Transactions in GIS, 2003, 7(4): 467–484

Research Article

Using the Relative Operating Characteristic to Quantify Certainty in Prediction of Location of Land Cover Change in India

R Gil Pontius Jr
Graduate School of Geography
Clark University

Kiran Batchu
Graduate School of Geography
Clark University

Abstract

This paper describes a methodology by which modelers, ecologists and planners can quantify the certainty in predicting the location of change for a given quantity of change. The specification of the quantity of a land cover category and the specification of the location of a land cover category are two distinct fundamental concepts in geographical analysis. It is crucial that scientists have appropriate quantitative tools to analyze each of these two concepts independently of one another. This paper gives methods whereby a scientist can convert a map of relative propensity for disturbance to a map of probability of future disturbance, based on a quantifiable validation of a map's predictive ability. The required inputs are: (1) maps that show a Boolean categorical variable at times 0, 1 and 2, (2) a technique to create a map that shows the relative propensity for membership in the Boolean category, and (3) a predicted proportion of the category at time 3.

1 Introduction

1.1 From thresholds to predictions

One of the most common procedures in Geographic Information Science is the conversion from a real variable on the interval [0, 1] to a Boolean categorical variable of zero or one. If the real value is greater than some specified threshold, then the categorical variable is assigned a value of one, else the categorical variable is assigned a value of zero. Examples abound. In the ecological and environmental sciences, researchers use this threshold procedure to generate raster maps of presence or absence of species (Wu and

Address for correspondence: R Gil Pontius Jr, Graduate School of Geography, Clark University, 950 Main Street, Worcester, MA 01610-1477, USA. E-mail: rpontius@clarku.edu

Huffer 1997). In remote sensing, scientists use thresholds to interpret satellite images in order to create a resultant image of categories of land cover (Mather 1999). In decision analysis, scientists create maps where each grid cell shows the suitability for a particular land use type, and then a threshold is set to create a map of recommended land use categories (Eastman 1995). In land-use change modeling, scientists generate maps of the likelihood of deforestation, then a threshold is set to determine which cells show the category of predicted deforestation (Veldkamp and Lambin 2001). This fourth case of predicting disturbance to a landscape serves as the example in the remainder of this paper. For this example, we call the real value on the interval [0, 1] a "propensity for disturbance". If the propensity of a cell in a raster map is greater than the threshold, then the cell is assigned to the "disturbed" category, else the cell is assigned to the "non-disturbed" category.

The decision concerning the level of the threshold is a decision that determines the magnitude of the quantity of area of land assigned to each category. If the threshold is high, then a small quantity of land is assigned to the disturbed category because only a small number of cells have a propensity for disturbance above a high threshold. If the threshold is lower, then a larger quantity of land is assigned to the disturbed category. In this sense, a map of propensity for disturbance contains information about only the relative geographical location of predicted disturbance, because the propensity values have meaning only in terms of their relative ordering, not in terms of their magnitudes. A map of propensity for disturbance claims nothing concerning the predicted quantity of disturbance. After the scientist sets the threshold and performs the reclassification, the categorical map of the predicted landscape specifies both the location and quantity of predicted disturbance versus non-disturbance.

The specification of the quantity of a category and the specification of the location of a category are two distinct fundamental concepts in geographical analysis. It is important that scientists have appropriate quantitative tools to analyze each of these two concepts independently of one another. This paper presents statistical tools to examine the specification of location, independently from the specification of quantity, for cases where a scientist converts a real number on the interval [0, 1] to a Boolean variable for the purpose of prediction and/or accuracy assessment.

1.2 Probability of disturbance on a landscape

Logistic regression is perhaps the most common method to create a raster map of propensity for disturbance (Ludeke et al. 1990, Irwin and Geoghegan 2001, Geoghegan et al. 2001). Each observation of the regression analysis is a grid cell that is non-disturbed at time 0. The dependent variable takes a value of one if disturbance happened in a cell between time 0 and time 1, and zero if disturbance did not happen. Typical dependent variables are distance to roads, distance to markets, and slope. Logistic regression produces a map where each grid cell has a fitted value on the interval (0, 1).

How should we interpret the fitted values that logistic regression produces? Some scientists commonly refer to these fitted values as predicted probabilities. However, if the purpose of the logistic regression is to predict future disturbance beyond time 1, then there are two reasons why it is faulty to interpret these fitted values as predicted probabilities. First, the fitted values are not predictions because the data from time 0 and time 1 are known and are used to generate the values. It is an oxymoron to predict something that is known and that happened in the past. Second, the fitted values are not probabilities because the data from time 0 and time 1 show definitively which cells became disturbed and which did not, therefore the probability of disturbance in those cells that became disturbed between time 0 and time 1 is one, and the probability of disturbance in those cells that were not disturbed between time 0 and time 1 is zero.

We should find meaning in the relative ordering of the fitted values, not in their magnitudes. If the fitted value of a particular cell is ŷ, we should not assume that the probability of cell becoming disturbed in the future is ŷ. If we were to insist that the magnitude of the fitted values is meaningful, then we would attribute characteristics to the fitted values that they are not designed to portray. Specifically, if we interpret the fitted values as probabilities, then the average of the fitted values among the non-disturbed cells of time 1 implies a specific quantity of predicted future disturbance between time 1 and some future time 2. The fitted values are not designed to predict the quantity of land that will become disturbed between time 1 and time 2.

However, it is possible to use the fitted values to help to predict which cells will become disturbed between time 1 and some future time 2. The method should be to examine the cells that remained non-disturbed at time 1, then to predict the cells that have the largest fitted values will become disturbed before the cells that have relatively smaller fitted values. We would need an independent calculation to predict the quantity of cells that would be disturbed between time 1 and time 2.

Ultimately, decision-makers need maps that show interpretable probabilities of future disturbance, not simply propensities for future disturbance. More importantly, decision-makers must know the extent to which they can trust predictions. Therefore, scientists must create maps where each cell shows the probability that the cell will become disturbed by some specific time. These maps should communicate clearly the level of certainty in terms of two important questions: (1) What is the level of certainty of the predicted quantity of disturbance? and (2) What is the level of certainty of the predicted location of disturbance? This paper offers methods to assess the second of these two questions. To clarify the distinction between the two questions, consider the example in the following section.

1.3 Certainty of location versus certainty of quantity

Suppose a scientist predicts that *P* is the proportion of a landscape that will be disturbed at some point in the future. This prediction of the quantity *P* alone gives no information concerning the spatial distribution of the disturbance on the future landscape. But suppose that the scientist would like to create a map of many grid cells for which each cell shows the probability of being disturbed in the future, given the estimate of *P*. If the scientist were completely uncertain of the location of the future disturbance, then every cell of the map should show a probability of *P* of being disturbed. If the scientist were completely certain of the location of the future disturbance, then the map of the predicted disturbance should contain a probability of 1 in *P* proportion of the cells and a probability of 0 in 1-*P* proportion of the cells. In most cases, the scientist's level of certainty concerning the predicted location of the future disturbance will be somewhere between no certainty and perfect certainty. If the scientist has an intermediate level of certainty, then the scientist should make a map such that the probability for each grid cell would be near *P* for locations that are very uncertain, and would be far from *P* (i.e. near 0 or 1) for locations that are very certain.

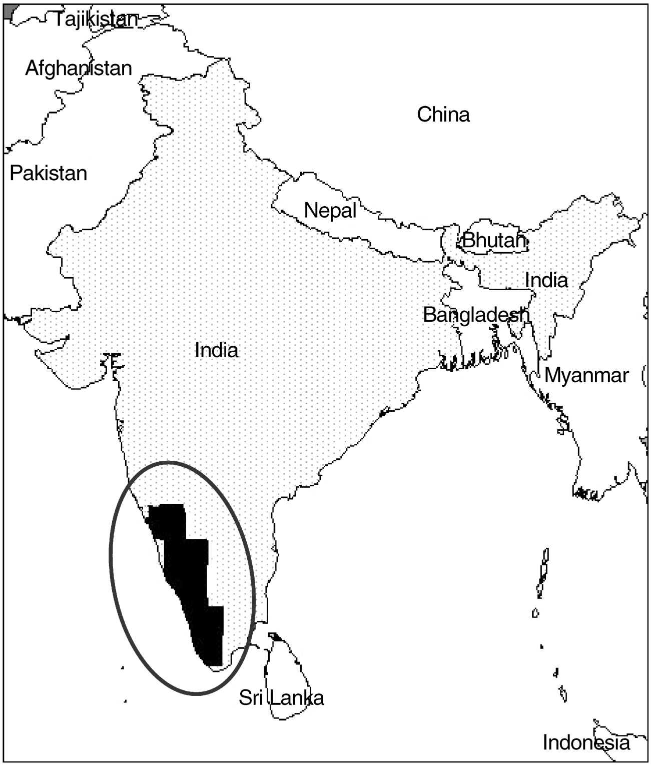

Figure 1 Map of the Western Ghats in India

This paper shows methods whereby a scientist can create a map of the probability of future disturbance, based on an estimated quantity of future disturbance and a measure of a map's ability to predict the location of future disturbance. We illustrate the procedure with an example of forest disturbance in India's Western Ghats.

1.4 Biodiversity and Human Disturbance in the Western Ghats

The Western Ghats consist of a chain of mountain ranges that stretch along the western coast of India, from the Vindhya-Satpura in the north to the tip of the Indian peninsula in the south (Figure 1). The Western Ghats span across four major states of India: Maharashtra, Karnataka, Kerala and Tamilnadu. Covering an estimated area of 160,000 km^2, the Western Ghats are an area of exceptional biological diversity and conservation interest (Rodgers and Panwar 1988). Almost one-third of all the flowering plant species in India are found in the Western Ghats region. The complex topography and heavy rainfall have made certain areas inaccessible and have helped the region retain its diversity.

Nevertheless, human disturbance has had a great influence on vegetation. Human activity has caused degradation, thus some forest areas have shrunk considerably (Menon and Bawa 1997). The Western Ghats has suffered rapid deforestation in the past few decades due to large-scale conversion of forests for fuelwood, roads, and plantations of tea and coffee. The loss of forest in sensitive regions such as steep slopes aggravates soil erosion and flooding. Some endemic species have disappeared entirely, and others are on the verge of extinction. Due to its richness and its vulnerability, the Western Ghats is among the world's biodiversity "hot spots" (Myers et al. 2000).

It is essential to understand and to anticipate the threat to biodiversity due to the cumulative effect of human disturbance. Therefore, we have applied a statistical technique to simulate cumulative forest disturbance between years for which we have maps (1920 and 1990) and to predict disturbance into the future. This paper uses the Western Ghats to demonstrate a technique to create a map that shows the probability of future disturbance at any specific location.

2 Methods

2.1 Data

Figure 2 shows cumulative forest disturbance, which is the dependent variable. The darkest shade denotes cells that were disturbed between time 0 and time 1. The medium shade shows cells that were disturbed between time 1 and time 2. The lightest shade shows cells that remain undisturbed at time 2. In this case, time 0 is an unspecified pre-human time. Time 1 is 1920 and time 2 is 1990. White denotes areas that are outside the Western Ghats, that are water, or that are locations for which data are not available. The resolution of each grid cell is approximately 1 km by 1 km. The accuracy of the maps for this analysis is not known precisely, but we accept the maps as points of reference because the purpose of this paper is to demonstrate a technique of analysis.

The entire Western Ghats would probably be forested, were it not for humans, because the entire study area has the biophysical potential to be forest (Kamaljit Bawa, personal communication). Therefore we assume a completely non-disturbed landscape at some time 0, called pre-human. This assumption allows the technique to be applied in cases where maps are available for only two points in time, e.g. 1920 and 1990.

We derive Figure 2 from a pair of individual maps of 1920 and 1990 in which each cell is categorized as forest or non-forest. The 1920 map was digitized from paper maps of Survey of India toposheets, printed by the Army Map Service, Corps of Engineers, surveyed in the years ranging from 1910 to 1930. The map of forest of 1990 was created from satellite imagery. The initial classification had the following land use types: Dense Forest, Less Dense Forest, Tree Plantations, Coffee, Tea, Scrub, Open, and Water. The first two categories are classified as forest. Non-forest areas are plantations, coffee, tea, scrub and open. All the non-forest areas are considered disturbed. The motivation for this categorization is to represent the cumulative threat to biodiversity.

A small number of cells exhibit forest re-growth between 1920 and 1990. Cells that are non-forest in 1920 and forest in 1990 are placed in the category "disturbed between pre-human and 1920" because our motivation is to assess the cumulative threat to biodiversity. Consequently, Figure 2 shows one-way cumulative disturbance from a pre-human time to 1920 to 1990. The assumption that the change is a one-way conversion of a Boolean variable is necessary for the technique of this paper. The need for this assumption is a limitation of the applicability of the technique.

Figure 3 shows an independent variable, slope. The slope map was derived from an elevation map that has pixels of 1 km by 1 km. The coarseness of the elevation map results in a modest range in slopes. If the resolution of the elevation map were finer,

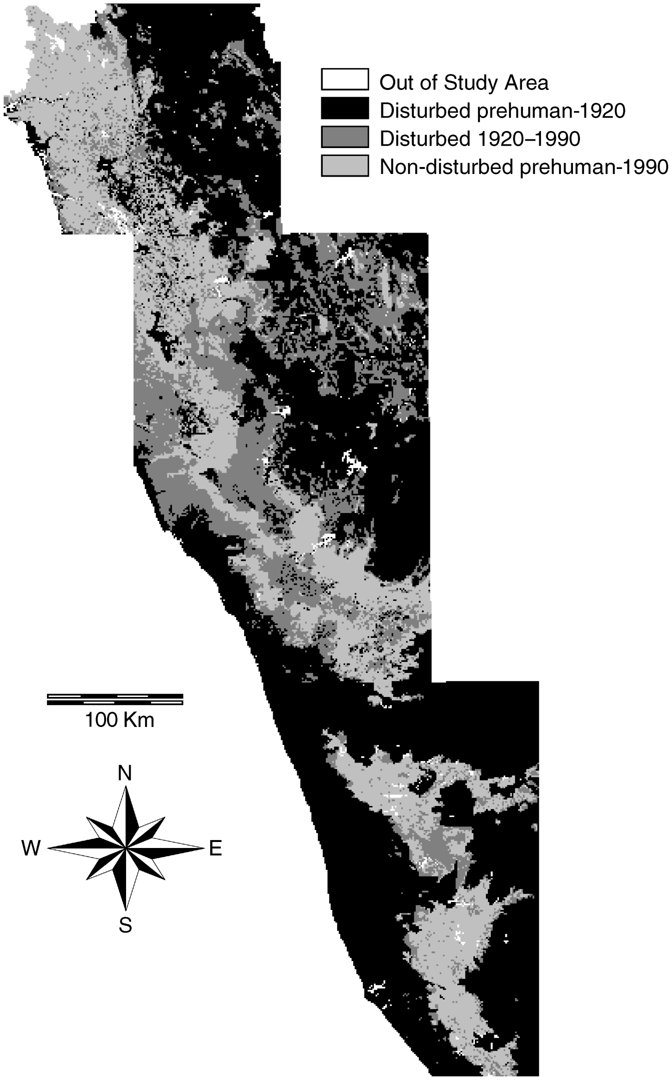

Figure 2 Map of cumulative forest disturbance at 1920 and 1990

then the range in slopes could be more extreme. Figure 3 distinguishes flatter regions from steeper regions at a resolution that is consistent with the forest map. We think that slope is important in predicting human disturbance to the landscape in the rugged terrain of the Western Ghats. We hypothesize that humans would prioritize disturbance on flatter slopes before steeper slopes.

2.2 *Strategy*

Figure 4 shows the sequence of operations of our analysis. The rectangles with solid borders denote maps. The rectangles with dotted borders denote non-map information.

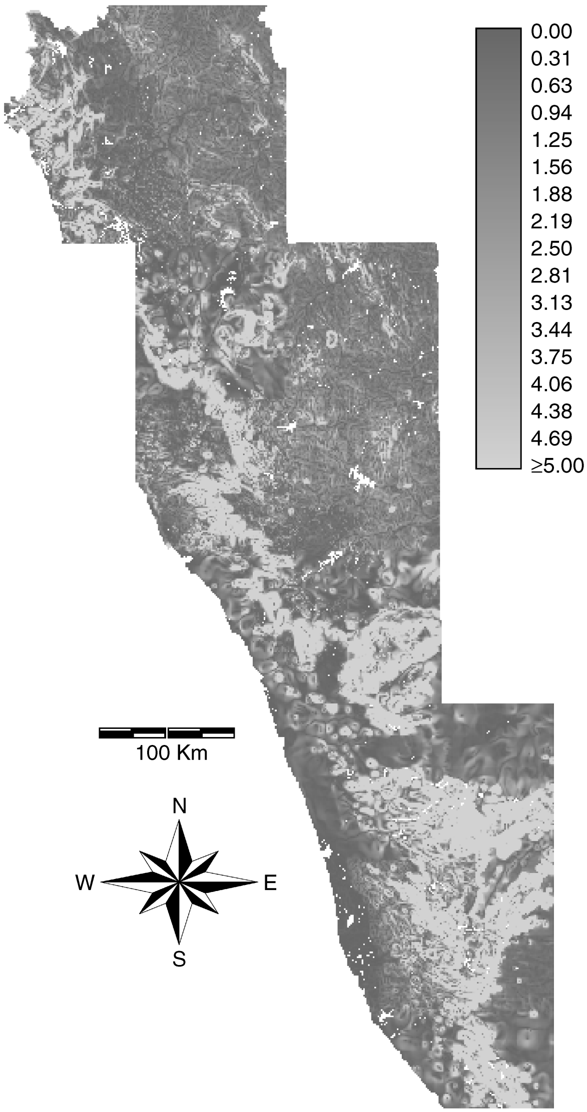

Figure 3 Map of slope in percent

The arrows show the flow of information and the sequence of steps. The first step is to create a propensity for post-1920 disturbance map from only the variables that would have been available in 1920. This can be accomplished by performing logistic regression to explain the disturbance between pre-human times and 1920 as a function of slope. The resulting fitted values give a propensity for post-1920 disturbance map, which is used to predict the disturbance between 1920 and 1990. Hence the propensity for post-1920 disturbance map is compared to the 1990 map, to compute a goodness-of-fit of validation for the predicted disturbance between 1920 and 1990. A statistic called the Relative Operating Characteristic (ROC) measures the goodness-of-fit of the validation.

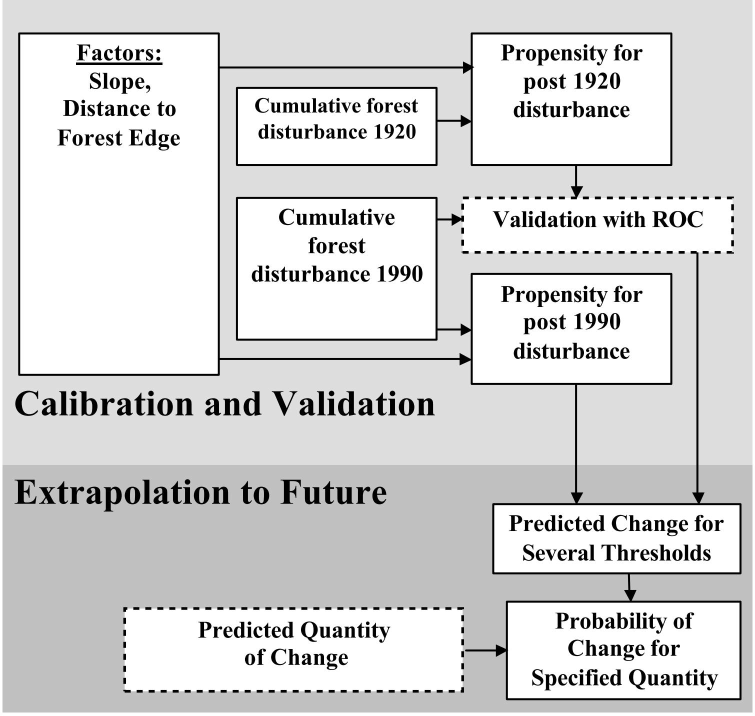
Factors:
Slope,
Distance to
Forest Edge
Cumulative forest disturbance 1920
Propensity for post 1920 disturbance
Cumulative forest disturbance 1990
Validation with ROC
Propensity for post 1990 disturbance
Calibration and Validation
Extrapolation to Future
Predicted Change for Several Thresholds
Predicted Quantity of Change
Probability of Change for Specified Quantity

Figure 4 Flow of methods

Then the model is recalibrated to generate a new propensity for disturbance map that shows the relative likelihood for post-1990 disturbance, which is subsequently used to predict post-1990 disturbance. Lastly, two types of adjustments convert the propensity for post-1990 disturbance map into a probability of post-1990 disturbance map. The two adjustments derive from the dotted rectangles, which are: (1) the ROC goodness-of-fit of validation, and (2) an independent prediction of the quantity of post-1990 disturbance. The subsequent subsections give details of the inner workings of each rectangle of Figure 4.

2.3 Specification of propensity for post-1920 disturbance

A calibration procedure uses data from 1920 to generate the propensity for post-1920 disturbance map. Logistic regression is one of the most commonly used calibration methods, which uses empirical analysis to establish a relationship between the independent variable and the propensity for post-1920 disturbance. In the regression, the dependent variable is 1 if the cell is disturbed between pre-human times and 1920, and 0 if the cell is non-disturbed by 1920. The independent variable is slope, which we assume does not change appreciably over time. The propensity for disturbance values are the regression's resulting fitted values. In order to predict post-1920 disturbance, the values are relevant for only those cells that are non-disturbed in 1920, since only those cells are candidates for post-1920 disturbance.

Alternatively, we can use distance to forest edge as the propensity for post-1920 disturbance map. In this case, the non-disturbed cells that touch disturbed cells receive a propensity value of 1 and the non-disturbed cell that is farthest from any disturbed cell receives a propensity value of 0. Using these two extremes, linear interpolation sets a relationship between propensity and distance to forest edge, thus assigns a propensity value to each non-disturbed cell of 1920 as a linear function of its distance to edge as defined by:

$$P(n) = \frac{\text{distance between cell } n \text{ and the nearest disturbed cell}}{\text{maximum distance between any cell } n \text{ and the nearest disturbed cell}} \tag{1}$$

where $P(n)$ = propensity for cell n to become disturbed after 1920; n = index of cell that is non-disturbed in 1920 = 1, 2, . . . , N; and N = number of cells that are non-disturbed in 1920.

The value of the propensity for disturbance in each cell shows the likelihood of that cell for disturbance, relative to the other cells. That is, we would predict that the cells with the larger propensity values would become disturbed before the cells with the smaller propensity values. Therefore we can test the validity of the propensity for post-1920 map by comparing it to the map of real disturbance between 1920 and 1990.

There are undoubtedly more complex methods, such as multiple logistic regression or multi-criteria analysis, to construct the propensity for disturbance map. However the purpose of this analysis is to show a technique to convert any propensity for disturbance map to a probability of future disturbance map, therefore we do not analyze more complicated methods of generating the propensity for disturbance map.

2.4 Validation for predicted change between 1920 and 1990

A validation procedure assesses how well the propensity for post-1920 disturbance map matches the map of real 1920–90 disturbance. The goodness-of-fit of the validation measures the predictive ability of the calibration procedure that created the propensity for post-1920 map. The propensity for post-1920 disturbance map has predictive power if the larger propensity values are concentrated at locations that truly became disturbed between 1920 and 1990. The Relative Operating Characteristic (ROC) is a statistic that measures the extent to which this is true. Swets (1986, 1988) describes the logic of the ROC in depth. Others show how to compute the ROC in the context of digital maps (Pontius and Schneider 2001, Pontius and Pacheco 2003). Here we give a brief description of the ROC.

The ROC is a method to compare a Boolean variable (e.g. real 1920–90 disturbance) versus an order variable (e.g. propensity for post-1920 disturbance). The ROC requires that we compute the accuracy of the prediction at several different threshold levels. For each threshold level, each cell of the propensity for post-1920 disturbance map is reclassified as either above or not above the threshold. Equation (2) shows how any particular threshold defines a predicted map of Boolean disturbance versus non-disturbance:

$$\begin{aligned} S_i(n) &= 1 \quad \text{if} \quad T_i < P(n) \\ &= 0 \quad \text{else} \end{aligned} \tag{2}$$

where $S_i(n)$ = predicted disturbance for cell n at threshold T_i; and i = index of threshold = 0, 1, . . . , I; T_i = threshold i; and I = number of bins created by the thresholds.

Table 1 Two-by-Two contingency table where 'i' is threshold index and A_i, B_i, C_i, and D_i are numbers of grid cells that are candidates for post-1920 disturbance in the map.

		Reality		
		Disturbed	Non disturbed	Total
Predicted	Disturbed	A_i (True Positives)	B_i (False Positives)	$A_i + B_i$
	Non disturbed	C_i (False Negatives)	D_i (True Negatives)	$C_i + D_i$
	Total	$A_i + C_i$	$B_i + D_i$	$A_i + B_i + C_i + D_i$

The initial threshold is one (i.e. $T_0 = 1$) such that all cells are predicted as non-disturbance. Each subsequent threshold is lower, so each subsequent predicted map has a larger quantity of cells in the disturbed category. At every threshold, we compute agreement between the predicted map and the map of real disturbance between 1920 and 1990 (Figure 2). The final threshold is less than 0 such that all cells are predicted as disturbance.

Let us define bin i as the union of cells that have a propensity for disturbance value that falls between threshold level i and i-1. Equation (3) shows that we set the threshold levels such that each bin i has the same number of cells for $i = 1, 2, \ldots, I$:

$$\left[\sum_{n=1}^{N} S_i(n)\right] - \left[\sum_{n=1}^{N} S_{i-1}(n)\right] = \frac{N}{I} \tag{3}$$

For each threshold level, we assess the accuracy of the prediction by using a contingency table such as Table 1. The columns of Table 1 refer to the category in the map of real disturbance between 1920 and 1990. Thus $A_i + C_i$ denotes the number of cells that are disturbed in the reality map, whereas $B_i + D_i$ denotes the number of cells that are non-disturbed in the reality map. The rows of Table 1 refer to the category in the map of predicted disturbance, as defined by the propensity for post-1920 disturbance map and threshold i. Thus the entry in the first row and first column of Table 1, denoted as A_i, is the number of non-disturbed cells of 1920 that are classified as disturbed in both the prediction map and the map of real disturbance between 1920 and 1990. A_i is the number of true positive cells for threshold i. Thus the entry in the first row and second column of Table 1, denoted as B_i, is the number of non-disturbed cells of 1920 that are classified as disturbed in the prediction and non-disturbed in the map of real disturbance between 1920 and 1990. B_i is the number of false positive cells for threshold i. A "positive" is a cell that is categorized as disturbed in the predicted landscape. C_i is the number of False Negatives and D_i is the number of True Negatives. For the ROC analysis, Equation (4) gives the rate of true positives and Equation (5) gives the rate of false positives for each threshold i:

$$Y_i = \frac{A_i}{A_i + C_i} \tag{4}$$

$$X_i = \frac{B_i}{B_i + D_i} \tag{5}$$

Figure 5 plots the rate of true positives versus the rate of false positives for each threshold. The point (0, 0) derives from the first threshold T_0 and the point (1, 1) derives

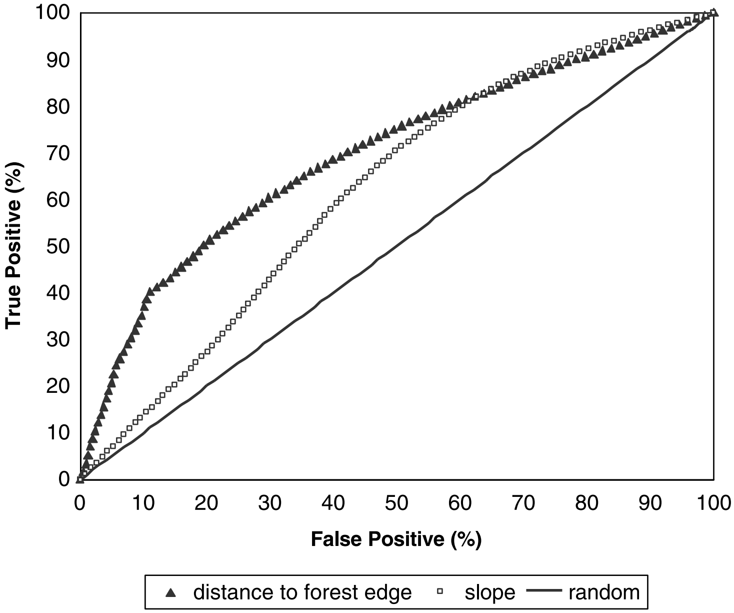

Figure 5 ROC curves to validate predicted additional disturbance 1920–1990 for three propensity maps based on: (1) distance to forest edge, (2) logistic regression using slope, and (3) random location

from the last threshold T_I. Points near the origin derive from high thresholds and points near (1, 1) derive from low thresholds. When all the points are connected with line segments, the area under the resulting curve is the ROC value. If the propensity for disturbance values were distributed at random locations, then the expected ROC curve would be the straight diagonal line between (0, 0) and (1, 1), hence the area under the curve would be 0.5. Alternatively, if the propensity for disturbance map were perfect, then the ROC curve would begin at the point (0, 0), proceed straight to the point (0, 1), then straight to the point (1, 1), hence the area under the curve would be 1. A perfect propensity for disturbance map is a map in which all the largest propensity for post-1920 disturbance values are at locations of real post-1920 disturbance.

In this respect, the ROC reflects the producer's accuracy, because Equation (4) indicates the producer's accuracy for the disturbed category. The producer's accuracy answers the question: "Given that a cell is really disturbed, what is the probability that the cell is predicted as disturbed?" This type of accuracy is interesting to the map-producer.

However, a decision-maker who examines the predicted map is more interested in the question: "Given that the cell is predicted as disturbed, what is the probability that the cell really is disturbed?" This type of accuracy is the "user's accuracy". Congalton and Green (1999) describe the difference between producer's accuracy and user's accuracy. Equation (6) defines the user's accuracy, denoted Z_i, for the disturbed category for every bin i, where $i = 1, \ldots, N$:

$$Z_i = \frac{A_i - A_{i-1}}{(A_i - A_{i-1}) + (B_i - B_{i-1})} \tag{6}$$

Bin i consists of the cells that have a propensity for disturbance value that is greater than or equal to threshold i and less than threshold $i - 1$. We compute a value of Z_i for each bin i, during the validation of the prediction of the disturbance between 1920 and 1990. We store these values of Z_i for use in the extrapolation part of the analysis to predict post-1990 disturbance, as shown in the bottom of Figure 4.

2.5 *Respecification of propensity for post-1990 disturbance*

A recalibration procedure creates propensity for post-1990 disturbance maps in a manner similar to the calibration method used to create propensity for post-1920 disturbance maps. The only difference is that the number of cells that are non-disturbed in 1990, denoted N', is less than the number of cells that are non-disturbed in 1920, denoted N. Hence, the number of cells that are candidates for post-1990 disturbance is less than the number of cells that are candidates for post-1920 disturbance. Otherwise the calibration techniques for post-1990 prediction are nearly identical to the calibration techniques for post-1920 prediction.

One method is to use logistic regression where the independent variable is slope. The dependent variable is 1 if the cell is disturbed between pre-human and 1990, and 0 if the cell is non-disturbed in 1990. The propensity values are the fitted values for the cells that are non-disturbed in 1990.

The method to create the second propensity for post-1990 disturbance map is to create a map of distance to forest edge of 1990. For the non-disturbed cells of 1990, the larger propensities are closer to 1990 forest edge and smaller propensities are farther from forest edge as described by Equation (7):

$$P'(n') = \frac{\text{distance between cell } n' \text{ and the nearest disturbed cell}}{\text{maximum distance between any cell } n' \text{ and the nearest disturbed cell}} \tag{7}$$

where $P'(n')$ = propensity of cell n' to become disturbed after 1990; n' = index of cell that is non-disturbed in 1990 = 1, 2, . . . , N'; and N' = number of cells that are non-disturbed in 1990.

The propensity for post-1990 disturbance map shows the locations of the relative likelihood of post-1990 disturbance. It says nothing about the quantity of future disturbance, thus it says nothing about the probability of future disturbance. The next section shows how to convert the propensity for post-1990 disturbance map to a probability for post-1990 disturbance map.

2.6 *Extrapolation of post-1990 disturbance*

In a manner similar to the prediction of post-1920 disturbance, each propensity for post-1990 disturbance map is sliced by several thresholds, to generate I bins. The first threshold is one (i.e. $T_0' = 1$) and each subsequent threshold T_i' is such that Equations (8) and (9) hold:

$$\begin{aligned} S_i'(n') &= 1 \quad \text{if} \quad T_i' < P'(n') \\ &= 0 \quad \text{else} \end{aligned} \tag{8}$$

$$\left[\sum_{n'=1}^{N'} S_i'(n')\right] - \left[\sum_{n'=1}^{N'} S_{i-1}'(n')\right] = \frac{N'}{I} \tag{9}$$

where $S_i'(n')$ = predicted disturbance for cell *n* at threshold T_i'; *i* = index of threshold = 0, 1, . . . , *I*; T_i' = threshold *i*; and *I* = number of bins created by the thresholds, which is the same as in Equation (3).

Then Equation (10) assigns Z_i to each cell in bin *i* of the propensity for post-1990 disturbance map. Recall that Z_i is the user's accuracy for the disturbed category of bin *i* obtained from the validation between 1920 and 1990.

$$V(n') = Z_i \quad \text{if} \quad T_i' < P'(n') \le T_{i-1}' \tag{10}$$

After Equation (10) assigns a user's accuracy value, Z_i, to each non-disturbed cell of 1990, we can compute an implied proportion of disturbance using Equation (11):

$$Q_1 = \frac{\sum_{n'=1}^{N'} V(n')}{N'} = \frac{\sum_{i=1}^{I} Z_i}{I} \tag{11}$$

If we were to interpret each Z_i value as a probability, then the proportion of disturbance, Q_1, would be the proportion of expected disturbance. However, Q_1 is the proportion of disturbance in 1920, so it should not necessarily be used to make a statement about the proportion of disturbance predicted in some future landscape, especially when the time of that future landscape is not yet specified.

To create a map of the probability of future disturbance, it is necessary to specify a quantity of future disturbance. This quantity can be specified completely independently of the other components of this analysis. For example, this quantity could come from some completely different scenario model that gives a quantity of disturbance for the entire region. Assume this quantity of future disturbance is a proportion, Q_2, of the number of non-disturbed cells of 1990. Equation (12) gives the necessary adjustment to each value of $V(n')$, in order to construct a map of probability of future disturbance, $W(n')$, where the implied proportion of disturbance is Q_2.

$$\begin{aligned} W(n') &= V(n')\frac{Q_2}{Q_1} \quad \text{if} \quad Q_2 \le Q_1 \\ &= \left[1 - [1 - V(n')]\frac{Q_1}{Q_2}\right] \quad \text{else} \end{aligned} \tag{12}$$

where Q_2 = proportion of non-disturbed cells of 1990 that become disturbed after 1990, specified independently.

Specifically, if $Q_2 < Q_1$, then each probability of disturbance, $V(n')$, is shrunk by the ratio Q_2/Q_1. If $Q_1 < Q_2$, then each probability of disturbance is grown by a technique that is equivalent to shrinking the probability of non-disturbance, $1-V(n')$, by the ratio Q_1/Q_2. The technique of Equation (12) assures that the resulting probability is constrained to the interval [0, 1].

If the validation component of the analysis shows that the prediction of location is equivalent to random, then each Z_i is identical, hence each cell in the final probability map is a constant value, $W(n') = Q_2$. If the validation component of the model shows that the prediction of location is perfect, then each Z_i is either 0 or 1, hence each cell in the final probability map is a value far from Q_2 and much closer to 0 or 1. In our example, we use $Q_2 = 0.5$ for illustration, hence Figure 6 shows a map of probability of disturbance when half of the non-disturbed cells in 1990 become disturbed. Figure 6 shows the prediction of post-1990 disturbance with a level of certainty based on the validation of the prediction of the 1920–1990 disturbance.

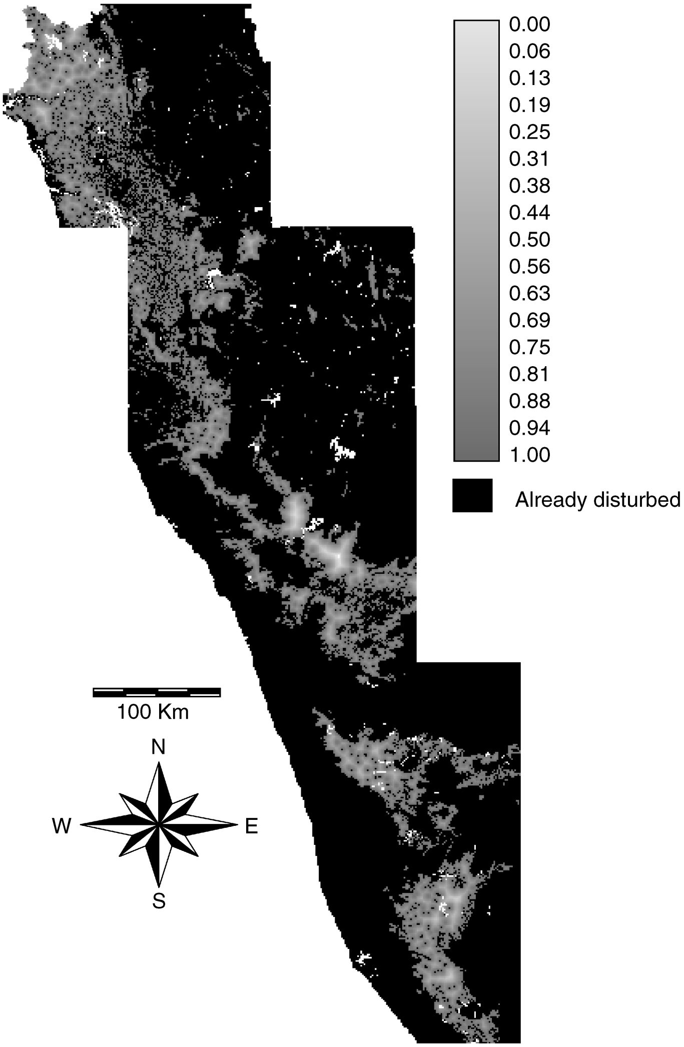

Figure 6 Map of user's accuracy of the newly disturbed category, adjusted to express 50% forest loss from 1990 based on distance to forest edge

3 Results

The most important results concern the range of the probability for disturbance among the cells in Figure 6. The probabilities of post-1990 disturbance range from 0.37 to 0.91 when the propensity for disturbance map is based on distance to forest edge. The probabilities of post-1990 disturbance range from 0.49 to 0.76 when the propensity for disturbance map is based on slope. These results indicate that the predictions of location of disturbance based on distance to forest edge are more certain than the predictions based on slope, since the former method yields a wider range of probabilities.

These results relate directly to Figure 5, which shows that 0.70 is the area under the ROC curve based on the distance to edge variable and 0.62 is the area under the ROC curve based on the slope variable. If the propensity for disturbance values were located at random, then the area under the ROC curve would be 0.5. A perfect ROC value is 1.0. Thus, the measure of the goodness-of-fit of validation for the location of the 1920–1990 disturbance is slightly less than half way between random and perfect.

The ROC ranged from 0.49 to 0.51 among 10,000 Monte Carlo runs, where each run simulated a predicted landscape where the location of the predicted 1920–1990 disturbance was randomized. Therefore, the ROC values of 0.62 and 0.70 are significantly greater than random.

The monotonic logistic regression gives larger fitted values on flatter slopes and smaller fitted values on steeper slopes. Therefore, if we were to use a simple heuristic rule to generate the propensity for disturbance map by putting larger propensity values on flatter slopes, then the results would have been the same as the results from the logistic regression.

4 Discussion

4.1 ROC and Bin Size

Like any statistical analysis, this analysis has its subjective components. The two most subjective components of this analysis are: (1) the selection of the ROC as the criterion to determine the best method to generate the propensity for disturbance map, and (2) the specification of the thresholds, which determines the number of cells in each bin.

There are two reasons why the ROC is well suited to validate predictive models. First, the ROC allows analysis of propensity for disturbance values, where a value has meaning in terms of its order in relation to the other values. ROC is an excellent method for analyzing propensity of disturbance values where there is not a natural interpretation of a value's magnitude, independent of the other values. Second, ROC analyzes the specification of location independently from the specification of any particular quantity of predicted disturbance. ROC accomplishes this by examining the goodness-of-fit of the validation at many thresholds, then aggregating the information at all thresholds into one measure of agreement. When modelers attempt to improve the predictive ability of models, it is helpful to have separate information concerning the goodness-of-fit of location versus the goodness-of-fit of quantity.

The subjective decision concerning the number of bins, hence the number of cells per bin, can influence the results of this type of ROC-based analysis. At an extreme, the smallest number of bins is one, which would be the case if there were only two thresholds, at 1 and $-\infty$. In this extreme case, the analysis could not show useful results, because all cells would be in the same bin, therefore the analysis could not distinguish among cells. At the other extreme, the maximum number of bins is the number of individual cells, which would occur when the number of thresholds equals the number of cells plus one. This would mean that each cell would be its own bin. However, if each cell were its own bin, then each Z_i would be either one or zero. That is, Z_i would be one if the cell were disturbed between time 1 and time 2, and Z_i would be zero if the cell were not disturbed between time 1 and time 2. We use Z_i as the probability that bin *i* will become disturbed after time 2. It is a violation of the concept of prediction to claim that the future state of any bin is guaranteed, especially when each bin is a specific grid cell. Therefore, when the number of cells in each bin is extremely small, the apparent certainty in the results is artificially inflated, since each Z_i would be close to zero or one. Therefore, there should be many cells in each bin.

What do we mean by "many"? In the Western Ghats example, there are 635 cells in each bin. We think this number is more than sufficient for this particular application. As a rule of thumb, we recommend at least 100 cells per bin, so that the proportion (Z_i) will have two significant digits. Ultimately, the decision concerning the number of cells per bin and the number of bins is analogous to the subjective decision a scientist makes when creating bins for a histogram. The size of the bins can influence to a limited extent the spatial distribution of the probability of the category. The size of the bins cannot influence the average probability in the study area because that average probability is determined solely by the proportion of the category in the map. The scientist must use the knowledge of the phenomenon in question to decide the appropriate bin size for the specific application.

4.2 *Quantity of Future Disturbance*

A major strength of this analysis is that it assesses information of location independently from information of quantity. Therefore, it is easy to combine this analysis with other models that focus on specification of quantity only. For example, some models specify the quantity of land required for scenarios of future economic growth (Raskin et al. 1996). The techniques presented in this paper would allow a scientist to create a map of probability of disturbance at specific locations, given any particular quantity of disturbance specified by another model.

Specification of the time when the disturbance will occur can be set independently from the location and quantity of disturbance, since only the specification of the quantity of disturbance is necessary to create a map of probability of future disturbance. If the scientist thinks that the disturbance will occur rapidly, then the future time will occur sooner than if the scientist thinks the disturbance will occur slowly. In our Western Ghats example, Figure 6 portrays a landscape in which 50% of the non-disturbed cells of 1990 are disturbed after 1990. To estimate when this quantity of disturbance will be reached, we interpolate a line using the quantity of non-forest land in 1920 and 1990. Assuming a constant annual amount of additional disturbance in the future, 50% of the non-disturbed land of 1990 will become disturbed by approximately 2030.

The quantity of predicted future disturbance has implications for the selection of the method to create the propensity for disturbance map. Figure 5 shows that distance to forest edge is better than slope at predicting disturbance when the proportion of disturbance is small, since the ROC curve for the distance to edge is higher than the ROC curve for the slope, near the origin. However, when the proportion of predicted future disturbance is large, then slope is slightly better than the distance to edge map at specifying the locations of disturbance between 1920 and 1990. So if the goal is to predict a landscape with a small to medium amount of post-1990 disturbance, then the propensity for disturbance map should be based on distance to edge. If the goal is to predict a landscape with a large proportion of post-1990 disturbance, then the propensity for disturbance map should be based on slope.

4.3 *Maximum Certainty*

The spread in the probabilities reflects the level of certainty in the model's predictive ability. If the computed probabilities are near zero or one, then the level of certainty of location is high. If the computed probabilities are closer to the proportion of predicted disturbance, then the level of certainty of location is lower.

The amount of total certainty in the prediction is a combination of the certainty associated with the specification of location and the certainty associated with the specification of quantity. This paper examines only the certainty in the specification of location. The next important step in the development of this methodology is to combine the certainty of the predicted location with the certainty of the predicted quantity. As a consequence, the predicted probabilities of Figure 6 reflect a maximum level of certainty, since they fail to consider the certainty in the prediction of the quantity of disturbance.

There is an additional reason why the computed probabilities of Figure 6 reflect the maximum level of certainty that the decision-maker should have in the model's predictive ability. The technique assumes that the model's predictive ability between time 1 and time 2 will be the same as its predictive ability beyond time 2. This should be true if the basic mechanisms of disturbance between time 1 and time 2 are the same as the mechanisms beyond time 2. For example, between 1920 and 1990, the technology in the Western Ghats was such that proximity to forest edge was important in deciding which forested locations to disturb. However, the change in future technology may change the predictive ability of the distance to forest edge variable. If the change in future technology makes distance to edge even more important than it was in the past, then the methodology will underestimate the level of certainty. If the change in future technology makes distance to edge less important than it was in the past, then the methodology will overestimate the level of certainty. To err on the side of caution, we should interpret Figure 6 as showing the maximum level of certainty that we should have in the prediction. Decision-makers should find the information about this upper bound on the certainty helpful, because it will enable them to decide on the maximum level of trust to put in the predictions.

5 Conclusions

We have presented a method whereby a scientist can generate a map that shows the probability of a category appearing at a specific location, given a particular predicted quantity of the category. The required inputs are: (1) maps that show a Boolean categorical variable at times 0, 1 and 2, (2) a technique to create a map that shows the relative propensity for membership in the Boolean category, and (3) a predicted proportion of the category at time 3. The technique shows an upper bound on the certainty that a scientist should have in a predicted spatial distribution.

Acknowledgements

We thank Drs. Kamal Bawa and Shaily Menon for supplying the data for this analysis. Additional help was provided by Pablo Pacheco who was funded by the Center for Integrated Study of the Human Dimensions of Global Change. The Center has been created through a cooperative agreement between the National Science Foundation (SRB-9521914) and Carnegie Mellon University, and has been generously supported by additional grants from the Electric Power Research Institute, the ExxonMobil

Foundation, and the American Petroleum Institute. Clark Labs facilitated this research by incorporating the ROC statistic in the GIS software *Idrisi*®.

References

Congalton R and Green K 1999 *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. New York, Lewis Publishers

Eastman J R, Jin W, Kyem P A K and Toledano J 1995 Raster procedures for multi-criteria/multi-objective decisions. *Photogrammetric Engineering and Remote Sensing* 61: 539–47

Geoghegan J, Villar S C, Klepeis P, Mendoza P M, Ogneva-Himmelberger Y, Chowdhury R R, Turner II B L and Vance C 2001 Modeling tropical deforestation in the southern Yucatan peninsular region: Comparing survey and satellite data. *Agriculture, Ecosystems, and Environment* 85: 25–46

Irwin E G and Geoghegan J 2001 Theory, data, methods: Developing spatially explicit economic models for land use change. *Agriculture, Ecosystems, and Environment* 85: 7–23

Ludeke A K, Maggio R C, and Reid L M 1990 An analysis of anthropogenic deforestation using logistic regression and GIS. *Journal of Environmental Management* 13: 247–59

Mather P M 1999 *Computer Processing of Remotely-Sensed Images* (2nd Edition). New York, John Wiley and Sons

Menon S and Bawa K S 1997 Applications of geographic information systems, remote-sensing, and a landscape ecology approach to biodiversity conservation in the Western Ghats. *Current Science* 73: 134–45

Myers N, Mittermeier R A, Mittermeier C G, Fonseca G A and Kent J 2000 Biodiversity hotspots for conservation priorities. *Nature* 403(6772): 853

Pontius Jr R G and Pacheco P 2003 A multiple resolution ROC statistic to validate a GIS-based model of forest disturbance in the Western Ghats, India 1920–1990. *GeoJournal* 57: in press

Pontius Jr R G and Schneider L 2001 Land-use change model validation by a ROC method. *Agriculture, Ecosystems, and Environment* 85: 239–48

Raskin P, Chadwick M, Jackson T and Leach G 1996 *The Sustainability Transition: Beyond Conventional Development.* Stockholm, Stockholm Environment Institute, POLESTAR Series Report No 1

Rodgers W A and Panwar H S 1988 *Planning Wildlife Protected Area Network in India* (2 volumes). Dehra Dun, FAO Project FO IND/82/003

Swets J A 1986 Indices of discrimination or diagnostic accuracy: Their ROC's and implied models. *Psychological Bulletin* 99: 100–17

Swets J A 1988 Measuring the accuracy of diagnostic systems. *Science* 240(4857): 1285–93

Veldkamp A and Lambin E F 2001 Predicting land-use change. *Agriculture, Ecosystems, and Environment* 85: 1–6

Wu H and Huffer F W 1997 Modelling the distribution of plant species using the autologistic regression model. *Environmental and Ecological Statistics* 4: 49–64