Landscape Ecol (2007) 22:419–430
DOI 10.1007/s10980-006-9037-7

RESEARCH PAPER

Prediction of multinomial probability of land use change using a bisection decomposition and logistic regression

Shoufan Fang · George Z. Gertner · Alan B. Anderson

Received: 25 January 2006 / Accepted: 17 August 2006 / Published online: 12 December 2006

Abstract Land use change is an important research area in landscape ecology and urban development. Prediction of land use change (urban development) provides critical information for making the right policies and management plans in order to maintain and improve ecosystem and city functions. Logistic regression is a widely used method to predict binomial probabilities of land use change when just two responses (change and no-change) are considered. However, in practice, more than two types of change are encountered and multinomial probabilities are therefore needed. The existing methods for predicting multinomial probabilities have limits in building multinomial probability models and are often based on improper assumptions. This is due to the lack of proper methodology and inadequate software. In this study, a procedure has been developed for building models to predict the multinomial probabilities of land use change and urban development. The foundation of this procedure consists of a special bisection decomposition system for the decomposition of multiple-class systems to bi-class systems, conditional probability inference, and logistic regression for binomial probability models. A case study of urban development has been conducted to evaluate this procedure. The evaluation results demonstrated that different samples and bisection decomposition systems led to very similar quality and performance in the developed multinomial probability models, which indicates the high stability of the proposed procedure for this case study.

Keywords Bisection decomposition · Conditional probability · Land use · Logistic regression · Multinomial probability · Urban development

Introduction

Land use change and urban development are two areas of research that attract broad attention, since both can produce significant ecological impacts to the environment. However, urban development is a special type of land use change: the conversion of mainly agricultural and forested lands to residential, commercial-industrial, and recreational in cities and along their edges. The special driving forces and radiating impact of

S. Fang · G. Z. Gertner (✉)
Department of Natural Resources and Environmental Sciences, University of Illinois at Urbana-Champaign, W503 Turner Hall, 1102 S. Goodwin Avenue, Urbana, IL 61801, USA
e-mail: gertner@uiuc.edu

A. B. Anderson
US Army Corps of Engineers, Construction Engineering Research Laboratory (CERL), P.O. Box 9005, Champaign, IL 61822, USA

urban development distinguish it from other kinds of land use change. The major driving forces of urban development include cultural, social, and economic factors; and make the processes of urban development very complex (Cheng and Masser 2003; Fang et al. 2005; Gimblett et al. 2001; Ligtenberg et al. 2001; Rusk 1995; Weber 2003). Although the proportion of urbanization is small compared to the total land use change on the earth's land surface area (Grübler 1994), urbanization can cause very large changes in surrounding environmental conditions, more so than other land use changes (Folke et al. 1997; Heilig 1994; Lambin et al. 2001). Even though the two areas can be separated based on the above reasons, both land use change and urban development researchers eventually study the conversion of land use. Therefore, they can share techniques such as those which estimate the conversion probability and the mapping of the conversion. Logistic regression is one of the techniques shared by land use change and urban development researchers.

Logistic regression is a common method to build models for predicting the probabilities of categorical variables (responses or events) based on numerical (continuous and discrete) and categorical variables (Agresti 2002; Hosmer and Lemeshow 2000; McCullagh 1980). It is also widely used in predicting the conversion probabilities of both land use change (Geoghegan et al. 2001; Serneels and Lambin 2001; Verburg et al. 2002) and urban development (Cheng and Masser 2003; Fang et al. 2005; Wu 2002). In studies of land use and urban development, usually logistic regression is used to fit a probability model based on sampled data. The conversion probability of land use change and urbanization in a study area thus can be predicted using the fitted logistic models based on the attribute maps. The resulting probability maps can be used to indicate the ‘‘hot spots’’ where the highest probability for change will occur over a certain duration (Fang et al. 2005).

Currently, logistic regression has been mainly used in predicting binomial probability of two responses (‘‘yes’’ or ‘‘no’’ in general) of a dependent categorical variable (event) in studies of land use change and urban development (for example, Geoghegan et al. 2001; Verburg et al. 2002; Wu 2002). Often the conversion of land use change and urbanization possesses a multinomial distribution (Turner et al. 1996; Wear et al. 1998). The challenge of multinomial probability prediction is that the sum of the predicted probabilities of all responses should be equal to one. Usually, multinomial probabilities are predicted using a set of logistic models, whose dependent variables could be fractions of probabilities of paired responses [$\log(P_i / P_j) = A'X$, $i \neq j$, where A and X are vectors of coefficients and independent variables, respectively], and the models are adjusted to make the multinomial probabilities consistent (i.e., the sum of probabilities equal to one) after their coefficients are estimated (Allison 1999; Chomitz and Gray 1996; McCullagh 1980).

The estimation of multinomial probability has its limits. First, the adjustment to reach consistency of the multinomial probabilities is based on the assumption of independence among multiple responses (Chomitz and Gray 1996). This assumption does not hold for most of situations. Furthermore, when each response has its own unique explanatory variables (Deal et al. 2002), the pair-response-based logistic models and the resulting models (Allison 1999) may make no logical sense and can cause confusion, since the resulting model for a specific response contains explanatory variables which are not defined for that response. Another problem for this situation is over-parameterization, i.e., too many explanatory variables may be included in a model, even though some are not significant in terms of prediction.

Finally, in model development and calibration, there is a need for screening techniques, i.e., to select significant independent variables of the logistic models from a number of candidate variables. Although screening techniques are available for binomial logistic regression, they are usually not available for multinomial probability estimation in standard statistical packages (for instance, SAS® and SUDAAN®). Therefore, there is a need for methods that estimate multinomial probability without the theoretical and practical limitations as given above.

Since the estimation of a logistic model for binomial probability is well-established and any multi-element system (set) could be divided into two subsystems (subsets) at a time, a bisection system is promising to find a solution for the prediction of the multinomial probabilities of land use change. A bisection system has been widely used in developing regression trees to improve the quality of empirical models (Alexander and Grimshaw 1996; Chaudhuri et al. 1994; De'Ath and Fabricius 2000; Loh and Shih 1997); for machine learning in classification (Perlich et al. 2003); and for classification of vegetation patterns and land use conversion changes (not prediction of future change) (Lawrence and Wright 2001; McDonald and Urban 2006; Rogan et al. 2003; Taverna et al. 2005). In tree regression, although the dependent variables of the models can be numerical or categorical and individual probabilities of categories can be computed for classification, the consistency of multinomial probabilities (i.e., their sum equals to 1.0) is never considered. In addition, tree regression does not divide data according to the (categorical) dependent variables of the models in the estimation of multinomial probabilities, but instead by the independent variables.

The objective of this study is to develop a consistency-constrained procedure based on a bisection system for prediction of the multinomial probabilities of effect-specified land use conversion. The bisection system is based on the dependent variables of the probability models. The procedure is suitable for the general properties of land use change and urbanization, but can also be used for other landscape systems where multinomial probabilities are needed. It has been developed based on conditional probability inference and utilizes existing logistic regression statistical software. A case study will be used to evaluate this procedure and demonstrate its use in the prediction of the multiple probability maps.

Procedure development

When an event has two types of responses (such as, "yes" and "no"), it has a binomial distribution and the probability of one type of response can be calculated from that of the other one, i.e., $P_2 = 1 - P_1$. Logistic regression has been developed for modeling the probability of binomial distributions and most major statistical software packages have procedures for this purpose. When there are more than two types of responses (say, k responses) for an event, it has multinomial distribution and the sum of the probabilities for all types of responses should be equal to one, i.e., $\sum_{r=1}^{k} P_r = 1$, where P_r is the probability of the rth response.

Suppose a special bisection decomposition system as shown in Fig. 1 is constructed. In such a decomposition system, k types of responses of an event are decomposed into $k - 1$ decomposition levels, each decomposition level being a binomial structure. At the first level there are exactly two classes, one class containing a single response type and the other class containing all remaining $k - 1$ response types. For the second level, remove the data for the single response type used for the first level. Then the class containing $k - 1$ responses in first level is regrouped to have exactly two classes: one class containing a single response type and the other class containing all

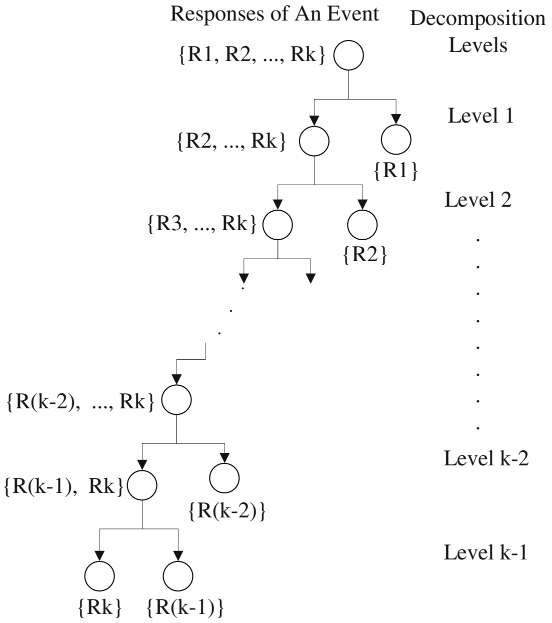


Fig. 1 Special bisection decomposition system designed for predicting multinomial probability of events which have more than two types of responses using logistic probability models and conditional probability inference

remaining $k-2$ response types. This general form of decomposition is continued for $k-1$ levels.

Therefore, at each decomposition level, a separate logistic regression for binomial distribution can be applied to estimate the (conditional) probabilities of the separated single response types at each level. The logistic model of the first decomposition level estimates the probability of the single response type, and the logistic models beyond the first level estimate the conditional probabilities of the corresponding single response types at each level. Based on this decomposition system (Fig. 1), a total of k-1 logistic models are needed to estimate the probabilities of all k types of responses of an event. In logistic model development, given a data set, the first model uses the entire data set in logistic regression. Thus the prediction of the established model is the probability of the first single response type. The second model based on the second level should use the sub data set, which excludes the data whose single response type belongs to the first level single response. Therefore, the predicted probabilities of the second model are the conditional probabilities of the second level single response type given the probability of the first level single response type. Following this pattern, the last ($k-1$ level) model uses only the sub data set which contains only data whose responses belong to the $k-1$ and kth types, and predicts the conditional probabilities of the $(k-1)$th type of response given the (conditional) probabilities of the 1st, 2nd, ..., and $(k-2)$th types of responses. After the probability model for each level is developed and the (conditional) probabilities of the first $k-1$ types of responses are predicted, the probabilities of all responses of the event can be computed according to the properties of binomial distribution and conditional probability:

$$\begin{cases} P_i = P'_i & i = 1 \\ P_i = P'_i \cdot \prod_{j=1}^{\mathrm{i}-1} (1 - P'_j) & i = 2, \cdots, k-1 \\ P_k = 1 - \sum_{r=1}^{k-1} P_r & i = k \end{cases} \tag{1}$$

where P' is the (conditional) probability predicted using the logistic models, P is the probability of response types, and subscripts (i, j, r, and k) indicate the types of the responses.

By using the bisection decomposition system, the logistic regression at each level can have specific explanatory variables, since there is only a single response type at each level. Therefore, there is no confusion with the explanatory variables and the probability models, which can be very meaningful. When a specific set of explanatory variables are not known or are not well defined for a particular response or responses, logistic regression with screening options (for example, forward, backward, and stepwise selection) can be used to develop the logistic models at each level.

Procedure evaluation

A case study of urbanization over a 10-year period was conducted in order to evaluate the properties of the procedure proposed in the previous section. In the case study, four types of urbanization land use conversion were considered.

Study area

The study area includes the cities of Columbus, Georgia, and Phenix City, Alabama; and their adjacent area. The geographical location of the study area is between latitude 32°25′00″–32°44′55″ N and longitude 84°34′18″–85°04′52″W. The land use of the study area in 1980 and 1990 is displayed in Fig. 2A and 2B. Outside the cities, the predominant land use category is forested. Development during 1980 and 1990 within the study area was mainly concentrated inside the cities and their suburbs (Fig. 2C). There are three categories (responses) of urban development: "Residential" (RES), "Commercial-Industrial" (CI), and "Open Space" (OS, urban/recreational grassy area). Adding the response of "No Change" (NCH) in development, there were a total four types of development (responses) considered in this study.

Among the total number of 493,600 pixels (177,696 ha) in the study area, 23,810 pixels (8,571 ha) had no data in the 1980 land use map

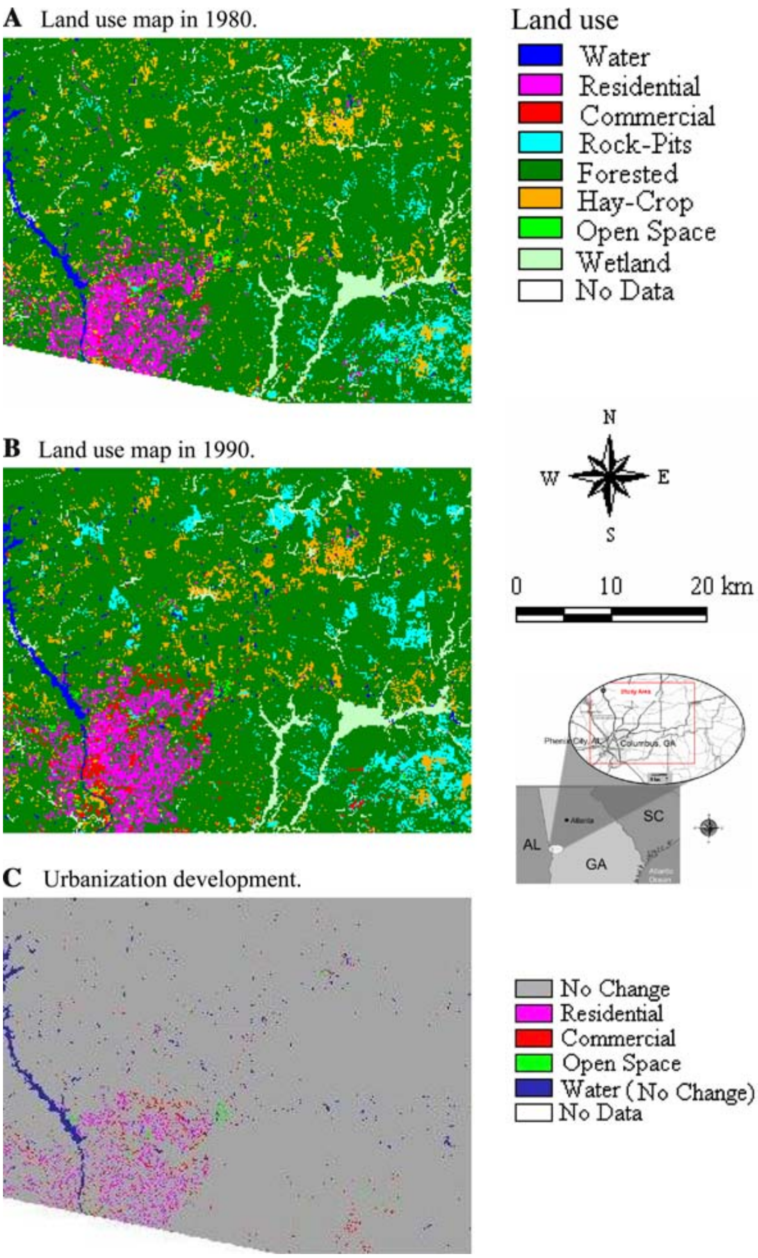


Fig. 2 Land use maps (**A** and **B**) of the study area (including cities of Columbus, Georgia (GA) and Phenix City, Alabama (AL), USA) and urbanization development during 1980 and 1990 (**C**). The land use categories come from the website of USGS (http://edcwww.cr.usgs.gov/programs/lccp/classes.html). Open Space represents urban/recreational grassy areas

(at the south-east corner) (Fig. 2B). Therefore, that corner was eliminated from the analysis.

Materials

The US Construction Engineering Research Laboratory (USACERL) (Lozar et al. 2003) provided land use maps. The pixel size of the land use maps was 60 × 60 m^2. The first three types of urban development (RES, CI, and OS) are modeled and predicted using the explanatory variables generated by ten factors. Those factors are City (X_1), County Road (X_2), Slope (X_3), Forest (X_4), Ramp (X_5), Road Intersection (RI, X_6), State Highway (SH, X_7), Water (X_8), Utilities (X_9), and the number of immediate neighbors (Neighbor, $X_1$0). They were defined by the LEAM (the Land Use Evolution and Impact Assessment Model, see URL "http://www.leam.uiuc.edu/") research group and their definitions are listed in Table A1 in the Appendix (Deal et al. 2002).

The LEAM group defined scores based on the ten factors that potentially could lead to the conversion to RES, CI, and OS. Those scores served as the direct explanatory variables for the logistic models. At a particular location (pixel), a factor could have different scores for different categories of conversion (Deal et al. 2002). For example, the factor Forest (X_4) for a pixel could have a higher score for RES and a lower score for CI, or visa-versa. The exceptions were the two factors, Utility (X_9) and Neighbor (X_{10}), which had the same scores for different categories at the same location. Therefore, there were a total of 26 (Neighbor + Utility + (8 factors × 3 categories)) unique scores across the categories, each category had ten scores (explanatory variables).

The score maps were also provided by the LEAM group. The original resolution of the score maps was 30×30 m^2. They were scaled up to 60×60 m^2 based on the average value of the merged pixels to correspond to the pixel size of land use maps. The score maps were used in two ways: (1) pixels from these score maps were sampled to calibrate the logistic probability models using the bisection system; and (2) using all pixels from the score maps as model inputs, the calibrated logistic probability models were used to predict the probabilities of the four types of urban development for the entire study area.

Methodology assessment

Three categories of urban development, RES, CI, and OS, were explicitly modeled in this study using the corresponding scores. The probability of the last category, NCH, was not explicitly modeled, since there were no scores defined for it and its probability can be calculated as the complement of the first three categories of urban development (see lower part of Eq. 1). For coefficient estimation, three independent samples were randomly drawn from the historical land use and score maps. For each of the random samples, 3% (14100 pixels) of the total pixels in the study area were sampled.

The order in which the response variables were considered in the bisection decomposition systems was evaluated. Three separate bisection decomposition system were constructed. Table 1 lists the decomposition levels and the order in which the response variables were considered in the bisections. According to the procedure, for each of the random samples, three logistic models (as one set) were needed to predict the multinomial probabilities of the categories of urban development for each of the decomposition systems. Thus, three sets of (nine) logistic models were developed with each of the three random samples, resulting in a total of nine sets of (27) logistic models.

Table 1 The order in which the categories were considered for three separate bisections systems

Decomposition Level	Decomposition System (Order)		
	1	2	3
1	CI	OS	RES
2	RES	CI	OS
3	OS	RES	CI

Categories CI, OS, and RES represent land use converted into Commercial-Industrial, Open Space, and Residential uses, respectively

The initial independent variables of the probability models included the scores of the ten explanatory factors and their cross product terms that are listed in Table 2. A stepwise logistic regression was used for selecting significant independent variables for the models and for estimating their coefficients using SAS® (PROCEDURE "LOGISTIC" with the model option "SELECTION=STEPWISE" and default significance level ($\alpha = 0.05$)). The quality of each model was assessed using (pseudo) R-square and concordance, which is a summary measure of association based on the number of pairs of observations whose predicted probability and response are consistent.

In order to evaluate the performance of the models, Relative Operating Characteristic (ROC) was used. ROC is an index used to measure the accuracy of predicted probability compared to the actual condition (Swets 1988). Pontius and Schneider (2001) interpreted ROC in measuring the quality of the prediction of spatial ecological changes, and provided the details for drawing empirical ROC curves and computing their

Table 2 The initial independent variables (scores generated from the factors and the combinations of scores)

Factor*	Code	Cross product	Code	Cross product	Code	Cross product	Code
City	X_1	$X_1 \times X_2$	U_1	$X_2 \times X_8$	U_{13}	$X_5 \times X_8$	U_{25}
County Road (Road)	X_2	$X_1 \times X_3$	U_2	$X_3 \times X_4$	U_{14}	$X_6 \times X_7$	U_{26}
Forest	X_3	$X_1 \times X_4$	U_3	$X_3 \times X_5$	U_{15}	$X_6 \times X_8$	U_{27}
Slope	X_4	$X_1 \times X_5$	U_4	$X_3 \times X_6$	U_{16}	$X_7 \times X_8$	U_{28}
Ramp	X_5	$X_1 \times X_6$	U_5	$X_3 \times X_7$	U_{17}	$X_{10} \times X_1$	U_{29}
Road Intersection (RI)	X_6	$X_1 \times X_7$	U_6	$X_3 \times X_8$	U_{18}	$X_{10} \times X_2$	U_{30}
State Highway (SH)	X_7	$X_1 \times X_8$	U_7	$X_4 \times X_5$	U_{19}	$X_{10} \times X_3$	U_{31}
Water	X_8	$X_2 \times X_3$	U_8	$X_4 \times X_6$	U_{20}	$X_{10} \times X_4$	U_{32}
Utilities	X_9	$X_2 \times X_4$	U_9	$X_4 \times X_7$	U_{21}	$X_{10} \times X_5$	U_{33}
Neighbor	X_{10}	$X_2 \times X_5$	U_{10}	$X_4 \times X_8$	U_{22}	$X_{10} \times X_6$	U_{34}
		$X_2 \times X_6$	U_{11}	$X_5 \times X_6$	U_{23}	$X_{10} \times X_7$	U_{35}
		$X_2 \times X_7$	U_{12}	$X_5 \times X_7$	U_{24}	$X_{10} \times X_8$	U_{36}

* The descriptions of the factors are listed in Table A1 in the Appendix

values. Fang et al. (2005) used ROC in comparison of the performance of urbanization models. In this study, both predicted probability and historical land use maps were used to draw Relative Operating Characteristic (ROC) curves and to compute the ROC value for each category. The details on how this was done can be found in Fang et al. (2005) and Pontius and Schneider (2001). In drawing the ROC curves and computing their values, 11 probability scenarios were adapted: 0.0, 0.1, ..., 0.9, and 1.0.

The quality and performance of the logistic probability models were statistically analyzed using analysis of variance (ANOVA). There were three random samples (Sample), three orders (Order) in which the response variables were considered in the decomposition, and three response variables (Category). Therefore, a three-way ANOVA was used to identify the effects of these variables on the quality measures of (pseudo) *R*-squares, percent concordance, and ROC values used for assessing performance of the logistic regressions. In the ANOVA, the null hypotheses were that the different levels of Sample, Order, and Category individually had the same effect on the three quality measures. The ANOVA's were conducted using SAS© PROCEDURE "ANOVA".

Results and analysis

From 1980 to 1990, among the 469,790 pixels (169,124 ha) considered in the study area (excluding the south-east corner as shown in Fig. 2C), 458,139 pixels (164,930 ha) had no change. The major type of urbanization was the conversion to Residential (RES), which consisted of 7,276 pixels (2,619 ha). The number of pixels converted to Commercial-Industrial (CI) were 3,130 (1,127 ha). Open Space (OS) was the least converted, only 1,245 pixels (448 ha) changed. The models calibrated with sampled data from the study area reflected the pattern of urbanization.

Table 3 lists the descriptive statistics of the quality measures of the (conditional) probability models, which were fitted based on the three selected bisection systems and three random samples. The *R*-squares of all models were between 0.72 and 0.75 . The mean *R*-squares of the models for the three categories with all samples were also within this interval (Table 3). The standard error

Table 3 The means and standard errors (SE) of the *R*-squares and percent concordance of the probability models based on the three random samples

Modeled category	Random sample					
	1		2		3	
	Mean	SE	Mean	SE	Mean	SE
R-squares						
RES	0.729	0.00040	0.729	0.00049	0.729	0.00044
CI	0.736	0.00013	0.734	0.00017	0.727	0.00009
OS	0.746	0.00007	0.744	0.00003	0.743	0.00009
Percent concordance						
RES	96.60	0.1000	96.00	0.1000	97.33	0.0667
CI	89.73	0.5841	87.67	0.3667	36.67	0.8333
OS	89.83	0.8172	90.03	1.1681	91.97	0.3283

of the *R*-squares of any individual sample was very small (<0.001), which indicates very small variation when different samples were used. The concordance was also larger than 87% for all models, except for CI with the third random sample. The standard errors of concordance for all models were smaller than 1.2 (Table 3). Since the means of the *R*-squares and concordance based on variables Sample and Category had relatively large variation compared to the overall standard errors, the corresponding *F*-values were large, thus leading to small *P*-values (<0.03) for the variables Sample and Category (Table 4). Therefore, the results of ANOVA led to rejecting the null hypotheses for the effect of these two variables at a 0.05 significance level. However, the variable Order had small *F*-values (<0.11) and very large *P*-values (> 0.9) (Table 4) for both the *R*-squares and the concordance from modeling. Thus, the null hypotheses for the effect of Order could not be rejected at even a 0.50 significance level.

The means and standard errors of the ROC values, computed from the predicted probability maps and historical land use maps, are listed in Table 5. The means of the ROC values for the category RES were the highest in comparison to those of the other categories. The logistic models for the category CI provided the lowest means of the ROC values and their highest mean was still smaller than 0.53, which was very close to the base ROC value of 0.50 . The higher ROC values of RES indicated that RES probability models could explain more land use conversion than CI and OS models based on their explanatory variables. The ROC curves visually demonstrate the performance of the logistic models built using the first random sample according to the locations of the categories with the third decomposition system (Table 1 and Fig. 3). The ANOVA results based on ROC values showed that both the variables Sample and Order had very small *F*-values (0.37 and 0.08) and large *P*-values (0.6983 and 0.9219). Therefore, the null hypotheses for the effect of these two variables could not be rejected at a 0.05 significance level. Thus, the difference of ROC values based on different sample and order was not significant. However, the variable Category had a very large *F*-value (693.74) and a very small *P*-value (<0.0001), which led to a rejection of the corresponding hypothesis at any significant level larger than 0.0001. A pairwise multiple comparison test was performed and the difference in the mean ROC values for each of the categories were all found to be significantly different from each other at the 0.05 level (Table 5). The ANOVA model of ROC was highly significant (*P*-value <0.0001).

Table 4 ANOVA results of the *R*-squares and percent concordance of the probability models

Source	DF*	*R*-square		Percent concordance	
		F-value	*P*-value	*F*-value	*P*-value
Sample	2	11.3	0.0005	4.16	0.0308
Order	2	0.1	0.9088	0.01	0.9878
Category	2	186.42	<0.0001	8.16	0.0026
Overall ANOVA	6, 26*	65.94	<0.0001	4.11	0.0075

* DF = degree of freedom. For the overall ANOVA degrees of freedom, the first and second values are model and corrected total degrees of freedom, respectively

Table 5 The means and standard errors (SE) of ROC values computed based on the historical land use maps and the predicted probabilities according to the three random samples

Modeled Category	Random Sample					
	1		2		3	
	Mean	SE	Mean	SE	Mean	SE
RES	0.777	0.00208	0.768	0.00176	0.781	0.00737
CI	0.527	0.01035	0.510	0.00623	0.500	0.00000
OS	0.590	0.00240	0.608	0.00203	0.622	0.00939

Figure 4 displays the predicted probability maps of all categories using the probability models based on the first random sample and the last (third) bisection decomposition system given in Table 1. Comparing the probability maps with the actual change during 1980 and 1990 (see Figs. 2C and 4), the predicted maps of RES and OS (especially RES) provided a reasonable prediction of the spatial pattern of the development. The quality of probability maps was measured by the ROC values and curves (Fig. 3 and Table 5). As these figures show, the predicted CI probability map captured a small portion of the actual CI conversion. The probability of No Change (NCH)

Fig. 3 ROC graphs based on the historical land use maps and predicted multinomial probabilities of land use converted to Residential (**A**), Commercial-Industrial (**B**), and Open Space (**C**) between 1980 and 1990 in the study area. Models were built using the first random sample according to the location of the categories in the third decomposition system (Table 3)

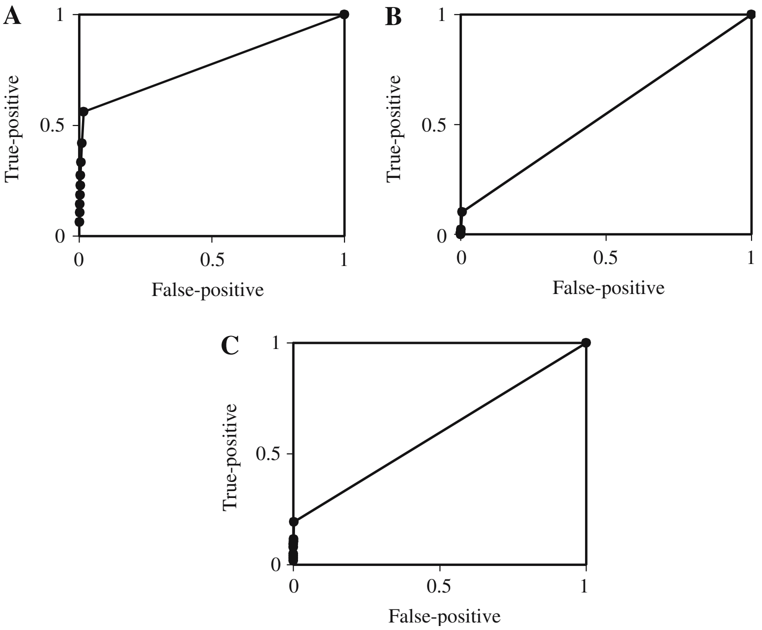
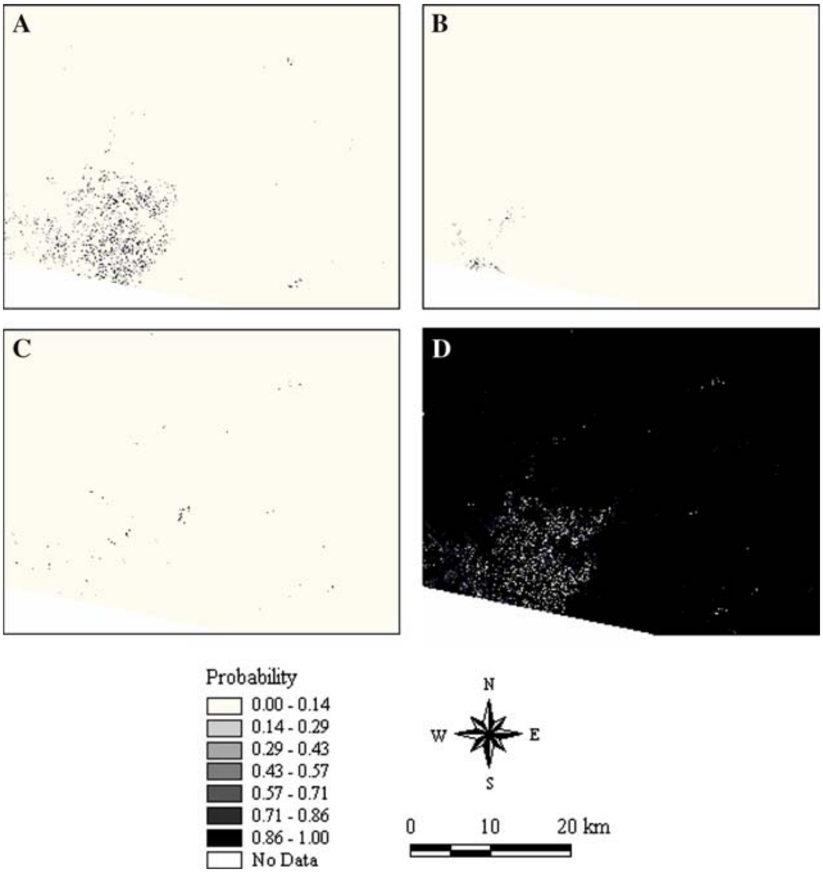


Fig. 4 The maps of the predicted probabilities of city development between 1980 and 1990. Maps **A** to **D** are the probabilities of land use converted to Residential, Commercial-industrial, Open Space, and no conversion (No Change), respectively



could be considered as a view of the probability of overall urban development. Comparing Figs. 2C and 4D, the major development during 1980 and 1990 was captured by lower probability of NCH in its probability map. The ROC value from the probability map of NCH was 0.851, which was considerably better than that of RES. This indicated that some CI or OS pixels which were not captured by their corresponding probability models were predicted to have higher probability of RES, or *vice versa*. This might have been caused by the similarity among the three categories.

Discussion and conclusion

The procedure presented for modeling multinomial probabilities has been developed based on conditional probability inference and a special bisection decomposition system. As long as such a decomposition system can be established, multinomial probability problems can be decomposed into binomial and conditional probability problems. Once decomposed, classical methods/techniques can be used for estimation. Theoretically, there was no assumption implied in the procedure, and a special bisection decomposition system could always be built for any multi-response event. The decomposition of multinomial probability into binomial probability made it much more convenient to use screening techniques with logistic regression and to structure specific models for specific responses and their corresponding explanatory factors.

In the evaluation of this procedure, the responses of interest had different numbers of observations in the samples. The largest number of responses, Residential (RES), had five times the number of observations as Open Space (OS). With this large difference, evaluation of this procedure showed that the impact of decomposition order to either the quality or performance of logistic models was not significant. It also implies that modeling uncertainty will impact the prediction of individual pixels, but not reduce the accuracy of predictions for the entire population (across entire case study area). This property of the procedure adds more flexibility in practice: researchers could decompose a multinomial system into a series of binomial systems according to their preference.

The performance of the probability models built with the procedure developed in this study was comparable to similar studies. The highest ROC mean of the probability maps of the modeled categories was RES with a value of 0.78, which was higher than that (0.72) of the RES probability map in Peoria, Illinois, USA (Fang et al. 2005). The comparison of the performance of the probability models built in different study areas showed that the procedure developed in this study would not cause difficulties in terms of model performance. Due to technical reasons mentioned in the Introduction, there is no comparison between this procedure and other estimation methods for multinomial probability.

The measures of model quality and performance were not consistent in this study. Concordance, which indicated the quality of probability models, had very different values when different random samples were used in model development. With the third random sample, logistic models of CI had concordance values less than half of that obtained from the first two samples. However, the *R*-squares and ROC values of all models for all categories across samples were very stable. The *R*-squares of the models of all categories were concentrated within a narrow interval (0.72 to 0.75), although statistically there was significant difference based on the ANOVA. The ROC values of different categories had more noticeable differences. Since ROC value was computed directly based on probabilities and for the entire study area, it was more reliable than *R*-square, which is computed using likelihood in logistic regression, and concordance based on ordinal variables. The consistency of ROC values for any one of the categories indicated that all three random samples represented the study area well.

Acknowledgements The authors would like to express their deep appreciation to the U.S. Army Corps of Engineers, Construction Engineering Research Laboratory (CERL) for their support; Professors Deal and Hannon at the University of Illinois; the Land Use Evolution and Impact Assessment Model (LEAM) research group at the University of Illinois for providing LEAM model and data; and Mr. Bruce MacAllister for generating the study location maps.

Appendix: Descriptions of the factors for predicting urban sprawl

The factors used in the LEAM (the Land Use Evolution and Impact Assessment Model, see website ‘‘http://www.leam.uiuc.edu/’’) model to predict urban sprawl have been defined by the LEAM research group. Table A1 lists the factors and their descriptions. In the LEAM model, functions have been developed to convert the original values of these factors into scores valued between zero and one. For details about how to measure these factors and convert them into scores, please contact the LEAM research group.

Table A1

Factor	Description
City	Weighted travel time to city centers
Road	Proximity to county roads
Forest	Proximity to forests
Slope	Steepness (degree) of the topographic character
Ramp	Travel time to ramps of limited-access highways
Road Intersection (RI)	Proximity to major road intersections
State Highway (SH)	Proximity to state highway
Water	Proximity to lakes and rivers
Utilities	Proximity to sewage, water supply, electricity, etc.
Growth Trend (GT)	Historical growth trend
Growth Booster (GT)	City development policy, such as zoning
Agricultural Protection (AP)	Policy to preserve farm land
Neighbor	Number of immediate house/building neighbors

References

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York

Allison PD (1999) Logistic regression using SAS® system: theory and application. SAS Institute Inc, Cary, NC

Alexander WP, Grimshaw SD (1996) Treed regression. J Comput Graphical Stat 5:156–175

Chaudhuri P, Wang MC, Loh WY, Roa R (1994) Piecewise-polynomial regression trees. Statistica Sinica 4(1):143–167

Cheng J, Masser I (2003) Urban growth pattern modeling: a case study of Wuhan City, PR China. Landscape Urban Planning 62:199–217

Chomitz KM, Gray DA (1996) Roads, land, markets, and deforestation: A spatial model applied to Belize. The World Bank Econ Rev 10(3):487–512

De’Ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81(11):3178–3192

Deal BM, Fournier DF, Timlin DM, Jenicek EM (2002) An assessment of encroachment mitigation techniques for army land. Technical Report, 19 Dec 2002, USACERL, Champaign, IL, Report Number ERDC/CERL TR-02–27

Fang S, Gertner GZ, Sun Z, Anderson AB (2005) The impact of interactions in spatial simulation of the dynamics of urban sprawl. Landscape Urban Planning 73:294–306

Folke C, Jansson A, Larsson J, Costanza R (1997) Ecosystem appropriation by cities. Ambio 26:167–172

Geoghegan J, Viller SC, Klepeis P, Mendoza PM, Ogneva-Himmerlberger Y, Chowdhury RR, Turner BL, Vance C (2001) Modeling tropical deforestration in the southern Yucatán peninsular region: comparing survey and satellite data. Agric Ecosyst Environ 85:25–46

Gimblett R, Daniel T, Cherry S, Meitner MJ (2001) The simulation and visualization of complex human-environment interactions. Landscape Urban Planning 54:63–78

Grübler A (1994) Changes in land use and land cover: a global perspective. In: Meyer WB, Turner BL (eds) Technology. University of Cambridge Press, Cambridge, pp. 287–328

Heilig GK (1994) Neglected dimensions of global land use change: reflections and data. Population Development Rev 20 (4):831–859

Hosmer DW, Lemeshow S (2000) Applied Logistic Regression, 2nd edn. Wiley, New York

Lambin EF, Turner BI, Geist HJ (2001) The causes of land-use and land-cover change: moving beyond the myths. Global Environ Change 11:261–269

Lawrence RL, Wright A (2001) Rule-based classification systems using classification and regression tree (CART) analysis. Photogrammertric Eng Remote Sensing 67(10):1137–1142

Ligtenberg A, Bregt AK, van Lammeren R (2001) Multi-actor-based land use modeling: spatial planning using agents. Landscape Urban Planning 56:21–33

Loh WY, Shih YS (1997) Split selection methods for classification trees. Statistica Sinica 7(4):815–840

Lozar RC, Ehlschlaeger CR, Cox J (2003) A Geographic Information System (GIS) and Imagery approach to historical urban growth trends around military installations. Technical Report, May 2003, US-ACERL, Champaign, IL, Report Number ERDC/CERL TR-03-9

McCullagh P (1980) Regression models for ordinal data. J Roy Stat Society B 42(2):109–142

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

Perlich C, Provost F, Simonoff JS (2003) Tree induction vs. logistic regression: a learning-curve analysis. J Machine Learning Res 4:211–255

Pontius RG Jr, Schneider LC (2001) Land-cover change model validation by an ROC method for the Ipswich Watershed, Massachusetts, USA. Agric Ecosyst Environ 85:239–248

Rogan J, Miller J, Stow D, Franklin J, Levien L, Fischer C (2003) Land-cover monitoring with classification tree using Landsat TM and ancillary data. Photogrammertric Eng Remote Sensing 69(7):793–804

Rusk D (1995) Cities without Suburbs, 2nd ed. The Woodrow Wilson Center Press, Washington DC, USA

Serneels S, Lambin EF (2001) Proximate causes of land use change in Narok district Kenya: a spatial statistical model. Agric Ecosyst Environ 85:65–81

Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293

Taverna K, Urban DL, McDonald RI (2005) Modeling Landscape vegetation pattern in response to historic land-use: a hypothesis-driven approach for the North Carolina Piedmont, USA. Landscape Ecol 20(6):689–702

Turner MG, Wear DN, Flamm RO (1996) Land ownership and land-cover change in the southern Appalachian highlands and the Olympic Peninsula. Ecol Appl 6:1150–1172

Verburg PH, Soepboer W, Veldkamp A, Limpiada R, Espaldon V, Mastura SSA (2002) Modeling the spatial dynamics of regional land use: The CLUE-S model. Environ Manage 30(3):391–405

Wear DN, Turner MG, Naiman RJ (1998) Land cover along an urban-rural gradient – implications for water quality. Ecol Appl 8:619–630

Weber C (2003) Interaction model application for urban planning. Landscape Urban Planning 63:49–60

Wu F (2002) Calibration of stochastic cellular automata: the application to rural-urban land conversions. Int J Geograph Information Sci 16(8):795–818