Modeling Earth Systems and Environment
https://doi.org/10.1007/s40808-019-00584-0

ORIGINAL ARTICLE

A weighted normalized likelihood procedure for empirical land change modeling

J. Ronald Eastman[1,2] · Stefano C. Crema[2] · Hannah R. Rush[2] · Kaixi Zhang[3]

Received: 13 December 2018 / Accepted: 16 March 2019

Abstract
A critical foundation for empirical land change modeling is the mapping of transition potentials—quantitative evaluations of the readiness or suitability of land to go through a transition. This paper presents a procedure based on empirically determined normalized likelihoods of transition. It shows that these normalized likelihoods equate to posterior probabilities if case–control sampling is carried out among historical instances of change and persistence. The posterior probabilities can then be aggregated at the pixel level across multiple covariates using linear opinion pooling where the pixel-specific weight for each covariate is determined locally by its ability to distinguish between the alternatives of change or persistence. Thus, covariates with spatially varying diagnostic ability can be productively incorporated. The resulting algorithm is shown to have a skill comparable to that of a multi-layer perceptron approach with the advantage of high efficiency and amenability to distributed processing in a cloud environment.

Keywords Land-use change · Land-cover change · Model development · Model evaluation · Empirical modeling · Machine learning

Introduction

Land cover transformation has been one of the most profound influences of humans on the environment, with major consequent impacts on species habitat, ecosystem services and climate through changes in albedo and the inputs of carbon to the atmosphere (Turner et al. 2007). As a consequence, there has been considerable effort to develop land change models for purposes of explanation and predictive evaluation.

In a recent review of land change modeling approaches, the National Research Council (2014) of the National Academies in the United States identified six primary approaches to land change modeling:

1. machine-learning and statistical approaches;
2. cellular automata approaches;
3. sector-based economic approaches;
4. spatially disaggregate economic approaches;
5. agent-based approaches; and
6. hybrid approaches that include multiple components of the above.

This paper is concerned with the first of these, which we more generally refer to as empirical land change models (ELCM). Empirical land change models use machine learning or statistical characterization to describe the relationship between historical land cover change and one or more explanatory variables, and then optionally use that model to project the change into the future (National Research Council 2014). Such a projection assumes stationarity and constitutes a business-as-usual (BAU) baseline that is, for example, useful for planning (Iizuka et al. 2017) and is used as a reference level for the evaluation of success in reducing emissions with reducing emissions for deforestation and forest degradation (REDD) climate change mitigation projects (Angelsen et al. 2011; Verified Carbon Standard 2013). Examples of ELCM include LCM (Eastman 2006), GEOMOD2 (Pontius et al. 2001), CLUE-S (Verburg et al.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40808-019-00584-0) contains supplementary material, which is available to authorized users.

✉ Stefano C. Crema
screma@clarku.edu

[1] Graduate School of Geography, Clark University, 950 Main St., Worcester, MA 01610, USA

[2] Clark Labs, Clark University, Worcester, MA 01610, USA

[3] ESRI, Redlands, CA 92373, USA

Published online: 03 April 2019

2002), LUCAS (Berry et al. 1996), DINAMICA EGO (Soares-Filho et al. 2002) and LTM (Pijanowski et al. 2002).

At the foundation of an ELCM is a procedure for empirically modeling the relationships between historical instances of land cover change and one or more covariates such as proximity to roads, slope, elevation, etc. Common approaches include logistic regression (LUCAS, CLUE-S, LCM) and multi-layer perceptron neural networks (LTM, LCM), but a variety of other approaches have been tried including empirical probabilities (GEOMOD2), k-nearest neighbor (LCM) and weights-of-evidence Bayesian modeling (DINAMICA EGO). The output is typically a set of maps with continuous values from 0 to 1 that express the suitability or readiness of the land to go through each of the transitions being modeled. As the numerical basis for this value can vary, we adopt the terminology of Pijanowski et al. (2002) and refer to it simply as a *transition potential*.

Given a set of transition potentials, ELCM can then focus on determining the quantity of expected change (such as through Markov chain analysis) and the specific allocation, based on the transition potentials, to produce a forecast land cover map (Eastman et al. 2005). However, the most intensive part of the analysis is typically the development of the transition potentials.

The focus of this paper is the development of an algorithm for the empirical modeling of transition potentials that would be suitable for deployment in a distributed cloud-computing environment (Zhang et al. 2010). Cloud-computing opens the prospect of a web-based implementation and the ability to deploy a cluster of computers, working in parallel, to develop the transition potentials. The technique developed thus has to permit nodes of a computer cluster to work on separate parts of the problem while minimizing the transmission of data between them.

Materials and methods

Logical design of the algorithm

One of the interesting features of the transition potential modeling procedure in GEOMOD2 (Pontius et al. 2001) is the simplicity of the approach. In that procedure, explanatory variables are binned into discrete categories and the relative frequency of change within bins is tabulated to produce a discrete probability density function. This kind of counting procedure is very simple and one can easily divide the task between multiple computer nodes with a simple communication of tallies upon the completion of each sub-area.

While this counting procedure is simple, and highly amenable to a cloud implementation, there is still quite a bit of work since each bin needs to be enumerated in its entirety. A more efficient procedure would be to calculate the relative frequencies of each bin within a sample of pixels that changed and a corresponding relative frequency of each bin occurrence within a sample of areas that were eligible to go through the transition, but did not (i.e., persistence). Given that historical change is a fairly rare event, this can reduce the amount work by several orders of magnitude (for example, from billions of counts to thousands). In this case, one would be tabulating likelihoods rather than probabilities. Using b to signify a bin, C to represent change and P to represent persistence, GEOMOD2 is calculating $p(C|b)$ while this logic would be calculating $p(b|C)$ and $p(b|P)$—the likelihood of the bin given change and persistence.

Another issue with the GEOMOD2 approach is the aggregation of evidence across explanatory variables. In its current implementation, GEOMOD2 uses a weighted average with the user being required to specify the weights. Thus GEOMOD2 is not completely empirical (since the weights are not empirically determined). While the averaging of posterior probabilities is a reasonable means of aggregation (Roberts 1965; Ranjan and Gneiting 2010; Clements and Harvey 2011), we require a completely empirical solution to meet the strict baseline requirements of business-as-usual models such as those used for REDD projects.

It is also important to note that GEOMOD2 assumes independence between the explanatory variables. This is the same assumption behind a *naïve Bayes* classifier. It greatly reduces the amount of computation by assuming that joint probabilities can be estimated from marginal probabilities. In empirical testing, Rish (2001) found that while the independence assumption of naïve Bayes is often unrealistic, it is remarkably robust since the classification decision is often correct despite inaccuracies in the computed probabilities. Given this, and the considerable reduction in computational complexity afforded by it, we have also made the assumption of independence, with the understanding that the user should strive to use uncorrelated variables.

Sampling

Land change modeling typically deals with rare events. For example, in frontier areas, the number of pixels that change from forest to agriculture is minute compared to the number that remain forest over any reasonable time frame. While this example is extreme, virtually all instances of land cover change are still rare events. In statistics and machine learning, it has been established that working with samples that are proportional to their rate of occurrence (i.e., random samples) is problematic. With logistic regression of rare events, random samples lead to underestimation of probabilities (King and Zheng 2001a, b). More significantly, with neural networks, the use of random samples with rare events leads to unbalanced networks (Weiss 2004) where

the model preferences skill at predicting the dominant class at the expense of the rare class. The solution is to use case-control sampling (King and Zheng 2001a, b) where, in this instance, a sample of pixels that changed during the historical period (the *cases*) is matched with an equal sized sample of pixels that were eligible to change during the historical period, but which did not (the *control*). Depending on the nature of the allocation stage in land change prediction, the estimated probabilities may need correction for the true prior probability. In case-control sampling, the prior probability is effectively 0.5. Most land change models use a greedy selection process where it is assumed that the pixels that should be changed in the prediction are those with the highest transition potential (Eastman et al. 2005). In cases such as this, no prior correction is needed since the transition potentials are monotonically related to those with prior correction.

In the procedure developed here we use case-control sampling. Typically, all pixels that experienced change during the historical period are selected with an equal sized random sample of pixels that persisted. Each of these groups is then split in half to yield a training and a validation set each for the cases of change and persistence.

Normalized likelihoods

To tabulate the histograms associated with each covariate, all quantitative variables are binned into 200 levels (an arbitrarily chosen number). Qualitative variables are unbinned and used as it is. Histogram tallies are then initialized to zero for each covariate. Training pixels of change and persistence are then visited and the histogram tallies are updated for each covariate. If a distributed computing architecture is involved, then the histograms from all nodes are collected and aggregated into the final histograms.

From the histogram tallies, normalized likelihoods are calculated as:

$$\mathrm{NL} = \frac{p(b|C)}{p(b|C) + p(b|P)}, \tag{1}$$

where $p(b|C)$ is the probability of the bin given change and $p(b|P)$ is the probability of the bin given persistence, each derived by calculating the proportion of pixels occurring within that bin relative to the size of the training sample for change or persistence.

The normalized likelihood has an interesting property. The values range from 0 to 1. Values greater than 0.5 indicate support for the hypothesis of change while values less than 0.5 support persistence. Values equal to 0.5 indicate an inability to support either hypothesis while values near 0 or 1 indicate a strong conviction of one hypothesis over the other. In the implementation here, Kernel Density Estimation (Rosenblatt 1956; Parzen 1962) is used to refine these normalized likelihoods by passing a 13-point Gaussian filter over the histogram of normalized likelihoods associated with each quantitative covariate. Using the histograms directly is subject to overfitting. Filtering helps to make the values more general. Qualitative covariates are not filtered.

A second interesting property of the normalized likelihood is that when the prior probability of change is 0.5 (as is the case with case–control sampling), the normalized likelihood is equal to the posterior probability of change. Bayes' Theorum states that:

$$p(C|b) = \frac{p(b|C) \times p(C)}{p(b)}, \tag{2}$$

where $p(b)$ is equal to the sum of the prior-weighted likelihoods:

$$p(b) = p(b|C) \times p(C) + p(b|P) \times p(P). \tag{3}$$

Now, since with case–control sampling $p(C) = p(P) = 0.5$:

$$p(C|b) = \frac{p(b|C) \times 0.5}{0.5 \times [p(b|C) + p(b|P)]}, \tag{4}$$

which reduces to the formula for the normalized likelihood:

$$p(C|b) = \frac{p(b|C)}{p(b|C) + p(b|P)}. \tag{5}$$

Thus normalized likelihoods under case–control sampling are in fact posterior probabilities.

Aggregation of evidence

Each pixel can now be evaluated for its transition potential. First, its value on each of the covariates is accessed, converted to a bin number and then converted to a normalized likelihood based on the entry in the histogram for that bin. As a result, for a single pixel, we have n estimates of normalized likelihood (i.e., posterior probability) where n is the number of covariates. The most common means of aggregating probability forecasts is through the use of a linear opinion pool (Ranjan and Gneiting 2010; Clements and Harvey 2011) where probabilities are combined through a weighted linear combination (Roberts 1965).

In this implementation we have calculated the weight of each estimate of normalized likelihood (posterior probability) as the absolute difference from 0.5:

$$w = \mathrm{abs}(\mathrm{NL} - 0.5). \tag{6}$$

For a specific pixel, if the normalized likelihood is near 0.5, then that variable is signaling that it has no strong empirical evidence to suggest that the pixel is a candidate for change or persistence. However, if the value is quite distant from 0.5 (i.e., near 0 or near 1) then the variable is indicating that it has a strong opinion regarding the alternative hypothesis. Accordingly, it is given a higher weight.

With this procedure, weights are calculated separately at each pixel, and thus no universal weight applies to each covariate. This is unusual, but a distinct advantage. For example, consider a map of protected areas with two classes: protected versus unprotected. Assume also that the amount of protected land is small compared to unprotected land. It would be expected that it would have tabulated very few cases of change and many instances of persistence in the protected areas. Thus the normalized likelihood would be close to 0. However, in the unprotected areas, it would be expected to find roughly equal frequencies of change and persistence. Thus the normalized likelihood would be expected to be near 0.5. In the protected areas, then, it has a strong opinion about the probability of change while in the unprotected areas it has little basis for an opinion, and thus a low weight. Since most of the layer is occupied by unprotected areas, it would seem like, overall, this covariate has low predictive ability. However, this is not the case. It has strong predictive ability, but only in limited areas. This is a very different approach to the treatment of variables from something like regression where a universal weight (the slope coefficient) is applied to all pixels in a covariate. Here, the weights will vary across space, conditional on the bin each pixel belongs to. Thus, it is expected that the results of this approach will differ from approaches such as logistic regression or multi-layer perceptron.

The linear opinion pool procedure used is thus:

$$\overline{\mathrm{NL}} = \frac{\sum_1^n (w_i \, \mathrm{NL}_i)}{\sum_1^n w_i}, \tag{7}$$

with weights calculated as in Eq. 6.

Logistic transformation

The normalized likelihood and its aggregate over covariates as in Eq. 7 are effectively linear classifiers with values greater than 0.5 representing change and values less than 0.5 representing persistence. As discussed in Flach (2012, pp. 219–224) and Jordan (1995), normalized likelihoods of covariates supporting change would be expected to follow a normal distribution as would the distribution of those supporting persistence. In these cases, then, the posterior probability of change will be a logistic function. Indeed, comparison of the output based on Eq. 7 with the activation level output of a multi-layer perceptron in the TerrSet geospatial monitoring and modeling system (Eastman 2014) indicated that to maximize the relationship between the two requires a logistic transformation of Eq. 7 as follows:

$$T = \frac{1}{1 + e^{-\left(\left(\overline{\mathrm{NL}} - 0.5\right) \times \beta\right)}}, \tag{8}$$

where the location parameter 0.5 is the classification threshold between change and persistence and β is a scaling factor that controls the steepness of the curve.

Determining the value of β is known as calibration (Flach 2012, p. 220). Although not strictly needed if the transition potential is only to be used in a relative sense (such as a greedy allocation), for cases where the absolute value is of interest (for example, as a characterization of vulnerability to change) calibration will be necessary. The procedure we have chosen is a modified form of Platt Calibration (Platt 2000) in which the location parameter is fixed at 0.5 and the scaling parameter β is empirically fit to the raw normalized likelihoods computed from the training data using Platt's sigmoidal curve fitting algorithm. Platt (2000) provides pseudo-code for the algorithm which we modified by forcing the intercept of the linear equation for estimating the sigmoid exponent to 0, thereby ensuring that the resulting transition potential is centered at 0.5.

Evaluation

To evaluate this new procedure, a comparison was done with the results of a multi-layer perceptron (MLP) neural network. MLP neural networks have received considerable interest because of their ability to model complex non-linear relationships and their non-parametric character (Tso and Mather 2001; Lin et al. 2011). Furthermore, in testing a wide variety of approaches to the modeling of transition potentials, Eastman et al. (2005) and Lin et al. (2011) found MLP to be superior to other techniques tested in the skill of the prediction. The research questions guiding the evaluation were thus:

1. Is the skill of the WNL procedure comparable to that of a multi-layer perceptron in a test over multiple transitions?
2. Does the WNL procedure offer a substantial improvement in speed upon parallelization?

A multi-layer perceptron neural network (Fig. 1), in this context, is a network that connects each of the inputs (the nine explanatory variables) with each of the outputs (the degree of support for change or persistence) through an intermediary *hidden* layer of analogues to neurons, called *perceptrons*. Each perceptron in the hidden layer is connected to each of the inputs, and computes a weighted linear combination of the inputs which it propagates to each of the outputs after applying a logistic transformation to the sum (which forces the output to a 0–1 range). The output neurons similarly compute a weighted linear combination and logistic transformation of the hidden layer inputs. Initially, all connections are initialized with random weights from − 1 to + 1. MLP then iteratively learns how to adjust

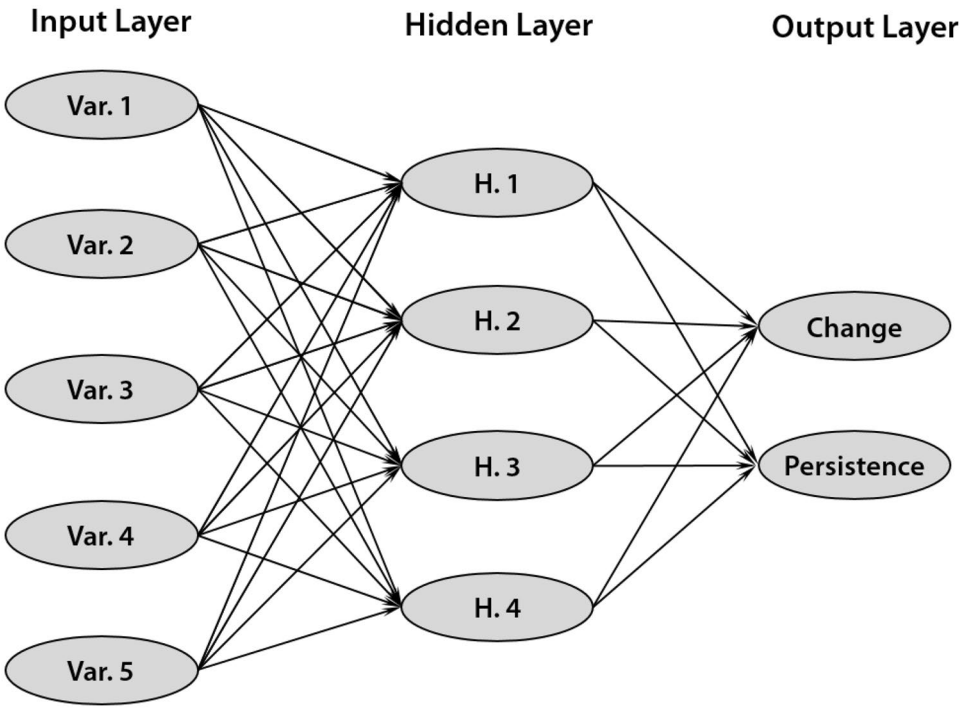

Fig. 1 The architecture of a multi-layer perceptron neural network. In this illustration, only five input neurons are shown for simplicity. However, in the tests undertaken, there were nine neurons representing the explanatory variables (as listed in Table 1). In a land change modeling context, the output for change is the transition potential

the weights based on a set of training data with examples of pixels that historically went through the transition being modeled (change pixels) and an equal sized set of pixels that were eligible to transition but did not (persistence pixels). After each small adjustment in the learning phase, it then tests itself on an independent sample of validation pixels (again with instances of change and persistence). Ideally, it should have an output near 1.0 for the correct outcome and a value near 0.0 for the wrong outcome. After many thousands of iterations, the modeling process is finished when acceptable skill on the validation data has been achieved. MLP is then ready to be used (with the connection weights determined in training) to make a prediction. The output for the change perceptron is then used as the transition potential.

The software used for the comparison was the Land Change Modeler (LCM) in the TerrSet geospatial monitoring and modeling system (Eastman 2014). A case study was conducted using the USGS National Land Cover Database (NLCD) 30-m land cover layers for 2001 and 2011 for the US State of Massachusetts. All land cover transitions between 2001 and 2011 in Massachusetts were modeled with both WNL and MLP, with the exception of transitions less than 2 km^2 in total area (for reasons of sample size) as well as transitions from open water to any other land cover, any land cover (except for woody wetlands and emergent herbaceous wetlands) to open water, deciduous forest to evergreen forest, and evergreen forest to deciduous forest. The latter two were omitted because they were deemed to be unlikely over a 10-year period and thus most likely arose from classification errors. In addition, a small area of 40 km^2 in the southwest of the state (less than 0.2% of the total area) was masked to remove the land cover change caused by a tornado in 2011—change that could not have been predicted from historical example. A total of 51 transitions were thus modeled using WNL and MLP (Table 1).

Nine covariates were used in the modeling. Table 2 lists these variables along with their sources. All covariates were co-registered and similarly masked with the land cover data at 30-m resolution. For MLP, the two qualitative variables (county subdivisions and protected areas) were converted to normalized likelihoods for each transition while WNL used them directly.

For MLP, the multi-layer perception procedure in the land change modeler was used with all of its defaults which included.

- Case–control sampling with all instances of change from 2001 to 2011 up to a maximum of 10,000 and an equal size sample of persistence, chosen at random from their respective populations.
- Division of the samples for change and persistence into equal sized subsamples for training and validation, yielding four equal sized sample groups. Because of differences in the prevalence of the transitions, samples within the four groups ranged from 1126 up to 5000.
- Nine input neurons (representing the nine explanatory variables as listed in Table 1), four hidden layer neurons determined automatically by the software, and training over 10,000 iterations (the software default)

Table 1 The transitions modeled for Massachusetts between 2001 and 2011

Barren land to developed, open space
Deciduous forest to developed, open space
Evergreen forest to developed, open space
Mixed forest to developed, open space
Shrub/scrub to developed, open space
Grassland/herbaceous to developed, open space
Pasture/hay to developed, open space
Cultivated crops to developed, open space
Woody wetlands to developed, open space
Emergent herbaceous wetlands to developed, open space
Developed, open space to developed, low intensity
Barren land to developed, low intensity
Deciduous forest to developed, low intensity
Evergreen forest to developed, low intensity
Mixed forest to developed, low intensity
Shrub/scrub to developed, low intensity
Grassland/herbaceous to developed, low intensity
Pasture/hay to developed, low intensity
Cultivated crops to developed, low intensity
Woody wetlands to developed, low intensity
Emergent herbaceous wetlands to developed, low intensity
Developed, open space to developed, medium intensity
Developed, low intensity to developed, medium intensity
Barren land to developed, medium intensity
Deciduous forest to developed, medium intensity
Evergreen forest to developed, medium intensity
Mixed forest to developed, medium intensity
Shrub/scrub to developed, medium intensity
Pasture/hay to developed, medium intensity
Cultivated crops to developed, medium intensity
Woody wetlands to developed, medium intensity
Developed, open space to developed, high intensity
Developed, low intensity to developed, high intensity
Developed, medium intensity to developed, high intensity
Barren land to developed, high intensity
Deciduous forest to developed, high intensity
Evergreen forest to developed, high intensity
Pasture/hay to developed, high intensity
Deciduous forest to barren land
Evergreen forest to barren land
Pasture/hay to barren land
Cultivated crops to barren land
Shrub/scrub to deciduous forest
Deciduous forest to shrub/scrub
Evergreen forest to shrub/scrub
Mixed forest to shrub/scrub
Grassland/herbaceous to shrub/scrub
Deciduous forest to grassland/herbaceous
Evergreen forest to grassland/herbaceous
Emergent herbaceous wetlands to woody wetlands
Woody wetlands to emergent herbaceous wetlands

For WNL, the procedure followed was similar with:

- Case-control sampling with all instances of change from 2001 to 2011 and an equal size sample of persistence, chosen at random from their respective populations.
- Division of the samples for change and persistence into equal sized subsamples for training and validation, yielding four equal sized sample groups. Because of differences in the prevalence of the transitions, samples within the four groups ranged from 1126 up to 29,830.
- Calculation of normalized likelihoods using 200 bins for each quantitative variable and unbinned for qualitative variables. Normalized likelihoods were also filtered using a 13-point Gaussian filter to reduce overfitting.

Figure 2 shows one of the 51 transition potential maps created by each of these techniques—in this case from deciduous forest to low-intensity development as modeled by WNL. Values have a potential range from 0 to 1 and are masked so that non-zero values can only occur in areas that were deciduous forest in 2011.

For both modeling techniques, skill was assessed using three different procedures. Both MLP and WNL incorporate a Heike Skill Score—HSS (Heidke 1926), also known as Kappa (Cohen 1960), calculated as:

$$\mathrm{HSS} = (\mathrm{PC} - E)/(1 - E), \tag{9}$$

where PC is the percent of validation pixels correctly classified by the model using a transition potential threshold of 0.5 to decide between change or persistence, and E is the expected accuracy by chance (which in this case is 50% since case–control sampling was used). The range of this skill index is from -1 to $+1$. A value of 0 indicates no skill—success is no better than what could be achieved by random guessing. Negative values indicate perverse skill (worse than guessing) and generally do not occur except for values very close to 0. All positive values indicate skill better than guessing with $+1$ indicating a case where the model correctly predicted the fate of all validation pixels. For MLP, this skill measure is integral to the process of learning, while in WNL it is provided as a rapid means of model assessment.

Some have criticized HSS/Kappa because many users do not distinguish between skill associated with establishing the correct quantity of change and skill that arises from allocation in the correct location (Pontius 2000). However, since case–control sampling has been used in establishing both the

Table 2 The covariates used in the modeling

Variable	Derivation	Source
Elevation		http://www.nationalmap.gov
Slope gradient	Derived from elevation	
Proximity to surface water	Derived from surface water	http://www.nhd.usgs.gov
Proximity to primary roads (up to 20 km)	2014 US census TIGER/Line roads	http://www.census.gov/geo/maps-data
Proximity to secondary roads	2014 US census TIGER/Line roads	http://www.census.gov/geo/maps-data
Proximity to local roads	2014 US census TIGER/Line roads	http://www.census.gov/geo/maps-data
Proximity to high intensity development	Derived from 2001 land cover	http://www.mrlc.gov
County subdivisions (towns)		http://www.census.gov/geo/maps-data
Protected areas		http://www.gapanalysis.usgs.gov/padus

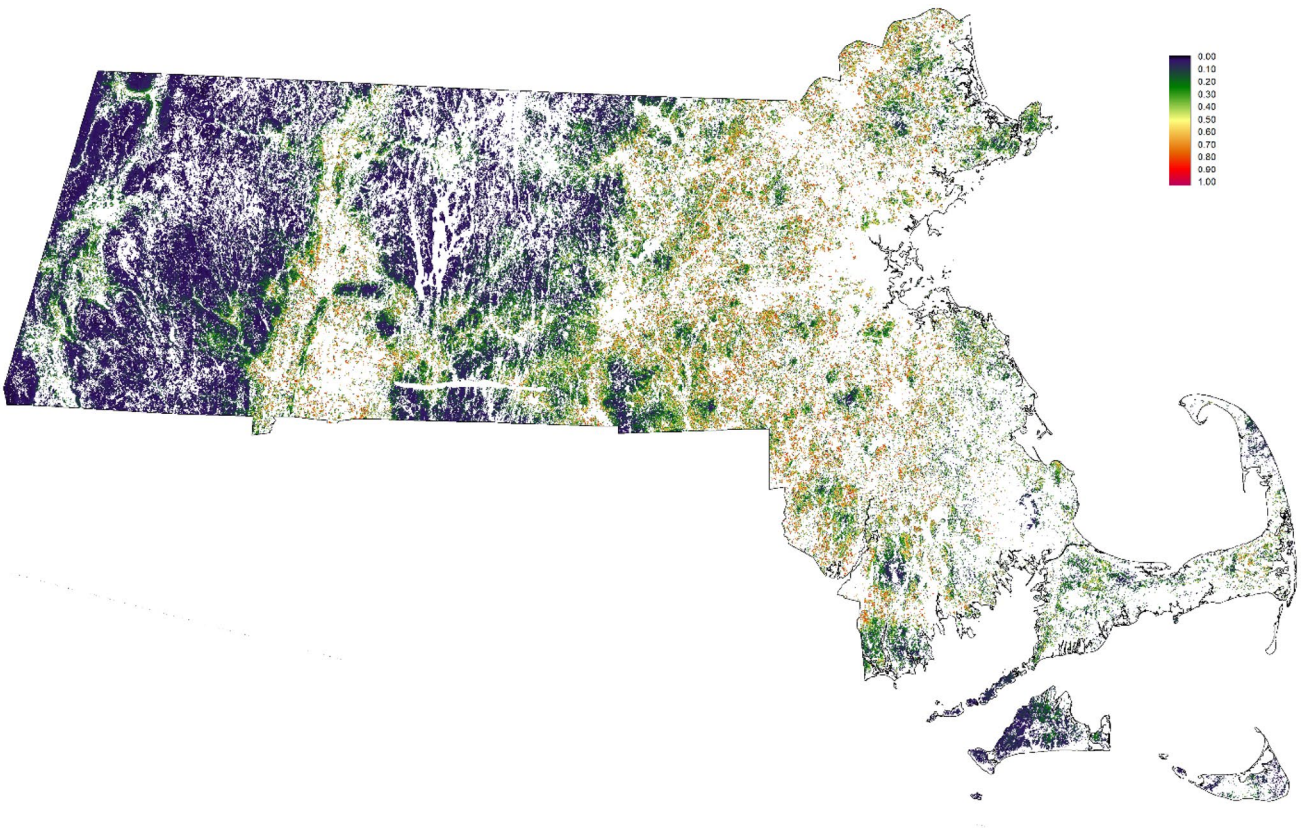

Fig. 2 An example of a transition potential map produced using the WNL procedure. This illustration shows the modeled potential in 2011 for transition from deciduous forest to low intensity development

training and validation sets, and thus fixing the quantity, the skill reported by MLP and WNL is effectively a measure of Kappa for location.

The second measure of skill assessed is based on the receiver operating characteristic (ROC). The ROC (Swets 1973; Pontius and Schneider 2001) assesses the ability of a continuous variable (such as the transition potentials generated here) to generate correct predictions (hits) rather than false predictions (false alarms) by a thresholding of that variable. Whereas, HSS/Kappa tests the skill based on a single threshold (0.5), ROC tests the skill over all thresholds. This is commonly graphed over a range of thresholds (e.g., percentiles) spanning the scope of the continuous variable, with the *x*-axis indicating the false alarm rate and the *y*-axis indicating the hit rate. A skillful forecast is dominated by hits as the thresholds are lowered, and accumulate false alarms at a slow rate. Thus the curve presents an arc that rises steeply from left to right and gradually becomes

asymptotic to 1 on the *y*-axis. In contrast, a random forecast would generate hits and false alarms at approximately the same rate, yielding a straight line from bottom-left to upper-right. A commonly used summary measure of skill is thus the Area Under the Curve (AUC) statistic. A random forecast has an AUC of 0.5 while a forecast with perfect skill has an AUC or 1.0. Values of less than 0.5 indicate forecasts with perverse skill. Figure 3 shows the ROC curves and their associated AUC statistics for the transition from deciduous forest to low density development as modeled by MLP and WNL. Corresponding graphs for all 51 transitions can be found in the supplementary materials.

The third measure of skill evaluated was the user's accuracy (Congalton 1991). The AUC evaluates a decision variable over all possible thresholds. Some have criticized the AUC statistic on the grounds that it is a relative measure and the fact that in applying the decision variable, we are rarely interested in all thresholds (Lobo et al. 2007; Pontius and Parmentier 2014). However, as noted earlier, in land change modeling, a greedy allocation is commonly undertaken whereby it is assumed that the areas that will change are those which have the highest transition potential. Thus the fact that the AUC is a relative measure is not a problem, and singularly appropriate in these cases. That said, it is true that the allocation of change is typically very small. For example, in Massachusetts, using the LCM Markov Chain estimation of the quantity change to the year 2050 (quite a long prediction) resulted in less than a 10% quantity of change for 50 out of the 51 transitions. Thus to test how the transition potential would perform in a typical application, we thresholded each result to extract the highest 10% of transition potentials among the validation pixels to determine the hit rate (where errors would constitute false alarms). This is known as the user's accuracy (UA) in that it tests the reliability of the allocation—i.e., the probability that the allocation of pixels to change or persistence is true according to the validation data (Congalton 1991). This then provides an evaluation of how the transition potential would be expected to perform with a land change model that uses greedy allocation (such as all of the options provided by LCM).

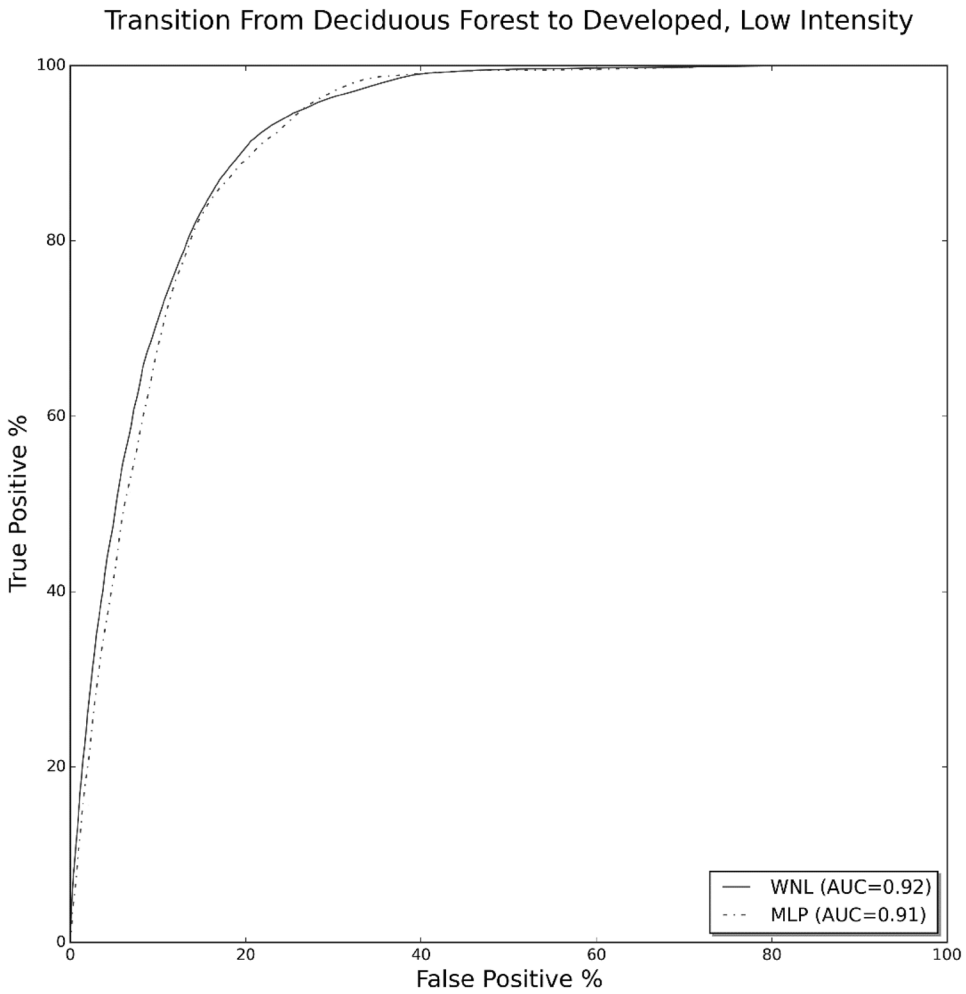

Fig. 3 The ROC curves and AUC statistics for the MLP and WNL modeled transition potentials from deciduous forest to low intensity development, assessed using out-of-sample validation data. Both curves are indicative of strong skill with WNL slightly more skillful at high thresholds found in the lower-left of the curves

Results

Transition potential maps such as that illustrated in Fig. 2 were created for all 51 transitions in Massachusetts using both WNL and MLP. The computer used for the tests had an Intel Xeon CPU running at 3.47 GHz with six cores with hyperthreading (yielding 12 logical processors) and 12 GB of RAM. The WNL procedure using multithreading for parallelization (by assigning different transitions to different threads) took only 283 s while MLP in batch mode required 6437s. Thus, WNL showed a 22-fold improvement over MLP. Even without parallelization, WNL was faster, requiring only 1083 s. Clearly with a greater number of processors, the improvement in speed would be even higher.

Figures 4, 5 and 6 show comparisons of the validation skills achieved by WNL and MLP using the HSS/Kappa, AUC and UA measures over all 51 transitions. In all three, the diagonal line indicates the position of equal skill. The correlation between the HSS/Kappa skills of WNL and MLP in Fig. 4 is $r=0.99$, indicating that the two had extremely similar validation skills over all transitions. Overall, the mean HSS/Kappa skills for WNL and MLP were 0.65 and 0.66, respectively, and were not significantly different ($p>0.10$) on a two-sample two-tailed t test. All tests of means reported here use this procedure since the training and validation samples are randomly chosen with each model run and MLP initializes its learning process with a random assignment of connection weights. Thus, the outcomes do not meet the requirements of a paired test.

Figure 5 shows the AUC skill comparison between WNL and MLP. The correlation between them was also 0.99. Mean AUC was 0.90 for both techniques and there was no significant difference in the means ($p>0.10$).

Figure 6 shows the UA skill comparison between WNL and MLP. Here there is a difference between the two. The correlation between them is lower at $r=0.88$. The mean UA skill was 0.94 for WNL and 0.93 for MLP, which was significant at the $p<0.001$ level.

Finally, a variant of the WNL procedure was written that used global weights for all pixels across each variable rather than the locally determined weights described above. These global weights were calculated as the mean of all the local weights. The purpose was to test whether there was a significant difference in skill between the two approaches.

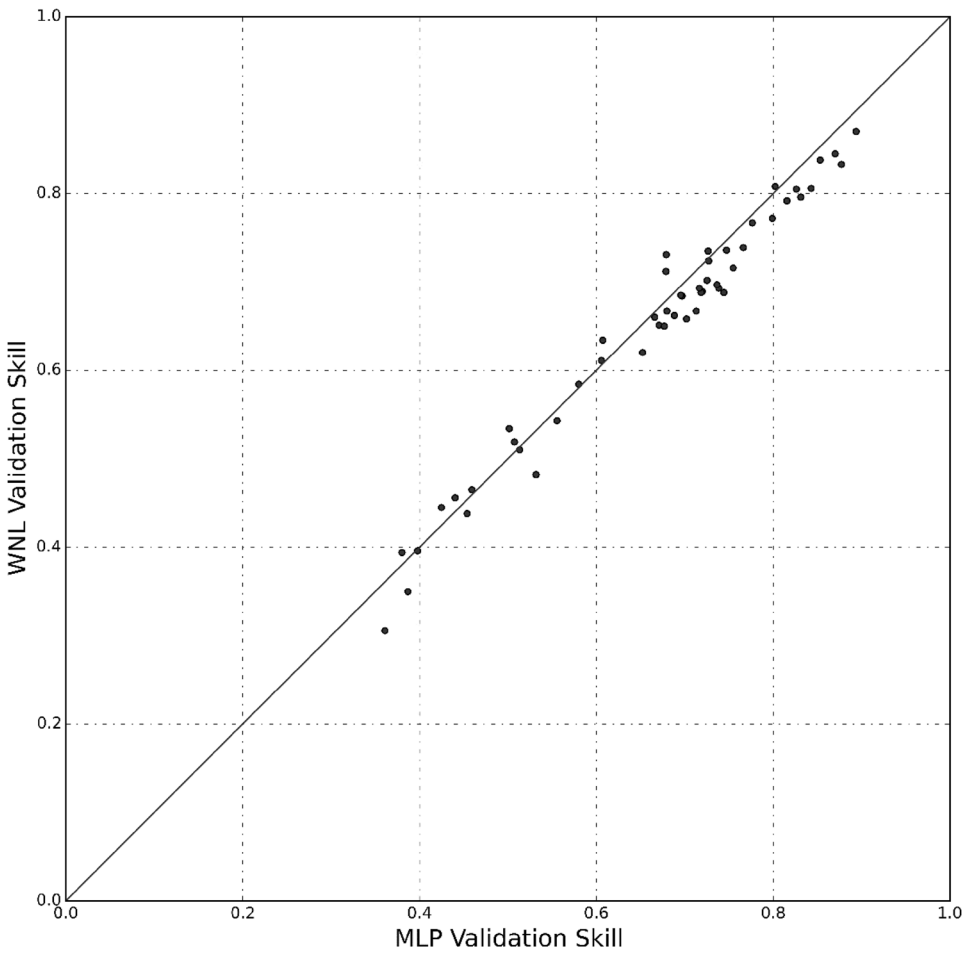

Fig. 4 A comparison of the HSS/Kappa skills of MLP and WNL for the 51 transitions modeled in Massachusetts. The diagonal line shows the position of equal skill

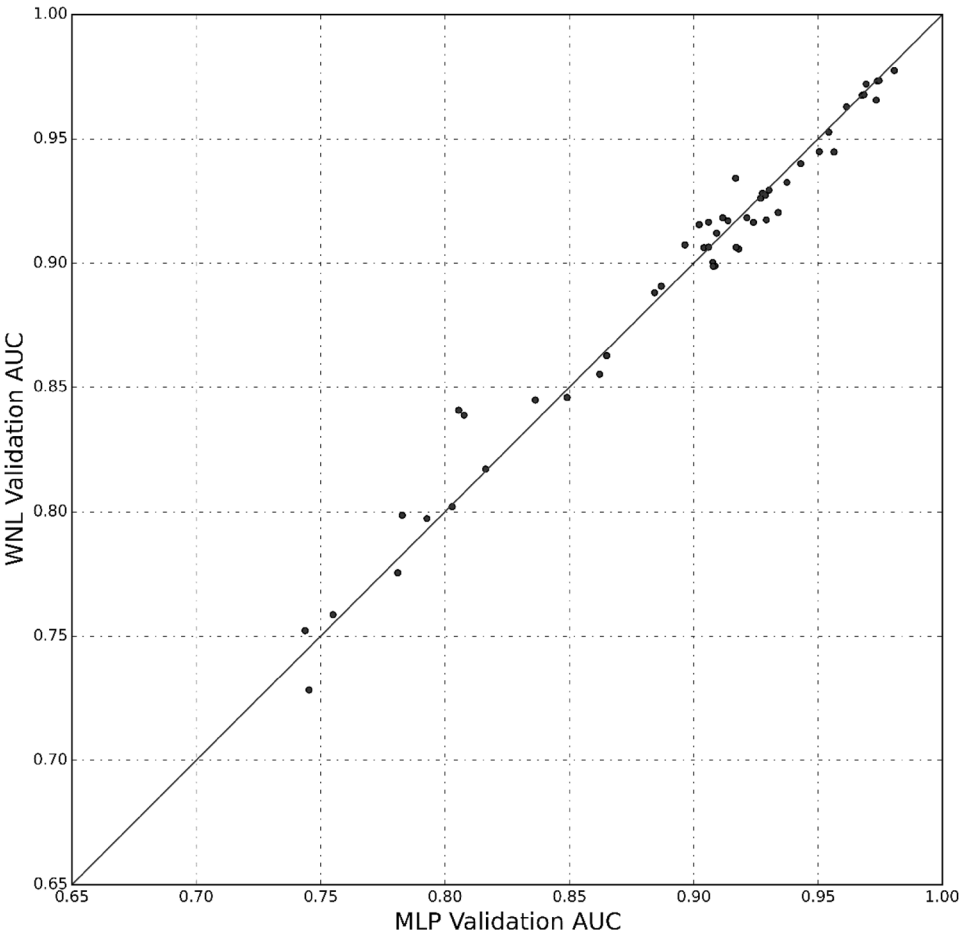

Fig. 5 A comparison of the AUC skills of MLP and WNL for the 51 transitions modeled in Massachusetts. The diagonal line shows the position of equal skill. Note that the origin of the graph starts at 0.65

However, no significant difference was found ($p > 0.10$) on any of the HSS/Kappa, AUC or UA skill tests.

Discussion

From the Massachusetts case study it is clear that WNL produces comparable skill to MLP in modeling transition potentials. The HSS/Kappa and AUC skill tests showed no significant differences. Although the user's accuracy (UA) was significantly higher for WNL, it was not dramatically so. They both performed very well. The big difference, however, was in the speed. WNL is a simpler algorithm. The training stage simply involves the tabulation of relative frequencies of pixels and the storage of histograms as opposed to the laborious iterative learning process of MLP. WNL is also capable of being parallelized in a very simple manner. In this test, parallelization was achieved by multithreading across transitions, yielding almost a fourfold improvement in speed over a single processor implementation (and more than a 22-fold speed improvement over MLP). However, parallelization could also be achieved by logically breaking the problem down into geographical tiles, with different nodes processing the separate tiles followed by an aggregation of tabulations—a very small amount of data that would need to be communicated between nodes. Either approach, or a combination would be highly suitable for distributed computing in a cloud environment.

Regarding the robustness of the procedure, the technique does make the assumption that the independent variables are uncorrelated. How sensitive it is to multicollinearity will be examined in future research. Additionally, because the procedure is linear, interaction effects would need to be tested explicitly by adding additional product variables such as is done with MaxEnt (Phillips et al. 2006). Interestingly, the procedure is not sensitive to irrelevant data. Because the weight assigned to a variable is determined locally based on how strong an opinion that variable has about that specific pixel, irrelevant data will always have a low weight.

Future testing will also focus on the unusual feature of local rather than global variable weighting. We implemented this approach to gain maximum utility from variables that are partially diagnostic (such as the protected lands layer, where there is good diagnostic information in protected areas but not in the bulk of areas that are unprotected). In the tests conducted here there was no significant difference

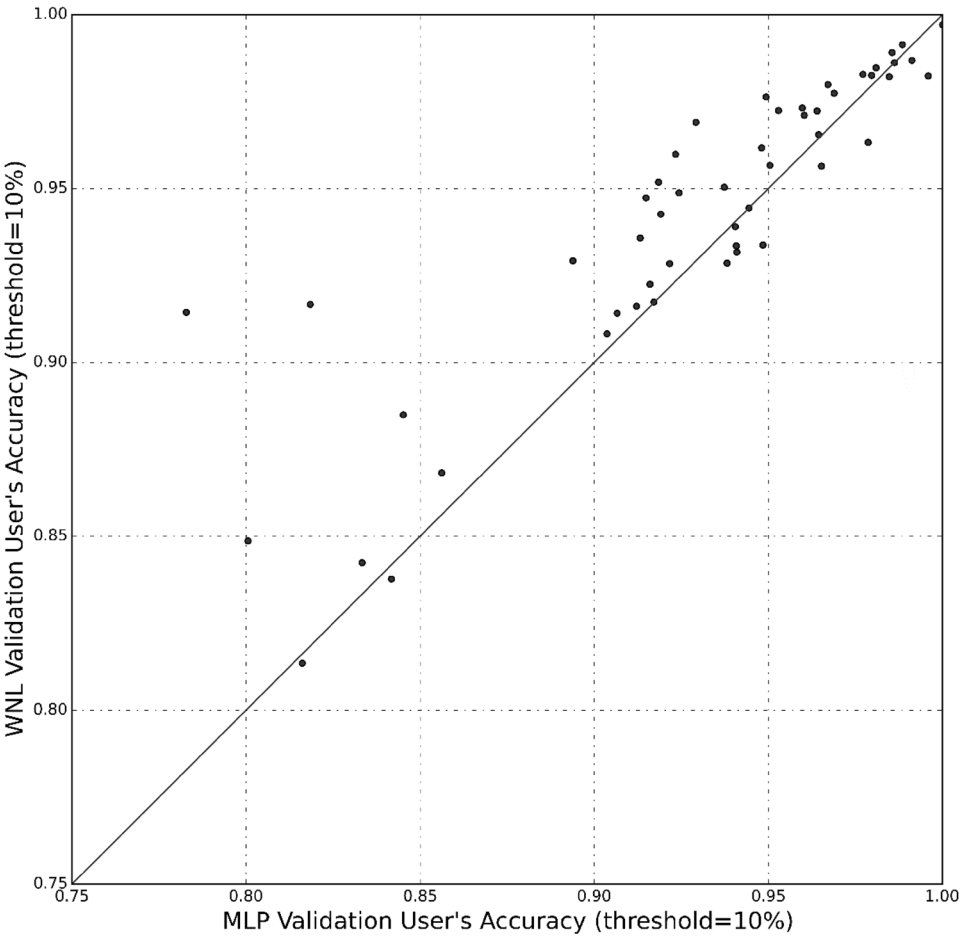

Fig. 6 A comparison of the UA skills of MLP and WNL for the 51 transitions modeled in Massachusetts. The diagonal line shows the position of equal skill. Note that the origin of the graph starts at 0.75

in the results using the local weighting procedure. However, only one of the explanatory variables used (protected lands) exhibited a clearly partial diagnostic character. Regardless, using the local weights does avoid the aggregation step necessary for computing the global variable-specific weights, which also yields a time improvement.

Finally, it should be noted that the procedure is not suited to cases with limited training data. In our use of the procedure we have found that when the sample size falls below about 2000 pixels each for change and persistence (which are subsequently split into 1000 pixels each for training and testing), the model exhibits minor overfitting. This overfitting is evident as a fine-grained striping associated with the covariate bins. This can be reduced by increasing the filtering described in section "Normalized likelihoods". However, the best solution is a larger sample.

Conclusions

The weighted normalized likelihood (WNL) procedure was developed specifically to support empirical land change modeling in a cloud environment. The algorithm combines elements of kernel density estimation of posterior probabilities, aggregation of evidence by linear opinion pooling and local rather than global variable weighting. An initial case study evaluation was achieved by modeling 51 land cover transitions from 2001 to 2011 for Massachusetts and comparing the skill with a multi-layer perceptron (MLP) with an independent validation sample. The WNL procedure was found to have predictive skill comparable to MLP, with the distinct advantage of being efficient to calculate and capable of parallelization.

Acknowledgements This research was conducted as part of a grant from Esri Inc. to develop a cloud-based implementation of Clark Labs' Land Change Modeler. We are very grateful to Esri for this support.

References

Angelsen A, Boucher D, Brown S, Valerie M, Streck C, Zarin D (2011) Modalities for REDD+ reference levels: technical and procedural issues. The Meridian Institute

Berry MW, Hazen BC, MacIntyre RL, Flamm RO (1996) LUCAS: a system for modeling land-use change. IEEE Comput Sci Eng 3(1):24–35

Clements MP, Harvey DI (2011) Combining probability forecasts. Int J Forecast 27(2):208–223
Cohen JA (1960) Coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46
Congalton RG (1991) A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens Environ 37:35–46
Eastman JR (2006) IDRISI 15.0: The, Andes edn. Clark University, Worcester
Eastman JR (2014) The TerrSet geospatial monitoring and modeling system. Clark University, Worcester
Eastman JR, Solorzano LA, Van Fossen ME (2005) Transition potential modeling for land-cover change. In: Maguire DJ, Batty M, Goodchild MF (eds) GIS, spatial analysis, and modeling. ESRI Press, California, pp 357–385
Flach P (2012) Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press, Cambridge, England
Heidke P (1926) Berechnung des erfolges und der gute der windstarkvorhersagen im sturmwarnungsdienst. Geogr Ann 8:301–349
Iizuka K, Johnson BA, Onishi A, Magcale-Macandog DB, Endo I, Bragais M (2017) Modeling future urban sprawl and landscape change in the Laguna de Bay Area, Philippines. Land 6:2. https://doi.org/10.3390/land6020026
Jordan MI (1995) Why the logistic function? A tutorial discussion on probabilities and neural networks, MIT computational cognitive science, MIT Computational Science Report 9503
King G, Zeng L (2001a) Explaining rare events in international relations. Int Org 55:693–715
King G, Zeng L (2001b) Logistic regression in rare events data. Polit Anal 9:137–163
Lin Y-P, Chu H-J, Wu C-F, Verburg PH (2011) Predictive ability of logistic regression, auto-logistic regression and neural network models in empirical land-use change modeling—a case study. Int J Geogr Inf Sci 25(1):65–87
Lobo JM, Jiménez-Valverdel A, Real R (2007) AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr 17:145–151
National Research Council (2014) Advancing land change modeling: opportunities and research requirements. The National Academies Press, Washington. https://doi.org/10.17226/18385
Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33(3):1065. https://doi.org/10.1214/aoms/1177704472
Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190:231–259
Pijanowski BC, Brown DG, Shellito BA, Manik GA (2002) Using neural networks and GIS to forecast land use changes: a land transformation model. Comput Environ Urban Syst 26(6):553–575
Platt J (2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola A, Bartlett P, Scholkopf B, Schuurmans D (eds) Advances in large margin classifiers. MIT Press, Cambridge
Pontius RG Jr (2000) Quantification error versus location error in the comparison of Tso, B., and Mather, P.M., 394 (2011) categorical maps. Photogram Eng Remote Sens 66(8):1011–1016
Pontius RG Jr, Schneider L (2001) Land-use change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. Agric Ecosyst Environ 85:239–248
Pontius R Jr, Parmentier B (2014) Recommendations for using the relative operating characteristic (ROC). Landsc Ecol 29:367–382
Pontius RG Jr, Cornell JD, Hall CAS (2001) Modeling the spatial pattern of land use change with GEOMOD2: application and validation for Costa Rica. Agric Ecosyst Environ 1775:1–13
Ranjan R, Gneiting T (2010) Combining probability forecasts. Stat Methodol 72(1):71–91
Rish I (2001) An empirical study of the naïve Bayes classifier. IBM Research Report RC 22230
Roberts HV (1965) Probabilistic prediction. J Am Stat Assoc 60(309):50–62
Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. Ann Math Stat 27(3):832. https://doi.org/10.1214/aoms/1177728190
Soares-Filho BS, Pennachin CL, Cerqueira G (2002) DINAMICA EGO—a stochastic cellular automata model designed to simulate the landscape dynamics in an Amazonian colonization frontier. Ecol Model 154(3):217–235
Swets JA (1973) The relative operating characteristic in psychology. Science 240:1285–1293
Tso B, Mather PM (2001) classification methods for remotely sensed data. Taylor and Francis, New York
Turner BL II, Lambin EF, Reenberg A (2007) The emergence of land change science for global environmental change and sustainability. Proc Natl Acad Sci USA 104:20666–20671
Verburg PH, Soepboer W, Veldkamp A, Limpiada R, Espaldon V, Sharifah Mastura SA (2002) Modeling the spatial dynamics of regional land use: the CLUE-S Model. Environ Manage 30(3):391–405
Verified Carbon Standard (2013) VCS module VMD0007 REDD methodological module: estimation of baseline carbon stock changes and greenhouse gas emissions from unplanned deforestation (BL-UP) version 3.1. http://database.v-c-s.org/sites/vcs.benfredaconsulting.com/files/VMD0007%20BL-UP%20v3.2.pdf. Accessed 8 Dec 2018
Weiss GG (2004) Mining with rarity: a unifying framework. SIGKDD Explorations 6(1):7–19
Zhang Q, Cheng L, Boutaba R (2010) Cloud computing: state-of-the-art and research challenges. J Internet Ser Appl 1(1):7–18. https://doi.org/10.1007/s13174-010-0007-6