Available online at www.sciencedirect.com
SCIENCE DIRECT®
Landscape and Urban Planning 74 (2006) 7–20
LANDSCAPE AND URBAN PLANNING
This article is also available online at:
www.elsevier.com/locate/landurbplan

Spatially varying rules of landscape change: lessons from a case study

Robert I. McDonald*, Dean L. Urban

Landscape Ecology Laboratory, Nicholas School of the Environment and Earth Sciences, Duke University, Durham, NC 27708-0328, USA

Received 9 February 2004; received in revised form 22 July 2004; accepted 25 August 2004
Available online 27 October 2004

Abstract

Land-cover and land-use change modeling have become increasingly common, and myriad different modeling techniques are now available. Many techniques assume that the rules of landscape change are the same everywhere within the study area, an assumption that contrasts with reality in many municipal regions, which have spatially varying development restrictions. In this paper, we provide a case study from the Raleigh–Durham area of North Carolina (USA) showing the consequences of using a model with a spatially homogeneous form when the rules of landscape change are spatially heterogeneous. Using classified Thematic Mapper images of 1990 and 2000, we fit two models relating probability of deforestation to a large set of potentially explanatory variables. Potential autocorrelation in the error term of our models was avoided by sampling outside the zone of spatial autocorrelation. The first model, a logistic regression (GLM), was used as an example of a simple, spatially homogeneous model, where the probability of deforestation is a function of a set of explanatory variables. The second model was a classification and regression tree analysis (CART), a spatially heterogeneous model in which the data were recursively partitioned on the same explanatory variables plus spatially explicit indicator variables, to create a binary decision tree that adequately captured the pattern in deforestation. Overall, the CART model (15.2% misclassification rate) performed significantly better than the GLM model (33.1% misclassification rate). When the residuals of both models were examined spatially, the CART model appears to perform better, more accurately predicting hotspots of development and predicting the baseline proportion of deforested pixels more accurately. Our results lend support to the importance of spatial heterogeneity in the rules of landscape change, and suggest that models that attend local variability in the forces driving landscape change can provide more useful predictions than models that assume these forces operate similarly throughout the landscape.

Keywords: Classification and regression trees; Deforestation; Land-use change; Logistic regression; North Carolina; Thematic Mapper

1. Introduction

Worldwide, the dramatic expansion of developed areas has raised concern about the associated loss of habitat (Houghton, 1994). In the United States 19,800,000

* Corresponding author. Present address: Harvard Forest, Box 68, Petersham, MA 01366-0068, USA. Tel.: +1 978 724 3302; fax: +1 978 724 3595.
E-mail address: rimcdon@fas.harvard.edu (R.I. McDonald).

0169-2046/$20.00 © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.landurbplan.2004.08.005

acres were converted into urban or suburban areas from 1992 to 2001 (USDA, 2003). The most fragmented patterns of development have occurred in a few rapidly growing areas in the Southern and Western US, cities that have primarily developed after the widespread availability of the automobile made dispersed development possible (Jenerette and Wu, 2001; Theobald, 2001; Waisanen and Bliss, 2002). In the Raleigh–Durham–Chapel Hill metropolitan area of North Carolina, USA (the Triangle), where this study was conducted, population grew by 38% from 1990 to 2000 (Triangle J Council of Governments, 2003a). At the same time, analysis of land-cover maps suggests that 20.5% of forested lands have been deforested (unpublished data). Given the likely negative ecological consequences of such widespread deforestation (cf., Harrison and Bruna, 1999; Fahrig, 2002; Laurance et al., 2002), understanding the factors driving sprawl and their dynamics over time is very important.

One way to understand land-cover and land-use change is through modeling, where the probability of development of some region in the study area is represented as a function of a set of explanatory variables. While the palette of available model types has increased dramatically in recent years (Veldkamp and Lambin, 2001), and there has been a tendency towards more complex modeling frameworks (Parker et al., 2003), most models have assumed that the functional form of the model is spatially homogeneous. This has certain conceptual and statistical implications, some of which may impede heuristic understanding of the factors driving sprawl. Note that our discussion and model classification scheme in the next section is not normative; different modeling approaches will be useful for different purposes. Our goal in this paper is thus merely to show that in many cases the degree of spatial flexibility in the functional form of a model will affect the results obtained.

1.1. Spatially homogeneous versus spatially heterogeneous models

Early models of land-cover and land-use change were essentially non-spatial, and the probability of change for all locations was estimated as the proportion of cells that changed state in the last time step (e.g., Johnson, 1977). This describes a Markov-chain process, where the probability of some change in land-cover (e.g., development, from forest to non-forest), $P(d)$, was constant and equal for all forested sites. Later models incorporated a set of explanatory variables that might themselves be spatially patterned, such that:

$$P(d)_i = f(\boldsymbol{x}_i) + \varepsilon \quad (1)$$

where (taking the example of a raster lattice) the probability of development at pixel i is a function of a vector $\boldsymbol{x}$ of explanatory variables measured at location i. This function f can be as complicated as desired, as can the explanatory variables contained within $\boldsymbol{x}$. For example, $\boldsymbol{x}$ can contain information on the state of nearby pixels, such as the proportion of neighboring cells that have already been deforested, giving the model attributes of a cellular automata (von Neumann, 1966; Wolfram, 1984; Hogeweg, 1988). In agent-based models, some of the most mechanistic models available, the decision function of the agents is analogous to the function f above. If different types of agents with different preferences are used, there are, in effect, multiple functions $f, g, h, \ldots$ for different types of agents (Parker et al., 2003).

All of the models described above are composed of functions that are spatially homogeneous; the function f does not vary between different pixels i. In essence, while the input into the model varies with space (and hence the model output varies with space), the *rules* of landscape change remain spatially constant. However, many of the rules that govern land-cover and land-use change vary from place to place. For example, zoning in some regions of the landscape is tightly enforced, while in other regions zoning plays a relatively unimportant role in controlling development. More generally, the desires of agents often vary between locations because of local interests or cultures. In effect, these commonplace examples suggest that the function f differs between places: the *rules* of landscape change are spatially heterogeneous.

Insights can be gained by exploring the implications of using spatially homogeneous models to describe a land-use change process that is spatial heterogeneous. One probable effect is an 'averaging' of landscape dynamics; in most statistical fitting procedures the function f will try to accommodate multiple variables that may be important in particular spots on the landscape (cf., Chambers and Hastie, 1992). The importance of certain variables (e.g., the absolute magnitude of a co-

efficient in a generalized linear model) is therefore underestimated in the particular places where they really matter and overestimated elsewhere (Vayssieres et al., 2000; Urban et al., 2002). A related effect is that the model's residuals should be spatially autocorrelated (Legendre and Legendre, 1998). This latter effect may be difficult to detect, as autocorrelation in residuals can result from other problems, and is apparent in spatial examination of residuals in almost all land-cover and land-use change models (Veldkamp et al., 2001).

Note that the distinction between spatially heterogeneous and spatially homogeneous models is fuzzy, and is really one of degree rather than kind. If the explanatory variables in $\boldsymbol{x}$ are highly structured spatially, then some of the spatial pattern that occurs in the actual probability of land-use change will be captured by the estimate of $P(d)_i$. In the limit, an arbitrarily complicated transition function f, combined with a sufficient set of explanatory variables $\boldsymbol{x}$, can be made to do arbitrarily well in capturing the dynamics of a system. Nevertheless, some modeling frameworks, such as classification trees (Moore et al., 1991; De'Ath and Fabricius, 2000) and neural nets (e.g., Fletcher and Goss, 1993; Miller et al., 1995), are extremely flexible in their estimation of f. In particular, if some of the variables in $\boldsymbol{x}$ effectively partition the study area spatially, then such flexible frameworks are essentially spatially heterogeneous: the *rules* of landscape change may differ between locations. Depending on the purposes of the land-use change modeling, this spatial heterogeneity might dramatically affect model performance.

1.2. Objectives

In this paper, we offer a comparison between a spatially homogeneous model and a spatially heterogeneous model, by parameterizing both models to replicate patterns of deforestation (a change in land-cover) in the Triangle region of the North Carolina Piedmont from 1990 to 2000. We model deforestation because of its relevance to forest ecologists and ease of measurement from satellite photos. The models described below will also interest urban planners and others in the Triangle (e.g., Mansfield et al., 2003). In addition, these models will help constrain estimates of successional forest processes discussed in a companion paper (McDonald and Urban, in preparation).

More broadly, we hope to compare and contrast these two classes of models, to allow for increased understanding of the consequences of assuming spatial homogeneity in the functional form of a model. Specifically, we describe the negative effects of the spatial 'averaging' that occurs with the spatially homogeneous model. Using a unique method to analyze the residuals in a spatial manner, we illustrate specific places where both of our models perform poorly, which yields insights into processes not adequately captured in our models.

2. Methods

2.1. Study area

Our study area is the Triangle metropolitan region of North Carolina, defined here as the three counties that include the cities of Raleigh, Durham, and Chapel Hill (Fig. 1). The population of the region has grown rapidly from 270,000 in the 1950s to greater than 1.5 million at present. The past decade experienced particularly high population growth rates of over 40%. The four main employment centers—Raleigh, Durham, Chapel Hill, and the Research Triangle Park, a high-technology industrial zone—are separate from the suburban municipalities that have experienced the most growth, creating a patchwork of different regulatory environments. Research Triangle Park has risen in prominence in recent years and now plays a pre-eminent economic role in the region, despite a lack of housing nearby. The dispersed style of development that has resulted concerns regional planners and reflects processes occurring in many other cities in the United States (Triangle J Council of Governments, 2003a).

The physical environment controls and modifies development in the region in a complex manner. The gentle rolling topography of the North Carolina Piedmont presents few direct limitations to deforestation and land-use change, although steep slopes in scattered localities limit development. Most municipalities limit development within a fixed buffer zone near streams, defined either by a fixed distance or by floodplain boundaries. Soils vary widely in plasticity and permeability, and thus the cost of building and the ease of installing septic tanks (for areas without sewer service) can vary considerably (Triangle J Council

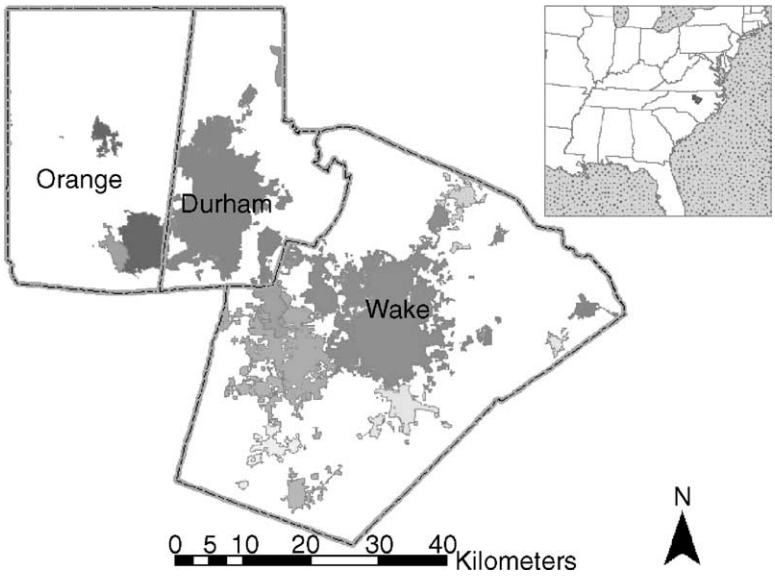

Fig. 1. Map of three-county study area with inset locator map. Different municipalities are colored differently, to show the diversity of municipalities in the region.

of Governments, 2003b). Former agricultural lands are now dominated by loblolly pine (*Pinus taeda)*, while forests that were never clear-cut remain dominated by a complex mix of hardwood species (Oosting, 1942; Christensen and Peet, 1984). The difference between these two forest states is not considered explicitly in this study, as discussions with developers suggested that the state of the forest played little role in deforestation decisions. A mosaic of protected federal, state, and local areas exists throughout the region, playing a crucial role in limiting development in some regions (Triangle J Council of Governments, 2003b).

The political environment of the region also controls and modifies development patterns. The various municipalities of the three-county region vary in their friendliness to development, and the area thus presents an interesting challenge to modelers. The economy is driven by high-tech product development and testing by pharmaceutical and computer companies that exist in a few centralized zones. However, like many rapidly growing cities the largest percentage gains in employment are in the retail and service sectors of the economy, which tend to be much more dispersed. The lack of adequate public transit (which, arguably, lowers incentive to live near developed areas) and the cheap price of land outside traditional city centers has caused most growth in the region to occur in a few municipalities that, until recently, were considered small towns. The heterogeneous wants and desires of different municipalities suggest that the constraints on development will vary in different locations (Triangle J Council of Governments, 2003a).

2.2. Data preparation

Two Thematic Mapper images (30 m pixels) of the region for May 1990 and 2000 were obtained, and georectified to other geographic data layers to within 10 m accuracy, as assessed with the root mean square (rms) error tool in Erdas Imagine. Atmospheric correction was conducted using dark-object subtraction (Song et al., 2001). The images were classified using a supervised maximum-likelihood classification in Erdas Imagine on log-transformed spectral data (to meet the assumptions of discriminent analysis), based on a training data set derived from high-resolution aerial photos. Originally, seven land cover classes were used (dark water, sediment-laden water, sparse vegetation, hardwood forest, mixed forest, pine forest, developed/barren), defined to maximize spectral discrimination between the classes. For the purposes of this study, this classification scheme was collapsed into forest ver-

sus non-forest (excluding water), to simplify the analysis and to focus most clearly on deforestation events.

Positive spatial autocorrelation is common in environmental datasets (Legendre and Fortin, 1989), and poses a problem for statistical testing because it tends to inflate the significance of most test metrics (Dale, 1999). The probability of deforestation in any pixel of our land-cover map can be seen as a function of a set of explanatory variables and the local small-scale autocorrelation in the process of deforestation. The autocorrelation in the process (i.e., the size of the average deforestation event) would be of interest if one wanted to model all the dynamics of land-cover change, but we chose not to fit this autocorrelated error term here because our interest for this study was to highlight the relationship between the explanatory variables and the probability of deforestation. Thus, we took the approach of estimating the scale of spatial autocorrelation of development events, and then made sure the samples used for statistical modeling are beyond the zone of autocorrelation (i.e., are statistically independent).

We sampled 10,000 pixels in areas that were forested in 1990, and assigned a 1 if they changed to another land cover type and 0 if they did not. Following the method of Sokal and Oden (1978), we calculated the number of similar states (joint-counts) between pairs of pixels at various distances from one another (0–300, 300–600 m, etc.). We tested the significance of any departures from the null expectation of no spatial autocorrelation (Sokal and Oden, 1978). As each distance class involves a separate statistical test, and we wanted to avoid the pitfalls of multiple statistical tests, we used the progressive Bonferroni correction of Legendre and Legendre (1998). Note that the correlogram approach used here is unlikely to reach the strict requirement of second-order stationarity (Legendre and Legendre, 1998), and so statistical tests should be interpreted with caution, although the approach has been widely used in similar situations and should be adequate to delineate the zone of spatial autocorrelation.

After determining the range of spatial autocorrelation, we randomly selected 1500 points (i.e., pixels) within the study region using a sequential interference design to exclude points within the zone of autocorrelation, with the constraint that all sample points must have been forested in 1990 (i.e., we are not considering deforestation events prior to this period). At each point, the value of the dependent variable (CHANGED) was extracted using GIS, as well as a set of potential explanatory variables (Table 1). Explanatory variables were included that accounted for several major factors that impact the probability of development. The county and municipality (1990 maps) of each sample point were recorded as a discrete variable, as different municipalities might have different probabilities of deforestation. A discrete variable, PROTECTED, was created, with a value of 1 if the sampling point fell into a protected area (e.g., State Park, conservation easement) and a 0 if it did not. Given information on the road (TIGER data) and stream (USGS data) network, we calculated the Euclidean distance to a road or stream, as a proxy for ease of development. To gain a more useful picture of how location might influence the probability of deforestation, we calculated the driving time from every sample point to the centers of Raleigh, Durham, Chapel Hill, and the Research Triangle Park, using the network functions of ArcGIS. Soil plasticity index for the B-horizon for each sample point was extracted from digitized soil survey maps (SSURGO data) and associated attribute databases available via the USDA Natural Resources Conservation Service, as a proxy for ease of construction and permeability for septic lines. Using a digital elevation model from the SRTM dataset (NASA), we calculated slope using predefined functions in ArcGIS. A topographic convergence index (TCI, Beven and Kirkby, 1979) was also calculated as an index for hydrologic inflow of water, as a proxy for soil wetness and site suitability for development. Finally, the proportion of developed cells that within a series of circular buffers from the sampled cell (0–30, 30–60, 60–120, 120–240, 240–480 and 480–960 m) was calculated, as a measure of the

Table 1
Explanatory variables used in both classes of models

Variables	Units
Driving time to UNC, Duke, Raleigh, and RTP	Minutes
Slope	Degrees
Distance to stream	Meters
TCI	Unit-less index of soil moisture
Proportion sparse vegetation cells within: 30, 60, 120, 240, 480 and 960 m	Proportion
Proportion development cells within: 30, 60, 120, 240, 480 and 960 m	Proportion

spatial contagion of the process of development (i.e., are forested areas near developed areas more likely to be developed in the future than forested areas far from developed areas). Similarly, the proportion of sparse vegetation within the same sets of buffer zones was calculated.

2.3. Generalized linear model

The probability of development was modeled with a logistic regression:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta\boldsymbol{x} + \varepsilon$$

where p is the probability of development, and $\boldsymbol{x}$ is the vector of explanatory variables. A logistic regression is a specific form of a generalized linear model (GLM), an example of a spatially homogeneous model: the per-unit effect of the explanatory variables is constant across the landscape. Of course, given spatially structured explanatory variables and numerous interaction terms, it is possible to achieve a degree of spatial heterogeneity in the rules of landscape change in a GLM. In practice, however, the search for a parsimonious set of explanatory variables means that most GLMs are mostly spatially homogeneous in the rules of landscape change.

The explanatory variables used here are a mix of continuous and discrete variables (Table 1). All Pearson's correlation coefficients between explanatory variables are less than 0.5, with the exception of the three driving time variables, which are moderately correlated for driving-times to UNC, Duke, and RTP ($R \sim 0.8$ in all cases). In addition, there is moderate correlation ($R < 0.7$ in all cases) for the circular buffer calculations when two buffers are of similar sizes (e.g., 0–30 and 30–60 m). To avoid multicollinearity problems, we used stepwise regression to select the most significant variables (Sokal and Rohlf, 1981). The linearity assumption implicit in this approach was examined graphically, and no major departures from this assumption were found. As the use of the logit link function here was arbitrary, we also examined the probit and complementary log–log link functions, but found no substantial differences between the models.

After fitting the logistic regression using SPLUS 6.1 (Insightful Corporation), the overall fit of the model was evaluated by classifying the training data as either predicted to have changed ($\hat{P} > b$) or to have remained forested ($\hat{P} < b$), where b is some threshold constant. The best value of b was determined for our data using receiver–operator characteristic (ROC) curves (Pearce and Ferrier, 2000). After classifying the training data, the accuracy of the classification was calculated using the Kappa statistic and a measure of available mutual information (AMI, Wilkie and Finn, 1996). This statistic calculates the increased information available from the classification scheme, relative to a null model of random classification.

Of potentially more interest for land-cover change modeling is examining *where* on the landscape the model does well and where it fails. This is challenging because the observed change between 1990 and 2000 is binary (deforestation either occurred, or it did not) while the predicted value from the model is a probability. Moreover, observed values that are close to one another have a tendency to be similar due to the small-scale spatial autocorrelation of the process of development, which we are not modeling here. To overcome this, we calculated for each pixel on the landscape the proportion of forested cells within a 1.5 km radius that were deforested. The distance 1.5 km was chosen to be twice the distance of significant spatial autocorrelation in the process (see below), and is somewhat arbitrary, although analyses using different ranges do not change the results significantly. This strategy scales the observed values from 0 to 1, and makes sure that the pattern displayed smoothes over the small-scale autocorrelation (see Section 3). We then subtracted the predicted probability of development from the actual proportion of the forest developed, and graphically examined the residuals.

2.4. Classification tree model

Classification tree models (CART) are a non-parametric approach that recursively partitions a dataset into subsets that are increasingly homogeneous with regard to a response variable, based on an optimal binary split on one of a set of explanatory variables (Moore et al., 1991; Vayssieres et al., 2000). This recursive partitioning means that with spatially partitioned explanatory variables, CART becomes essentially spatially heterogenous in the rules of landscape change. It should be noted that several other statistical techniques (e.g. neural nets) could have also

been used as an example of a spatially heterogeneous model; CART was chosen for its simplicity and ease of presentation.

To avoid overfitting the CART model (i.e., making it too sensitive to variation peculiar to the sample dataset), trees are usually pruned (the number of binary decisions is reduced) to find a consistent set of rules that has meaning beyond the specific sample used to create it. We used a 10-fold cross-validation to find an optimal level of complexity for our CART model. We tested other forms of cross-validation, and found that the final form of the tree was insensitive to the form of cross-validation. After the final form of the CART model was decided, we examined surrogate variables (i.e., explanatory variables that would have reduced deviance almost as much at a given split in the tree), to gain insight into other forms of the tree that would have similar values of reduction of deviance.

The deviance at each step is reduced as much as possible by a binary division of the data. Note that this is not the exact same form of the deviance statistic typically used in the GLM model, which makes model comparison difficult. As CART automatically generates classification of the training data, there was no need for the ROC curves optimization as described above with the GLM model. A Kappa statistic and AMI were calculated comparing predicted versus actual deforestation events. The results between the CART and GLM models are thus directly comparable by comparing their Kappa statistics.

As with the GLM model, a graphical comparison of residuals is often more interesting than global metrics of fit. We used the same smoothed observed data described above to summarize the actual deforestation events that occurred, and the proportion of sample points that were developed in each node as a continuous measure of probability of development. The residual then was calculated as predicted minus observed, and the results examined graphically. The spatial pattern between the residuals was thus easily comparable between the GLM and the CART models, and any arbitrariness in smoothing of empirical maps is at least equally arbitrary for both models.

3. Results

3.1. Autocorrelation

There is significant positive spatial autocorrelation in the process of deforestation between points that are closer than 750 m (Fig. 2). In the first two distance

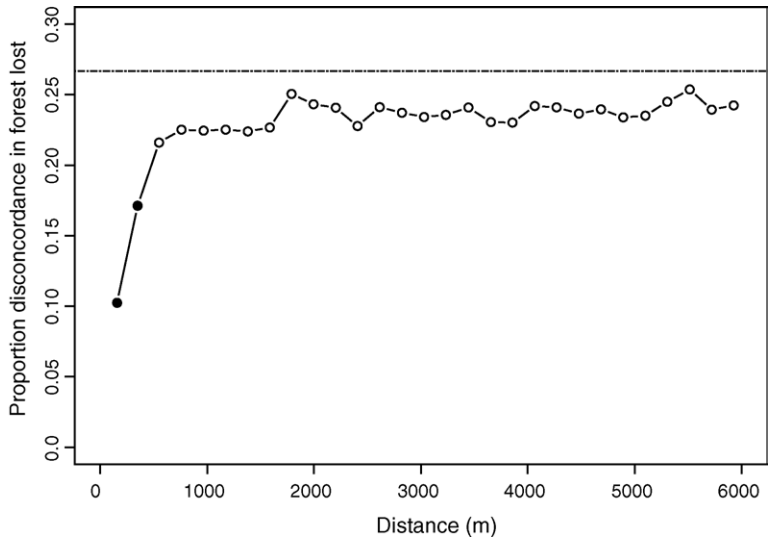

Fig. 2. Scale of spatial autocorrelation in the process of deforestation, as measured by joint-count statistics between pairs of points in various distance classes. A pair of points is considered discordant if one point was deforested and the other was not. The dotted line indicates the null expectation (see text).

classes tested (i.e., pairs of points less than 750 m), the proportion of sample pairs that were discordant (i.e., one sample point was deforested while the other was not) was significantly less than that expected if the process of deforestation was distributed at random. The proportion of sample pairs that were discordant for all other distance classes was not significantly different from the null hypothesis. However, the proportion discordant remains below that expected under a random spatial process, so some slight spatial autocorrelation seems to persist at larger spatial scales. One interpretation of our results is that the intense, small-scale positive spatial autocorrelation is the zone of autocorrelation in the deforestation process (i.e., the size of the deforestation patch), while the slight, large-scale positive autocorrelation is due to the slow gradient change in spatially patterned explanatory variables (Legendre and Fortin, 1989).

3.2. GLM

The overall logistic regression equation derived from stepwise regression is shown in Table 2. The first variable to enter the stepwise regression is the factor municipality, and it is highly significant ($P < 0.0001$). The probability of deforestation varies widely in different municipalities. For example, a sampling point in the town of Apex is twice as likely to be deforested as a point in an unincorporated area, while a sampling point in the town of Chapel Hill is five times less likely to be deforested. The next variable to enter into the stepwise regression was the driving time to Durham ($P < 0.0001$), where a 10% increase in drive time from the baseline case increases the probability of deforestation by 8.0%. This result is counterintuitive, as we expected sites close to a city to be more likely to be deforested. However, the remaining patches of forest cover near cities appear less likely to be deforested than forest patches in suburban regions further from the city center, perhaps because of different land-use regulations. Another important variable is the level of conservation protection; protected sites are 34.3% less likely to be deforested than the baseline case ($P = 0.0023$). The remaining variables that entered into the stepwise regression show lower sensitivities, but are in some cases highly statistically significant.

Table 2
Regression coefficients from the GLM model, as well as the sensitivity of each coefficient

Variable	Coefficient	Sensitivity (%)	d.f.	Deviance	Residual deviance	*P*
Intercept	−1.7160	−12.6	1	NA	1250.8	NA
Municipality			13	57.7	1193.1	<0.0001
Apex	1.0178	97.5				
Butner	0.6770	61.4				
Carrboro	0.1119	8.9				
Cary	0.0804	6.4				
Chapel Hill	−0.2916	−20.7				
Durham	−0.0846	−6.4				
Garner	−0.0987	−7.4				
Hillsborough	0.0754	5.9				
Morrisville	0.0776	6.1				
Raleigh	−0.0721	−5.6				
Wake Forest	−0.1183	−8.8				
Fuqay-Varina	−0.0464	−3.5				
Holly Springs	−0.0636	−4.8				
Driving time-duke	0.0375	8.0	1	29.7	1163.4	<0.0001
SV within 30 m	−2.3382	−1.1	1	14.1	1149.3	0.0002
Slope	−0.3999	−3.2	1	12.3	1137.0	0.0005
Protected	−0.5177	−34.4	1	9.3	1127.8	0.0023
Development within 120 m	−1.5056	−0.5	1	7.9	1120.0	0.0050
Development within 30 m	−3.4191	−0.4	1	4.0	1115.9	0.0463
Distance to road	0.0007	1.2	1	2.6	1113.3	0.1052

See text for details.

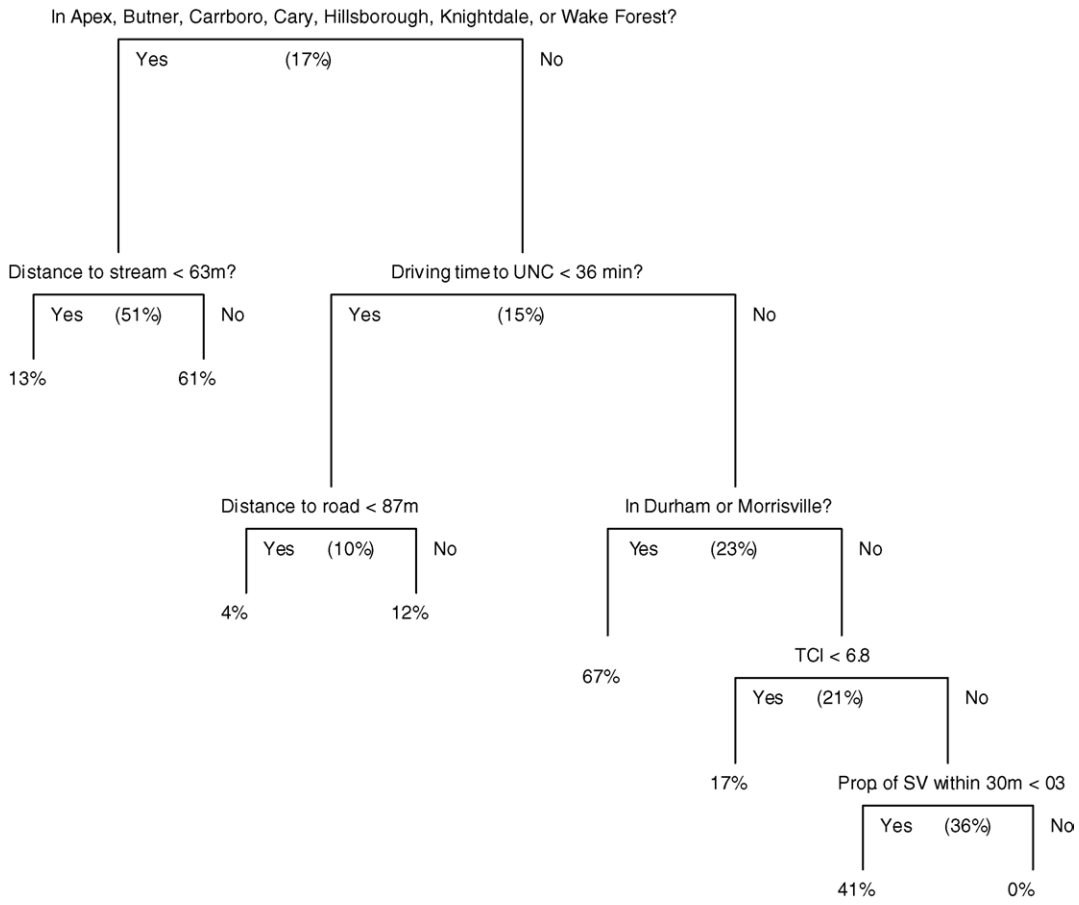

Fig. 3. Graphical representation of the CART tree. At each node, the relevant decision is shown, with the length of each vertical bar proportional to the proportion of the deviance explained by that split. Percentages are the proportion of cells in the training sample that are deforested at this point in the tree. For example, a terminal node with 61% signifies a branch of the tree in which 61% of the training cells were deforested.

3.3. CART

The first split in the CART model is on municipality, with Apex, Butner, Carrboro, Cary, Hillsborough, Knightdale, and Wake Forest having a higher proportion of sample points deforested than in sample points in unincorporated lands or in other municipalities (Fig. 3). Within this selected group of municipalities, the only remaining factor in the CART model is distance to stream, with sites close to rivers less likely to be deforested. In unincorporated lands or other municipalities, the next split is the drive time to Chapel Hill, with sites close to Chapel Hill having a lower proportion of sample points deforested than sites farther away. For sites relatively close to Chapel Hill, the next split is on distance to road, with sites close to a road having a lower probability of deforestation. For sites relatively far from Chapel Hill, the next split is on municipality again; Durham and Morrisville have relatively high probabilities of development, while outside these two municipalities probability of deforestation is lower. The final two splits are on TCI and the proportion of sparse vegetation within a 30 m buffer.

As with the GLM model, municipality is very important, and no other variable would be a good surrogate for the first split of the tree. Throughout the tree, distance to stream can be replaced by TCI, and visa versa, with only a small loss in deviance explained. The split on driving time to UNC can be replaced with a split on driving time to RTP or Durham with lit-

tle loss in deviance explained; all variables partition off the eastern portion of the study area. The proportion of sparse vegetation in a 30 m buffer can be adequately replaced by other small-scale (60 and 120 m) buffers of sparse vegetation, but not by large-scale buffers of sparse vegetation or any scale of buffers of development.

3.4. Comparison of the two models

Overall, the CART model does better at predicting deforestation events than does the GLM model (Table 3). The ROC-optimized GLM model predicts 66.9% of samples correctly ($\kappa = 0.213$, AMI = 58.91), while the CART model predicts 84.8% of samples correctly ($\kappa = 0.252$, AMI = 66.4). An examination of Table 3 shows that, on average, the GLM overpredicts deforestation events, while the CART model underpredicts deforestation events.

Mapping the predicted probability of deforestation reveals interesting trends over the study area (Fig. 4). The GLM model predicts the highest rates of deforestation in the eastern portion of Wake County, presumably because the time it takes to drive to Duke becomes quite high in this region of the map, and hence the predicted probability of development (Fig. 4a). A closer examination of the map reveals several other places where the form of the GLM models is too inflexible. In contrast, the map of predicted probability of deforestation from the CART model shows localized hotspots of defor-

Table 3
Confusion matrix for the GLM and CART techniques

Model prediction	True value		
	Deforested	Not deforested	Predicted totals
Deforested			
GLM	150 (10.8%)	379 (27.2%)	529 (38.0%)
CART	48 (3.5%)	29 (2.1%)	77 (5.5%)
Not deforested			
GLM	81 (5.8%)	781 (56.1%)	862 (62.0%)
CART	183 (13.2%)	1131 (81.3%)	1314 (94.5%)
Observed totals			
GLM	231 (16.7%)	1160 (83.3%)	1391 (100%)
CART	231 (16.7%)	1160 (83.3%)	1391 (100%)

Whole numbers are the number of training pixels that fell into that cell in the table, while numbers in parentheses are the percentage of the training pixels that fell into that cell in the table.

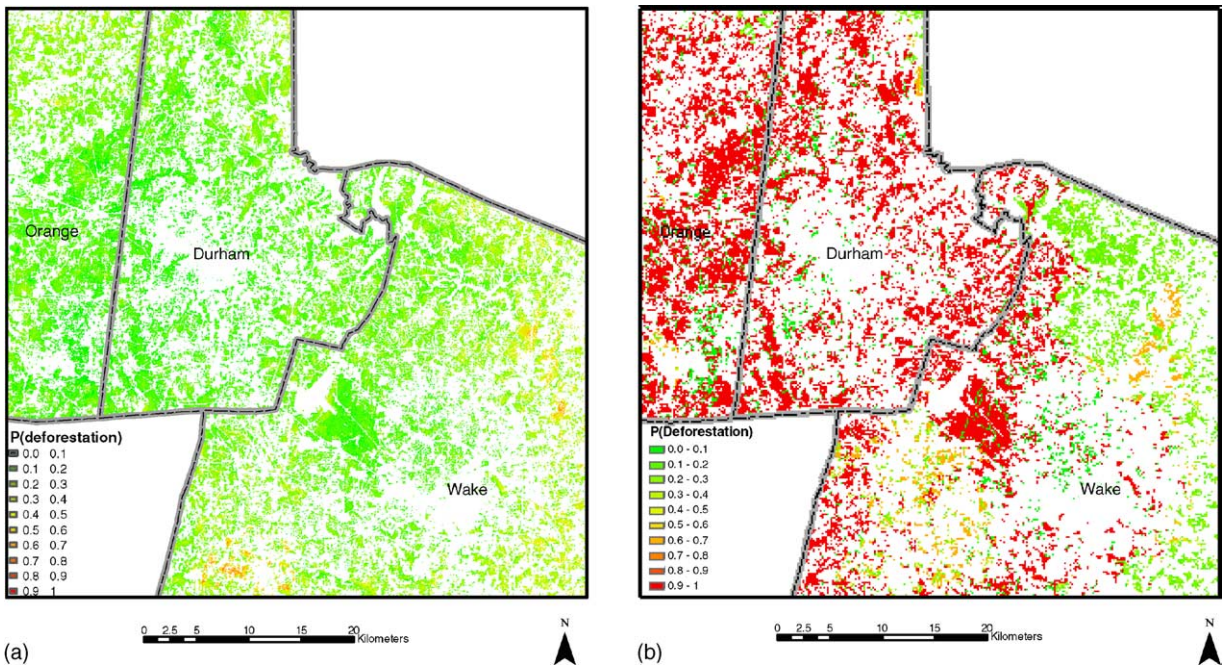

Fig. 4. Two panel figure of predicted probability of deforestation for both models. The left-hand panel is the predicted probability of deforestation for the GLM model, while the right-hand panel is the predicted probability of deforestation for the CART model.

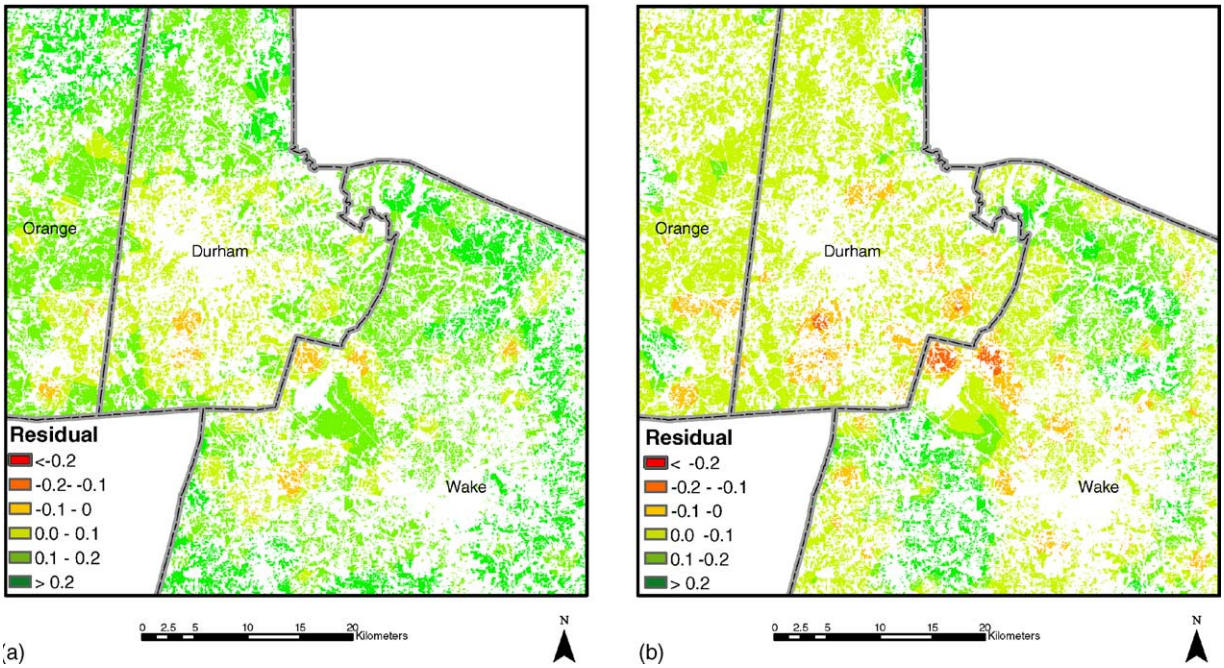

Fig. 5. Two panel figure of residuals (predicted minus actual) for both models. The left-hand panel is the residuals for the GLM model, while the right-hand panel is the residuals for the CART model.

estation based on municipalities and distance to stream (Fig. 4b).

The residuals from the model (Fig. 5) generally are positive, implying that large areas of the landscape are less prone to deforestation than predicted by either model. The problem is most severe for the GLM model (Fig. 5a), where deforestation outside of the city centers is overpredicted by 10–30%. The GLM model underpredicts development slightly in South Durham, Southern Chapel Hill, and in areas near the RDU airport. The CART model, in contrast, overpredicts deforestation outside of city centers by less than 10%, and underpredicts deforestation in some of the same hotspots of development that are problematic for the GLM model.

4. Discussion

4.1. Processes driving development in the Triangle

Both the GLM and the CART model confirm that differences between municipalities are one of the dominant sources of variance in the probability of deforestation. First, this result is partly a reflection of the highly fragmented nature of this metropolitan region, and it is not clear if different regions with a more spatially homogeneous political structure would have similar results. For at least this study area however, this result is reasonable, as each of the myriad municipalities has different goals and desires regarding development. Second, there are circular processes at work that may cloud the detection of municipality-level differences. The boundaries of municipalities change over time in response to development pressure, so that land that is very likely to be developed is just within the municipal boundaries. This trend will tend to cloud the signal of deforestation risk in unincorporated land near the current municipal boundaries, and future modeling work in the region should include some proxy to account for this process (e.g., distance to municipal boundary).

While both classes of models include one of the measures of driving time as a significant variable, the shape of the relationship is in a different direction than expected. In the classic economics model of city development, the spread of development is roughly ra-

dial, covering larger concentric circles of area over time (cf., the discussion of land-rent theories in Berry et al., 1993). This would imply that forested pixels that had lower driving times to the city center should be more likely to be developed (i.e., in a GLM the parameter should have a negative value). However, we find that the remaining forested areas near cities are less likely to be deforested, and that deforestation is more severe far away from cities. In part, this is due to our use of a land-cover map rather than a land-use map, as forested pixels within city centers may reflect big trees in the yards of established neighborhoods, rather than anything that might have the functional properties of an intact forest stand. However, our results also are in accord with the more dispersed style of development described by Jenerette and Wu (2001), and observed by others throughout the Southeast US (e.g., Yang, 2002).

Distance to stream (and its partial correlate TCI) proved to be an important variable in the CART model, but not the GLM. Intuitively, it seems logical that distance to stream would be an important variable, as many developing restrictions occur near streams and in floodplains. Therefore, it is surprising that the GLM does not pick up any relationship with distance to stream. This failure may be because of the 'averaging' phenomenon discussed previously, where the small proportion of the landscape in which distance to stream is important is swamped by the large proportion of the landscape in which it is not. Alternatively, the failure of the GLM model may simply be because of the arbitrariness of the process of forward-stepwise regression, which is not guaranteed to arrive at the optimal fit.

In contrast, slope proved to be an important variable in the GLM model but not in the CART model. Again, it seems logical that slope would be significant, because many restrictions on development are predicated on slope, and thus it is surprising that slope does not play a role in the CART model. Part of the explanation may be that slope is weakly correlated with TCI, and could be substituted for that split in the CART tree, and thus TCI explained part of the variance that slope otherwise would. However, the difference between the two models may simply be due to their radically different natures: the GLM model is fitting globally (to all the data), while the CART model partitions within local subsets of the data.

In both models, at least one of the various buffer zone calculations was important. In the GLM model the amount of development within 30 and 120 m is negatively related to the probability of deforestation. With the CART model, only the amount of sparse vegetation within 30 m plays a role in determining any portion of the tree. In both cases, the direction of the relationship is in the opposite direction expected. Instead of being contagious, the process of development is spatially repulsive at these small scales. This might occur if setbacks and other zoning regulations discouraged deforestation in regions directly adjacent to other developed or cleared land.

4.2. Spatially homogeneous versus spatially heterogeneous models

Based on an examination of overall error rates, the CART (16.2% misclassification) seems to do better than the GLM (34.1% misclassification). The same conclusion would also be reached using more sophisticated metrics of model performance, the Kappa statistics or a measure of the average mutual information. Most of the improved performance for the CART model occurred because the model simply predicted low probability of deforestation for most branches on the tree, except for two nodes that were 'hotspots' of development.

An analysis of the spatial patterns of model residuals is also interesting. The CART model seems to do marginally better than the GLM model over a broad spatial region, as it predicts a lower (and more accurate) average deforestation rate than the GLM model. The inferior performance of the GLM model could be due to the phenomenon of 'averaging', as the spatially homogeneous form tries to accommodate the few hotspots of deforestation by increasing the overall deforestation rate.

Interestingly, both models under-predicted deforestation in several regions. These hotspots correspond to very large development projects that occurred between 1990 and 2000. For example, the large Southern Village development to the South of Chapel Hill was not accurately predicted by either model. These unexplained hotspots might have been poorly predicted by our models for a couple of reasons. First, these development events are larger than the average scale of autocorrelation in the process of deforestation, and really are the result of a single unified decision by a planning board. It unlikely any model could predict the exact

spatial location of a particular development event, even one as spatially huge as these hotspots. Secondly, the failure of both models could also be attributed to the lack of the relevant explanatory variables in the training dataset. As these large development decisions are singular events, they will generally be more idiosyncratic than a collection of independent smaller development projects, and may require different explanatory variables than smaller deforestation events.

Our results suggest that a relatively simple, spatially heterogeneous model can outperform a relatively simple, spatially homogeneous model. While our results strictly only speak to a comparison between GLM and CART models, we believe that the general principle will hold in many cases of land-use change modeling. We recognize that spatially heterogeneous models of land-cover change will not (and should not) replace other modeling techniques that have a proven ability to address other issues in landscape modeling, especially models that aim to be more mechanistic. Nevertheless, it appears that, for our case study, accounting for spatial heterogeneity in the rules of development is important, and could arguably be more important than accounting for other variation in the agents of development (i.e., the different preference functions sometimes used in agent-based models). We believe, therefore, that models that explicitly quantify and explore spatial heterogeneity in the rules of land-cover and land-use change are a useful supplement to more traditional models.

5. Conclusion

For our study area, we found that our simple, spatially heterogeneous model (CART) outperforms our simple, spatially homogeneous model (GLM). Examined from a non-spatial perspective, the CART model (15.2% misclassification rate) did significantly better than the GLM model (33.1% misclassification rate). In a spatial analysis of the residuals of both models, the CART model appears to do better, more accurately predicting hotspots of development and predicting the baseline proportion of pixels deforested more accurately. We stress that our results are only intended to be complementary to other approaches, and not to replace other approaches. Nevertheless, our results suggest that significant gains can be made in landscape change modeling if spatial heterogeneity is utilized, and highlight the overall importance of spatially heterogeneous processes in landscape change. Moreover, our results suggest that the consequence of ignoring spatial heterogeneity in the rules of land-use and land-cover change is a form of 'averaging' in the model output, where the probability of change is over-predicted in most locations and under-predicted in hotspots of development.

Acknowledgements

We wish to thank Joseph Sexton for comments on the project, and Pete Harrell and Patrick Halpin for invaluable GIS assistance. The entire Landscape Ecology Laboratory at Duke made helpful comments on a draft of this manuscript. Financial support was provided by NSF grant SBR-9817755, and by a NSF Predoctoral Fellowship to RIM.

References

Berry, B.J.L., Conkling, E.C., Ray, D.M., 1993. The Spatial Organization of Land Use. Prentice-Hall, New York.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69.

Chambers, J.M., Hastie, T.J., 1992. Statistical Models in S. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Christensen, N.L., Peet, R.K., 1984. Convergence during secondary forest succession. J. Ecol. 72, 25–36.

Dale, M.R.T., 1999. Spatial Pattern Analysis in Plant Ecology. Cambridge University Press, Cambridge.

De'Ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.

Fahrig, L., 2002. Effect of habitat fragmentation on the extinction threshold: a synthesis. Ecol. Appl. 12, 346–353.

Fletcher, D., Goss, E., 1993. Forecasting with neural networks—an application using bankruptcy data. Inform. Manage. 24, 159–167.

Harrison, S., Bruna, E., 1999. Habitat fragmentation and large-scale conservation: what do we know for sure? Ecography 22, 225–232.

Hogeweg, P., 1988. Cellular automata as a paradigm for ecological modeling. Appl. Math. Comp. 27, 81–100.

Houghton, R.A., 1994. The worldwide extent of land-use change. Bioscience 44, 305–313.

Jenerette, G.D., Wu, J.G., 2001. Analysis and simulation of land-use change in the central Arizona-Phoenix region, USA. Landscape Ecol. 16, 611–626.

Johnson, W.C., 1977. A mathematical model of forest succession and land use for the North Carolina Piedmont. Bull. Torr. Bot. Club 104, 334–346.

Laurance, W.F., Lovejoy, T.E., Vasconcelos, H.L., Bruna, E.M., Didham, R.K., Stouffer, P.C., Gascon, C., Bierregaard, R.O., Laurance, S.G., Sampaio, E., 2002. Ecosystem decay of Amazonian forest fragments: a 22-year investigation. Conserv. Biol. 16, 605–618.

Legendre, P., Fortin, M.J., 1989. Spatial pattern and ecological analysis. Vegetatio 80, 107–138.

Legendre, P., Legendre, L., 1998. Numerical Ecology, second ed. Elsevier Science, Amsterdam.

Mansfield, C., Pattanayak, S., McDow, W., McDonald, R.I., Halpin, P.N., 2003. Shades of green: measuring the value of urban forests in the housing market, RTI, Durham, NC.

McDonald, R.I., Urban, D.L., in preparation. Succession from space: identification of changes in species composition from remote sensing imagery.

Miller, D.M., Kaminsky, E.J., Rana, S., 1995. Neural-network classification of remote-sensing data. Comput. Geosci. 21, 377–386.

Moore, I.D., Lee, B.G., Davey, S.M., 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. Environ. Manage. 15, 59–71.

Oosting, H.J., 1942. An ecological analysis of the plant communities of piedmont, North Carolina. Am. Midland Nat. 28, 1–126.

Parker, D.C., Manson, S.M., Janssen, M.A., Hoffmann, M.J., Deadman, P., 2003. Multi-agent systems for the simulation of land-use and land-cover change: a review. Ann. Assoc. Am. Geograph. 93, 314–337.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Modell. 133, 225–245.

Sokal, R.R., Oden, N.L., 1978. Spatial autocorrelation in biology—methodology. Biol. J. Linnean Soc. 10, 199–228.

Sokal, R.R., Rohlf, F.J., 1981. Biometry, second ed. W.H. Freeman and Company, New York.

Song, C., Woodcock, C.E., Seto, K.C., Lenney, M.P., Macomber, S.A., 2001. Classification and change detection using Landsat TM data: when and how to correct atmospheric effects? Remote Sens. Environ. 75, 230–244.

Theobald, D.M., 2001. Land-use dynamics beyond the American urban fringes. Geograph. Rev. 91, 544–564.

Triangle J Council of Governments, 2003a. Research Triangle Region Population, 1950–2020. In: Triangle J Council of Governments.

Triangle J Council of Governments. 2003b. Triangle GreenPrint Outreach Report. In: Triangle J Council of Governments. Triangle Land Conservancy, Department of Environment and Resources, Raleigh, North Carolina.

Urban, D.L., Goslee, S.C., Pierce, K.B., Lookingbill, T.R., 2002. Extending community ecology to landscapes. Ecoscience 9, 200–212.

USDA, 2003. National Resources Inventory Assessment. Natural Resources Conservation Service, Washington, DC.

Vayssieres, M.P., Plant, R.E., Allen-Diaz, B.H., 2000. Classification trees: an alternative non-parametric approach for predicting species distributions. J. Veg. Sci. 11, 679–694.

Veldkamp, A., Lambin, E.F., 2001. Predicting land-use change. Agric. Ecosys. Environ. 85, 1–6.

Veldkamp, A., Verburg, P.H., Kok, K., de Koning, G.H.J., Priess, J., Bergsma, A.R., 2001. The need for scale sensitive approaches in spatially explicit land use change modelling. Environ. Model. Assess. 6, 111–121.

von Neumann, J., 1966. Theory of Self-Reproducing Automata. University of Illinois Press.

Waisanen, P.J., Bliss, N.B., 1997. Changes in population and agricultural land in conterminous United States counties 1790 to 1997. Global Biogeochem. Cycles 16.

Wilkie, D.S., Finn, J.T., 1996. Remote Sensing Imagery for Natural Resources Monitoring. Columbia University Press, New York.

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

Yang, X.J., 2002. Satellite monitoring of urban spatial growth in the Atlanta metropolitan area. Photogrammetr. Eng. Remote Sens. 68, 725–734.