a common edge with the central parcel; (B) a distance neighborhood that includes parcels that fall completely or partially within a certain distance of buffer from the edge of the central parcel; and (C) a clipped distance neighborhood that includes all parcels that fall completely and portions of the parcels that are partially within a certain distance of buffer from the edge of the central parcel.

The neighborhood of a parcel in this study is defined by an external buffer with a thickness of 145 meters (m) around the edge of a parcel shown in Fig. 2. It is very similar to the neighborhood configuration (C) discussed, but it is different from the configuration (C) by excluding the central parcel itself. Assuming that all the parcels are squares in Hunterdon County, the side length of an average square-shaped parcel is estimated to be 145 m. The original land use GIS layers were overlaid with the buffer to identify the percentages of different land use types within the buffer, which are used as driving factors in the land use change prediction model to determine the future land use of the parcel.

4.3 Derivation of driving factors

Allen and Lu (2003) suggested three criteria in selecting an appropriate set of driving factors for modeling land use change. First, the factors should include all physical, economic, demographic and social factors that affect all types of land use change. Second, they must have spatial attributes. Third, they must reflect the properties and characteristics of the parcel. Jiao and Boerboom (2003) grouped various driving forces into five categories namely: neighborhood, accessibility, suitability, policy, and socio-economic factors. Following those

principles and examples, this study considers the following driving factors: land use type of a parcel, distribution of land uses in a parcel's neighborhood in terms of percentages, parcel size, amount of wetlands within a parcel, distances from the center of a parcel to its nearest streams, major roads, and urban centers, average slope, and number of soil restrictions for urban development. Table 1 presents various driving factors in selected publications and also in this study.

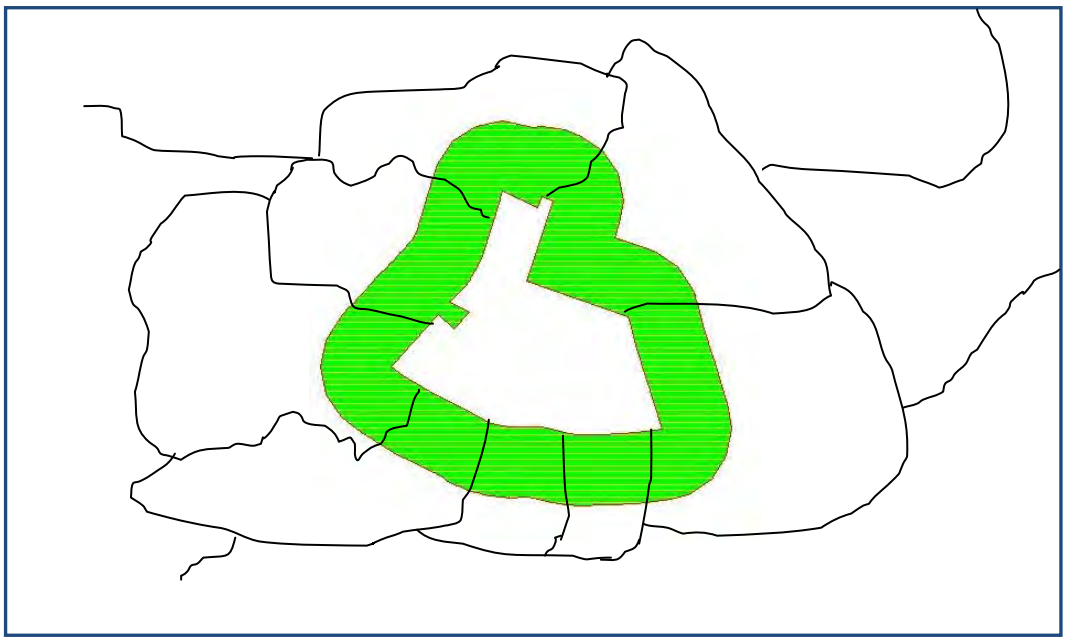

Fig. 2 A hypothetical configuration of the neighborhood used in Cellular Automata modeling

Table 1 The list of driving factors in land use changes

Authors	**Driving Factors Used**
Stevens et al. (2007)	Distance to parks; distance to commercial areas; distance to light industrial land; distance to heavy industrial land; adjacency to existing developed and undeveloped land
Li and Yeh (2004)	Urban conversion; distance to city proper; distance to town centers; distance to roads; distance to expressways; distance to railways; number of developed cells in the neighborhood; current land use agricultural suitability; slope
Allen & Lu (2001)	Distance to tourist attraction features; distance to roads; distance to sewer line; distance to central business district; distance to nearest neighborhood; elevation; slope; parcel size; parcel ownership; drainage; policy constraints: protected land, residential zone, commercial zone, subdivision, and urban boundary; population density; housing unit density; housing unit value; initial land use
Waddell (2000)	Current development in neighbor; policy constraint: zoning regulations; land and improvement values; distance to highways; distance to existing development; regional accessibility to population
Moreno and Marceau (2006)	Area of parcel; distance to adjacent polygon; transformation probability to neighbor land use type based on area
This Study	**Land use type of a parcel, distribution of land uses in a parcel's neighborhood in terms of percentages, parcel size, amount of wetlands within a parcel, distances from the center of a parcel to its nearest streams, major roads, and urban centers, average slope, and number of soil restrictions for urban development.**

The identification of the land use type for a parcel and the land use distribution in its neighborhood was discussed above. The parcel size was calculated from the land parcel layer. The amount of wetlands within a parcel was calculated by overlaying the wetlands layer extracted from the land use/cover data with the land parcel layer. A point layer for the centers of parcels was created by extracting the geometric centers of the polygons from the land parcel layer. The distances from the center of a parcel to its nearest streams, roads, and urban centers were calculated by conducting NEAR analyses in ArcGIS between the point layer and the respective line layers. The stream layer used is the 2002 stream, the newest stream data maintained by NJDEP. The major roads include all county roads and interstate highways. Population density from the Census data was used to identify the location of urban centers. The census tract having the highest population density in each municipality was then the designated location of its urban center. The slope was calculated from the 10-meter DEM maintained by NJDEP. The average slope for each parcel was calculated using the Spatial Analyst in ArcGIS by overlaying the slope raster with the parcel layer.

Soil suitability for development is an important factor that drives urban development. NJWSA (2002) evaluated all soils in the Raritan River Basin including Hunterdon County in term of their suitability for the six of 24 community development applications as defined by United State Department of Agriculture – Natural Resources Conservation Service (NRCS) soil survey. The six pertinent elements of development include septic tank absorption fields; foundation for dwellings with basement; foundation for dwellings without basement; foundation for small commercial buildings; local streets and roads; and lawns, landscaping, and golf fairways. The Soil Survey Geographic data was obtained from the NRCS. The dominant soil type in each parcel was identified by overlaying soil data layer over the parcel layer. The number of restrictions to these six development application in each parcel were identified based on the soil suitability assessment by NJWSA (2002).

5 Results

When applying the land use change prediction model, the transition rules were derived using 1986 and 1995 re-classified parcel-based land uses in Hunterdon County, New Jersey and the discussed driving factors in the DT module. The derived transition rules were then used in the CA module to predict future land use pattern using the parcel-based land uses in 1995. Since the transition rules were based on cumulative land use changes in a 9-year period from 1986 to 1996, the future land use changes predicted by the model should reflect the land use pattern in 2004, the 9^{th} year from 1995. Since no land use data for 2004 was available, the land use data in 2002 was used as a reference to evaluate he model performance using an array of accuracy measurements discussed previously.

5.1 Transition rules derived from land use changes from 1986 to 1995

A complicated DT structure was generated using J48 and was then converted to the transition rules in the subsequent CA modeling. It is difficult to report all the transition rules, but some significant rules are noticed below. The first split in the decision tree is current land use type. Although most urban parcels remain in urban uses, they could be converted to other uses such as agriculture and forest. This occurs when the parcels are large, have a big portion of wetlands within them, three and more soil restrictions to urban development, lower percentage of urban and barren land, and higher percentage of water, agricultural, forest, wetlands in the neighborhood, and are farther away from highways and urban centers with steeper slope.

Agricultural parcels could be converted into urban, forest, barren and wetlands depending primarily on their neighborhood land use distribution and parcel size. Agricultural parcels tend to be converted into urban uses when there is high percentage of urban land in their neighborhood and the parcel size is small. The conversion from agriculture to forest could occur to those parcels with steep slopes where there are severely

restricted soils for development and a high percentage of forest in their neighborhoods. Agricultural parcels with a significant amount of barren land in their neighborhood have the potential of becoming barren land. The conversion to wetlands usually occurs to the large agricultural parcels that have significant amount of wetlands in it.

Forest parcels with a high percentage of urban land and a low percentage of barren land in their neighborhood are usually converted to urban use. This type of conversion tends to occur in the case of small forest parcels. Forest parcels could be converted to agriculture or wetlands when there is a significant presence of agricultural land already in their neighborhood or wetlands within these parcels, respectively.

The actual amount of wetlands within a wetland parcel usually determines its future status. The wetland parcels with a large amount of wetlands within it always remain as wetlands. However, wetland parcels with smaller amounts of wetlands have higher likelihood of converting into urban in a high urban neighborhood or become barren lands if the parcel has three or more restrictions to urban development. Barren parcels can be developed into urban lands or remain as barren. Finally, the land parcels classified as water, or artificial and natural lakes usually stay as water.

The accuracy of the transition rules were evaluated by testing the derived transition rules against a testing dataset which is about two-thirds of the randomly selected land parcels in the county. To do this, all the attributes except the land use class in 1995 of an instance in the dataset were fed into the decision tree that consists of all the transition rules to predict the land use class in 1995. Such process was repeated for all instances in the dataset. The accuracy is computed by dividing the total number of instances with the correct predictions by the total number of instances in the dataset. The accuracy of these transition rules for predicting the land uses in 1995 from the land use in 1986 was 81.4 percent when using a training sample dataset.

Table 2 The predicted land use distribution based on the land uses in 1995 in Hunterdon County, New Jersey, USA

Land Use in 1995	Predicted Land Use						
	Agriculture	Barren	Forest	Urban	Water	Wetlands	Total
Agriculture	44,130 (4,463)		28 (24)	2,440 (1,434)			46,598 (5,921)
Barren		249 (318)		64 (80)			313 (398)
Forest			25,367 (6,872)	10,018 (3,306)			35,384 (10,178)
Urban		8 (38)	1 (1)	17,870 (33,827)			17,878 (33,866)
Water					2,722 (142)		2,722 (142)
Wetlands			115 (35)	2,078 (739)		4,332 (763)	6,525 (1,537)
Total	44,130 (4,463)	257 (356)	25,511 (6,932)	32,470 (39,386)	2,722 (142)	4,332 (763)	109,420 (52,042)

The numbers in parentheses indicates of number of parcels.

5.2 Predicted land use changes using the land use in 1995

Table 2 presents the predicted land use changes during the next simulation period, i.e. 1995-2004, from 1995. The model predicted that the urban areas increase from 17,878 hectares (33,866 parcels) in 1995 to 32,470 hectares (39,386 parcels) by the end of the simulation period, i.e. 2004, while the area of other land uses decrease except water which remains the same. Forest is the biggest contributor to the urban development in that period followed by agriculture and wetlands. The total forest loss to urban development is 10,018 hectares

(3,306 parcels). 2,440 hectares (1,434 parcels) of agricultural lands are given away to urban development. There are 2,078 hectares (739 parcels) of wetlands losses to urban development, which is quite significant since this amount represents almost a third of the county's remaining wetlands in 1995. The forest loss was partially offset with the reforestation of 144 hectares from wetlands (35 parcels), agriculture (24 parcels), and urban parcels (1 parcel).

Table 3 Confusion matrix for evaluating accuracy based on the land uses in 2002

2002 Land Uses	Predicted Land Uses in Terms of the Number of Parcels								
	Agriculture	Barren	Forest	Urban	Water	Wetlands	Total Parcels	Misclassified	Error of omission (%)
Agriculture	3,171		98	1,057		12	4,338	1,167	26.90
Barren	97	15	15	52		2	181	166	91.71
Forest	270	9	5,787	3,385	13	32	9,496	3,709	39.06
Urban	912	332	955	34,241	17	68	36,525	2,284	6.25
Water	4		25	40	105	14	188	83	44.15
Wetland	9		52	611	7	635	1,314	679	51.67
Total Parcels	4,463	356	6,932	39,386	142	763	52,042		
Misclassified	1,292	341	1,145	5,145	37	128		8,088	
Error of commission, %	28.95	95.79	16.52	13.06	26.06	16.78			
2002 Land Uses	Predicted Land Uses in Terms of the Total Area, Hectares								
	Agriculture	Barren	Forest	Urban	Water	Wetland	Total Area	Misclassified	Error of omission, %
Agriculture	40,264		490	2,829		130	43,714	3,449	7.89
Barren	372	107	131	77		2	689	582	84.47
Forest	1,942	46	23,629	9,651	18	156	35,444	11,814	33.33
Urban	1,400	104	1,069	17,929	10	85	20,597	2,669	12.96
Water	25		33	34	2,691	38	2,820	130	4.61
Wetland	127		159	1,948	4	3,919	6,156	2,238	36.35
Total Area	44,130	257	25,511	32,468	2,722	4,332	109,420		
Misclassified	3,865	150	1,881	14,540	32	413		20,881	
Error of commission, %	8.76	58.52	7.37	44.78	1.16	9.54			

5.3 Model evaluation using the land use in 2002

Table 3 presents the confusion matrix in terms of the number of parcels (the upper panel) and of the area (the lower panel) using the predicted land use distribution and the actual land uses in 2002. As shown in the upper panel of the table, the overall prediction accuracy in terms of the number of parcels, computed as the sum of the agreements in the diagonals (43,954 parcels) divided by the total number of parcels (52,042 parcels), is 84.46 percent. Similarly, the overall prediction accuracy is 80.92 percent (88,539 ha divided by 109,420 ha) in terms of the total acreage as shown in the lower panel of Table 3. These measurements are comparable to the values reported in the literature. Li and Yeh (2004) reported an overall accuracy of 82 percent using a DT-based CA for predicting land use change in an urbanizing city in Southern China. Allen and Lu (2003) developed a multinomial logistic land use change model with the parcel as the unit of analysis and achieved an overall accuracy of 80.76 percent in terms of number of parcels.

Table 3 also shows the error of omission and the error of commission. Omission error varies across land use categories. There are large error of omission for barren, wetlands and water, but they account for a very

small portion of the county. On the other hand, the urban and agriculture, two major land use categories have low error of omission. Similar observations are also found for the error of commission. Table 3 also shows that agricultural and urban lands are consistently underpredicted while water and wetlands are overpredicted in terms of both the total numbers of parcels and acreage.

The agreement between the predicted land use distribution and the actual land uses in 2002 was also evaluated by Cohen's Kappa Index and its two variants to eliminate the agreements by pure chance. The calculated Kappa Index is 0.644, which indicates that the two patterns are in a moderate agreement based on Congalton (2001) and Landis and Koch (1977).

Two variants of the standard Kappa Index were also calculated to evaluate the agreement between the predicted land use distribution and the actual land uses in 2002: $K_{location}$ and $K_{quantity}$ following Pontius (2000). $K_{location}$ is equal to 0.748, which indicates the model has good capacity to specify location correctly. $K_{quantity}$ is equal to 0.925, which indicates the model has excellent capacity to specify quantity correctly. These numbers suggested that the model has better capacity to predict the quantitative changes than the locational changes in Hunterdon County.

5.4 Sensitivity of neighborhood size

A change in the neighborhood size directly affects the values of the driving factors. The results presented above are based on the neighborhood size of 145 m further from the boundary of a parcel. The sensitivity analysis evaluates whether the different neighborhood sizes improve the modeling accuracy. Two neighborhood scenarios considered are 72 m and 217 m away from the boundary of a parcel. Table 4 presents the contingency table and Kappa indices when comparing the predicted land use pattern to the 2002 reference land use pattern for each scenario.

There are slight differences in the predicted future land use changes using the two neighborhood sizes. Take the three land use classes with large areas – agriculture, forest, and urban as examples. The model using the 72-m neighborhood predicts 3,217 agricultural parcels in 2004 while the model using the 217-m neighborhood 3,142 agricultural parcels. The difference is about 75 parcels, which is equivalent to 2.3 percent of total agricultural parcels. Predictions for forest and urban lands differed by 24 (0.41 percent) and 313 (0.92 percent) parcels, respectively.

The overall accuracy is 84 percent and the Kappa index is 0.64 when comparing the predicted land use pattern using the 72-m neighborhood to the land use pattern in 2002. The overall accuracy is 85 percent and the Kappa index is 0.65 when using the 217-m neighborhood. The relatively small discrepancy in the two values of the Kappa index confirms previous observations indicating that there is no marked difference in model outcomes between the two neighborhood sizes in the application. These accuracy measurements are very similar to the measurements using the 145-m neighborhood as discussed above (i.e. the 84.46 percent of overall accuracy and the Kappa Index of 0.644.

6 Summary and Conclusions

This study develops a DT-based CA model to predict future land use changes with parcel-level data. The model was evaluated using the historical land use changes in Hunterdon County, New Jersey. The model defines the modeling space as a collection of geographic objects of irregular shape that are spatially represented by land parcels and defining the transition rules using a knowledge discovery algorithm DT. The neighborhood of each parcel was defined as an external buffer along the boundary of the parcel. A DT elicits land use patterns from a large set of driving factors and is free from the subjectivity biases often encountered in expert knowledge based methods. The DT approach also offers the convenience of incorporating the land

use policies such as down-zoning, open space and farmland preservation when predicting the future land use changes.

Table 4 The Impacts of Two Different Sizes of Neighborhoods on the Overall Accuracy and the Resulting Kappa Index

The Number of Parcels in 2002	The Number of the Simulated Parcels with The Neighborhood Buffer Width of 72 m						
	Agriculture	Barren	Forest	Urban	Water	Wetlands	Total
Agriculture	3,217		104	1,006		11	4,338
Barren	97	15	17	50		2	181
Forest	297	9	5,785	3,359	13	33	9,496
Urban	1,013	376	982	34,077	17	60	36,525
Water	4		25	39	105	15	188
Wetlands	9		44	644	7	610	1,314
Total	4,637	400	6,957	39,175	142	731	52,042
Overall Accuracy							84.18
Kappa Index							0.64
The Number of Parcels in 2002	The Number of the Simulated Parcels with The Neighborhood Buffer Width of 217 m						
	Agriculture	Barren	Forest	Urban	Water	Wetlands	Total
Agriculture	3,142		91	1,094		11	4,338
Barren	88	13	15	63		2	181
Forest	256	8	5,809	3,376	13	34	9,496
Urban	804	297	944	34,390	17	73	36,525
Water	3		28	38	105	14	188
Wetlands	10		47	590	7	660	1,314
Total	4,303	318	6,934	39,551	142	794	52,042
Overall Accuracy							84.78
Kappa Index							0.65

The coupled DT-based CA model reasonably predicts the land use changes in the Hunterdon County, New Jersey, where substantial land use changes have taken place during the last three decades. Using the historical land use changes during the period 1995-2002 as a reference, the model achieves an overall accuracy of 80.92 in terms of the total areas and of 84.46 percent in terms of the total number of land parcels. The Kappa Index, the conventional statistics for comparing similarity of two spatial patterns, is measured at 0.644. Two variants of the Kappa Index are also calculated to evaluate the model's ability to correctly predict location and quantity and are 0.748 and 0.925, respectively. Such results indicate the model has the higher capacity of predicting the quantitative changes than the locational changes in land uses in the study area. This study defines the neighborhood of a parcel by a 145-m buffer from the boundary of the parcel. The sensitivity analyses using the 72-m and 217-m buffers shows the definition of the neighborhood has no significant impacts on the model's prediction accuracy in this study area. Caution should be given when generalizing this result to other studies and areas. Some studies showed that the CA model was sensitive to changes in model elements such as the neighborhood configuration (Chen and Mynett, 2003).

The application of the coupled model in Hunterdon County, New Jersey demonstrates the feasibility and effectiveness of using parcel-level data in land use change modeling. However, there are still challenges that

need to be addressed in the future land use change modeling using the parcel-level data. First, the model assumes a single land use for each parcel. It is a challenging task to assign a single land use to a parcel based on a land use map compiled from satellite images and/or aerial photography especially when the study area is too large for detailed field verification. As discussed previously some land use classes were overestimated, while others were underestimated. The accuracy of assigning the correct land uses would have significant impacts on the overall accuracy of the modeling. Although the transition rules on the derived parcel-based land uses achieves reasonable prediction accuracy, the overall accuracy could be further improved by improving the accuracy of assigning a single land use to a parcel based on the current land use data derived from aerial and remote sensing imagery. Second, the model assumes the parcel boundary stays the same during the modeling process, which is ideal. A parcel itself may evolve over time. For example a large agricultural or forest parcel could be divided into several smaller parcels in urban development. Potential improvement could be made by recent developments in using the dynamic vector agent in CA-based models (Ménard and Marceau, 2005; Hammam et al., 2007; and Moreno et al., 2009).

It should be recognized that assessing the accuracy of the land use change model is a fast evolving science. As argued by White (2006), the cell-to-cell comparison methods as discussed above for assessing the accuracy of the simulation results are useful, but limited since most land use change models especially CA-based models emphasize similarity in spatial patterns rather than attribute matching at a specific location. Future research may consider using more advanced pattern-based map comparison techniques.

References

Aalders IH, Aitkenhead MJ. 2006. Agricultural census data and land use modeling. Computers, Environment and Urban Systems, 30(6): 799-814

Allen J, Lu K. 2003. Modeling and prediction of future urban growth in the Charleston region of South Carolina: A GIS-based integrated approach. Conservation Ecology, 8(2): 2

Baker WL. 1989. A review of models of landscape change. Landscape Ecology, 2(2): 111-133

Batty M, Xie Y, Sun Z. 1999. Modeling urban dynamics through GIS-based cellular automata. Computers, Environment and Urban Systems, 23(3): 205-233

Breiman L, Friedman JH, Olshen R, et al. 1984. Classification and Regression Trees. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, USA

Briassoulis H. 2000. Analysis of Land Use Change: Theoretical and Modeling Approaches, the Web Book of Regional Science. http://www.rri.wvu.edu/WebBook/Briassoulis/contents.htm.

Chen J, Gong P, He C, et al. 2002. Assessment of the urban development plan of Beijing by using a CA-based urban growth model. Photogrammetric Engineering and Remote Sensing, 68(10): 1063-1071

Chen Q, Mynett AE. 2003. Effects of cell size and configuration in cellular automata based prey-predator modelling. Simulation Modelling Practice and Theory, 11(7-8): 609-625

Cheng J, Masser I. 2004. Understanding spatial and temporal processes of urban growth: Cellular automata modeling. Environment and Planning B: Planning and Design, 31(2): 167-194

Clarke KC, 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(2): 247-261

Cohen J. 1960. A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20(1): 37-46

Congalton RG. 2001. Accuracy assessment and validation of remotely sensed and other spatial information. International Journal of Wildland Fire, 10(3-4): 321-328

de Kok, J, Engelen G, White R, et al. 2001. Modeling land-use change in a decision-support system for coastal-zone management. Environmental Modeling and Assessment, 6(2): 123-132

Foody GM. 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment. 80(1): 185-201

Goodman RM, Smyth P. 1988. Decision tree design from a communication theory standpoint. IEEE Transactions on Information Theory, 34(5): 979-994

Groff ER. 2007. 'Situating' simulation to model human spatio-temporal interactions: An example using crime events. Transactions in GIS, 11(4): 507-530

Guan Q, Wang L, Clarke KC. 2005. An artificial-neural-network-based, constrained CA model for simulating urban growth. Cartography and Geographic Information Science, 32(4): 369-380

Hammam Y, Moore A, Whigham P. 2007. The dynamic geometry of geographical vector agents. Computers, Environment and Urban Systems, 31(5): 502-519

Hughes JW, Seneca JJ. 2006. New Jersey's New Economy: Growth Challenges. Rutgers Regional Report No. 25 (Edward J, ed). Bloustein School of Planning and Public Policy, Rutgers University, New Brunswick, New Jersey, USA. http://www.policy.rutgers.edu/news/reports/RRR/RRR_July_2006.pdf

Hunterdon County Planning Board (HCPB). 2007. Growth Management Plan. http://www.co.hunterdon.nj.us/pdf/hcpb/2007GrowthManagmentPlan.pdf.

Irwin EG, Geoghegan J. 2001. Theory, data, methods: Developing spatially explicit economic models of land use change. Agriculture, Ecosystems and Environment, 85(1-3): 7-23

Jiao J, Boerboom L. 2003. Transition Rule Elicitation Methods for Urban Cellular Automata Models in Innovations. In: Design & Decision Support Systems in Architecture and Urban Planning (Van Leeuwen J, Timmermans H, eds). Springer, Netherlands

Kalnay E, Cai M. 2003. Impact of urbanization and land-use change on climate. Nature, 423(6939): 528-531

Landis JR, Koch GG. 1977. A one-way components of variance model for categorical data. Biometrics, 33(4): 671-679

Landis J, Zhang M. 1998a. The second generation of the California urban futures model. part 1: Model logic and theory. Environment and Planning B: Planning and Design, 25(5): 657-666

Landis J, Zhang M. 1998b. The second generation of the california urban futures model. Part 2: Specification and calibration results of the land-use change submodel. Environment and Planning B: Planning and Design, 25(6): 795-824

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

Li X, Yeh AG. 2004. Data mining of cellular automata's transition rules. International Journal of Geographical Information Science, 18(8): 723-744

Lillesand TM, Keifer RW. 2000. Remote Sensing and Image Interpretation. John Wiley and Sons, New York, USA

Liu XP, Li X, Yeh AG, et al. 2007. Discovery of transition rules for geographical cellular automata by using ant colony optimization. Science in China, Series D: Earth Sciences, 50(10): 1578-1588

McDonald RI, Urban DL. 2006. Spatially varying rules of landscape change: Lessons from a case study. Landscape and Urban Planning, 74(1): 7-20

Ménard A, Marceau DJ. 2005. Exploration of spatial scale sensitivity in geographic cellular automata. Environment and Planning B: Planning and Design, 32: 693-714

Monserud RA, Leemans R. 1992. Comparing global vegetation maps with the kappa statistic. Ecological Modelling, 62(4): 275-293

Moore DM, Lees BG, Davey SM. 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. Environmental Management, 15(1): 59-71

Moreno N, Wang F, Marceau DJ. 2009. Implementation of a dynamic neighborhood in a land-use vector-based geographic cellular automata model. Computers, Environment and Urban Systems, 33: 44-54

New Jersey Department of Environmental Protection (NJDEP). 2005. Guidelines for the Highlands Protection and Planning Act. http://www.nj.gov/dep/highlands/faq_info.htm.

New Jersey Department of Labor and Workforce Development (NJDLWD). 2006. Northern Regional Community Fact Book. Hunterdon County Edition.

New Jersey Water Supply Authority (NJWSA). 2002. Landscape of the Raritan River Basin: A Technical Report for the Raritan Basin Watershed Management Project. Somerset, New Jersey: Watershed Protection Unit, NJWSA, USA

New Jersey Water Supply Authority (NJWSA). 2003. Watershed Model for Watersheds of the Spruce Run Reservoir. Somerset, New Jersey: Watershed Protection Unit, NJWSA, USA

North MJ, Howe TR, Collier NT, et al. 2005. Repast Simphony Development Environment, In: Proceedings of the Agent 2005 Conference on Generative Social Processes, Models, and Mechanisms, ANL/DIS-06-1 (Macal CM, North MJ, Sallach D, eds). Co-sponsored by Argonne National Laboratory and The University of Chicago, USA

Pal M, Mather PM. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86(4): 554-565

Pielke Sr RA, Walko RL, Steyaert LT, et al. 1999. The influence of anthropogenic landscape changes on weather in south Florida. Monthly Weather Review, 127(7): 1663-1672

Pijanowski BC, Brown DG, Shellito BA, et al. 2002. Using neural networks and GIS to forecast land use changes: A land transformation model. Computers, Environment and Urban Systems, 26(6): 553-575

Pontius Jr RG. 2000. Quantification error versus location error in comparison of categorical maps. Photogrammetric Engineering and Remote Sensing, 66(8): 1011-1016

Pontius Jr RG, Cheuk ML. 2006. A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. International Journal of Geographical Information Science, 20(1): 1-30

Quinlan JR. 1996. Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, 4: 77-90

Razi MA, Athappilly K. 2005. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert Systems with Applications, 29(1): 65-74

Reid RS, Thornton PK, McCrabb GJ, et al. 2004. Is it possible to mitigate greenhouse gas emissions in pastoral ecosystems of the tropics? Environment, Development and Sustainability, 6(1-2): 91-109

Riebsame WE, Meyer WB, Turner II BL. 1994. Modeling land use and cover as part of global environmental change. Climatic Change, 28(1-2): 45-64

Salmun H, Molod A. 2006. Progress in modeling the impact of land cover change on the global climate. Progress in Physical Geography, 30(6): 737-749

Speybroeck N, Berkvens D, Mfoukou-Ntsakala A, et al. 2004. Classification trees versus multinomial models in the analysis of urban farming systems in central Africa. Agricultural Systems. 80(2), 133-149.

Stevens D, Dragicevic S, Rothley K. 2007. iCity: A GIS-CA modelling tool for urban planning and decision making. Environmental Modelling and Software, 22(6): 761-773

Story M, Congalton RG. 1986. Accuracy assessment: A user's perspective. Photogrammetric Engineering & Remote Sensing, 52(3): 397-399

Verburg PH, Schot PP, Dijst MJ, et al. 2004. Land use change modeling: current practice and research priorities. GeoJournal, 61(4), 309-324

Waddell P. 2000. A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim. Environment and Planning B: Planning and Design, 27: 247-263

Walker R. 2003. Evaluating the performance of spatially explicit models. Photogrammetric Engineering and Remote Sensing, 69(11): 1271-1278

Ward DP, Murray AT, Phinn SR. 2000. A stochastically constrained cellular model of urban growth. Computers, Environment and Urban Systems, 24(6): 539-558

White R. 2006. Pattern based map comparisons. Journal of Geographical Systems, 8(2): 145-164

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

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

Wu S, Silván-Cárdenas J, Wang L. 2007. Per-field urban land use classification based on tax parcel boundaries. International Journal of Remote Sensing, 28(12): 2777-2801