Landscape Ecol (2014) 29:367–382
DOI 10.1007/s10980-013-9984-8

REPORT

Recommendations for using the relative operating characteristic (ROC)

Robert Gilmore Pontius Jr. · Benoit Parmentier

Received: 23 July 2013 / Accepted: 31 December 2013 / Published online: 28 January 2014
© Springer Science+Business Media Dordrecht 2014

Abstract The relative operating characteristic (ROC) is a widely-used method to measure diagnostic signals including predictions of land changes, species distributions, and ecological niches. The ROC measures the degree to which presence for a Boolean variable is associated with high ranks of an index. The ROC curve plots the rate of true positives versus the rate of false positives obtained from the comparison between the Boolean variable and multiple diagnoses derived from thresholds applied to the index. The area under the ROC curve (AUC) is a summary metric, which is commonly reported and frequently criticized. Our manuscript recommends four improvements in the use and interpretation of the ROC curve and its AUC by: (1) highlighting important threshold points on the ROC curve, (2) interpreting the shape of the ROC curve, (3) defining lower and upper bounds for the AUC, and (4) mapping the density of the presence within each bin of the ROC curve. These recommendations encourage scientists to interpret the rich information that the ROC curve can reveal, in a manner that goes far beyond the potentially misleading AUC. We illustrate the benefit of our recommendations by assessing the prediction of land change in a suburban landscape.

Keywords Accuracy · AUC · Index · Land change · Map · Prediction · ROC · Threshold · Uncertainty

Introduction

Our manuscript offers recommendations for the use and interpretation of the relative operating characteristic (ROC), which is also known as the receiver operating characteristic (Swets 2010). The ROC is a quantitative method to compare a reference Boolean variable versus an index. We use the word "index" because index is general and has meaning in terms of ranking, while other authors have used alternative words such as: activation level, probability, propensity, likelihood, and suitability. An index displays higher values for observations that are deemed more likely for the presence of a Boolean feature. The reference Boolean variable shows presence versus absence of a feature, where each observation is usually coded as 1 for presence and 0 for absence. If the index value for an observation is greater than a threshold, then the observation is diagnosed as presence, otherwise the observation is diagnosed as absence for the particular threshold. Therefore, each threshold produces a binary diagnosis. Each diagnosis produces

Electronic supplementary material The online version of this article (doi:10.1007/s10980-013-9984-8) contains supplementary material, which is available to authorized users.

R. G. Pontius Jr. · B. Parmentier (✉)
Graduate School of Geography, Clark University,
950 Main Street, Worcester, MA 01610-1477, USA
e-mail: benoit.parmentier@gmail.com

R. G. Pontius Jr.
e-mail: rpontius@clarku.edu

a contingency table. Figure 1 shows the format of the contingency table, where H_t, F_t, M_t, and C_t are called respectively hits, false alarms, misses and correct rejections. Some literature uses the terms true positives, false positives, false negatives, and true negatives respectively for these four entries in the contingency table (Fielding and Bell 1997). Hits H_t plus misses M_t is the quantity of the feature according to the reference Boolean variable, i.e. prevalence denoted as P, thus hits plus misses is a constant that does not change with the threshold. Hits H_t plus false alarms F_t is the quantity of the diagnosed presence according to the threshold applied to the index, thus hits plus false alarms increases as the threshold lowers. A ROC curve is obtained by comparing the reference Boolean variable to successive diagnoses, i.e. successive thresholds of the index.

Equations 1 and 2 define the coordinates (X_t, Y_t) for each point t on the ROC curve. Some authors describe X_t as the rate of false positives, while other authors describe X_t as 1 minus specificity, because specificity is defined as $C_t/(F_t + C_t)$. Y_t is known as both the rate of true positives and as sensitivity (Fielding and Bell 1997). Each profession has its own jargon for these terms, which can be easily confused by readers; therefore our figures present each ratio in a manner that indicates clearly its numerator and denominator. A ROC curve is drawn by first plotting points with Y_t on the vertical axis versus X_t on horizontal axis, and then joining the successive threshold points by straight line segments.

$$X_t = F_t/(F_t + C_t) \quad (1)$$

$$Y_t = H_t/(H_t + M_t) \quad (2)$$

The ROC technique is used in remote sensing to evaluate classifications (Saatchi et al. 2008), in land change science to validate simulations (Eastman et al. 2005), in atmospheric science to verify forecasts (Jolliffe and Stephenson 2003) and in species distribution modeling to evaluate model outputs (Fielding and Bell 1997). For geographical applications, the Boolean variable and index are frequently spatially explicit raster maps, in which case the observations are pixels in: (1) a Boolean map of presence versus absence, and (2) an index map that indicates relative ranking as derived from a variety of possible algorithms (Pontius and Pacheco 2004; Verburg et al. 2004; Lesschen et al. 2005). It is usually necessary to use a third map that defines the candidate region. The candidate region is a subset of the study area that consists of observations that are candidates to have presence of the Boolean feature. For example, if the goal is to predict the gain of Built land between times 1 and 2, then the set of Non-built pixels at time 1 constitutes the candidate region. The index map can be used to predict the spatial allocation of the presences, where higher ranking index values determine the order of priority in which observations are selected as presence.

The ROC measures the degree to which high ranking index values are concentrated on the observations where the reference Boolean variable shows presence. A frequently-reported metric is the area under the ROC curve, denoted AUC, which can range from 0 to 1, where larger AUCs indicate stronger positive association. AUC is a unitless summary metric that synthesizes the association between the reference Boolean feature and several diagnoses by the index. AUC is widely reported, even when the ROC curve is not shown. In our literature review of 63 papers that reported the AUC, 37 of the papers did not show the ROC curve (see Supplementary materials). This is a concern because the AUC does not communicate the rich information that the entire ROC curve can reveal.

Our manuscript addresses a central question: how can an investigator interpret the entire ROC curve to

		Reference		
		Presence	Absence	Diagnosis_t Total
Diagnosis_t	Presence	H_t	F_t	$H_t + F_t$
	Absence	M_t	C_t	$M_t + C_t$
	Reference Total	$H_t + M_t = P$	$F_t + C_t = 1\text{-}P$	1

Fig. 1 Contingency table that serves as the basis to compute a threshold point on the ROC curve. The units of H_t, F_t, M_t, and C_t are proportions of the candidate region, which sum to 1

gain insights that go beyond interpretation of the AUC? Our manuscript answers this question with an illustration of the prediction of the gain of Built land in a suburban landscape of the USA, while our manuscript's principles apply generally to many professions.

Lobo et al. (2008) give five specific reasons to recommend against using the AUC: "(1) it ignores the predicted probability values and the goodness-of-fit of the model; (2) it summarizes the test performance over regions of the ROC space in which one would rarely operate; (3) it weights omission and commission errors equally; (4) it does not give information about the spatial distribution of model errors; and, most importantly, (5) the total extent to which models are carried out highly influences the rate of well-predicted absences and the AUC scores". These five ideas explain correctly the nature of the AUC metric, while some of these criticisms apply also to other metrics. Our manuscript offers methods to address all these criticisms. To address criticisms 1–3 and 5, we recommend labeling the points on the ROC curve with both the threshold values and the percent of the observations that have an index value greater than the threshold. To address criticisms 2, 4, and 5, we recommend mapping the reference Boolean variable, the index, and the density of presence in each bin of the ROC curve.

Peterson et al. (2008) propose using partial ROC curves to address some of the perceived undesirable properties of the full ROC curve and its AUC. Partial ROC curves are designed to deal with situations when some parts of the curve are more important than other parts of the curve, or when the index values have a highly skewed distribution, such as when a large percent of the candidate region has the same minimum index value. We prefer to address these concerns by using labels for the threshold points on the full ROC curves and interpreting the shape of the full ROC curves. Furthermore, Peterson et al. (2008) propose to use $H_t + F_t$ for the horizontal coordinate, rather than the conventional $F_t/(F_t + C_t)$. For both the conventional ROC and Peterson's proposed modification, the curves fail to reveal the entries in the diagnosed contingency table for each threshold, which is a topic that we address in the "Next steps" section of this manuscript.

Science needs widely-understood standards to express the uncertainty in the AUC and to test whether two AUCs are different in a meaningful manner. Some authors use inferential statistics to define a confidence interval around the ROC curve (Cortes and Mohri 2004; Muñoz and Felicísimo 2004; Mcskassy et al. 2005; Fang et al. 2006). Other investigators have computed confidence intervals and p-values to determine whether one AUC is statistically different than another AUC or different than 0.5 (Cortes and Mohri 2004; Fang et al. 2006; Phillips et al. 2006; Saatchi et al. 2008; Taylor et al. 2008; Robin et al. 2011). The use of inferential statistics to create confidence intervals applies to cases where the uncertainty derives from random sampling. However, the selection of the thresholds is another potentially important source of uncertainty that must be considered when interpreting an AUC, even for applications that do not use sampling. Our manuscript offers a method to consider the uncertainty due to threshold selection when comparing two AUCs.

Some investigators use universal rules to anoint particular levels of AUC as low, good, high, or excellent (Westphal et al. 2003; Rizkalla et al. 2008; Boscolo and Metzger 2009; Evans and Cushman 2009; Dlamini 2010; Lin et al. 2010; Mochizuki and Murakami 2011). These universal rules are not related to any particular research question or case study, and thus damage the clarity of communication among scientists. It is more helpful to establish a baseline that relates to the particular research question and case study. Our manuscript demonstrates how to establish a baseline that is more relevant to each particular case study, compared to a baseline that claims to be universal.

Some scientists desire to report a single metric to describe the diagnostic ability of models, and the AUC of the ROC is frequently the single number reported. This reporting has led to criticism of the use of the ROC and its AUC. Our manuscript provides researchers with advice to address these criticisms by showing how to design and interpret the entire ROC curve.

Case study

The case study focuses on the Plum Island Ecosystems (PIE), which is in Northeastern Massachusetts, USA (Fig. 2). PIE spans 26 towns and covers 1,134 km^2. The landscape has a gentle topography, and is dominated by forest cover and residential use. PIE is an area of crucial importance for watershed management due to its proximity to estuarine ecosystems. Consequently, the US National Science Foundation established PIE in 1998 as a long term ecological research (LTER) site.

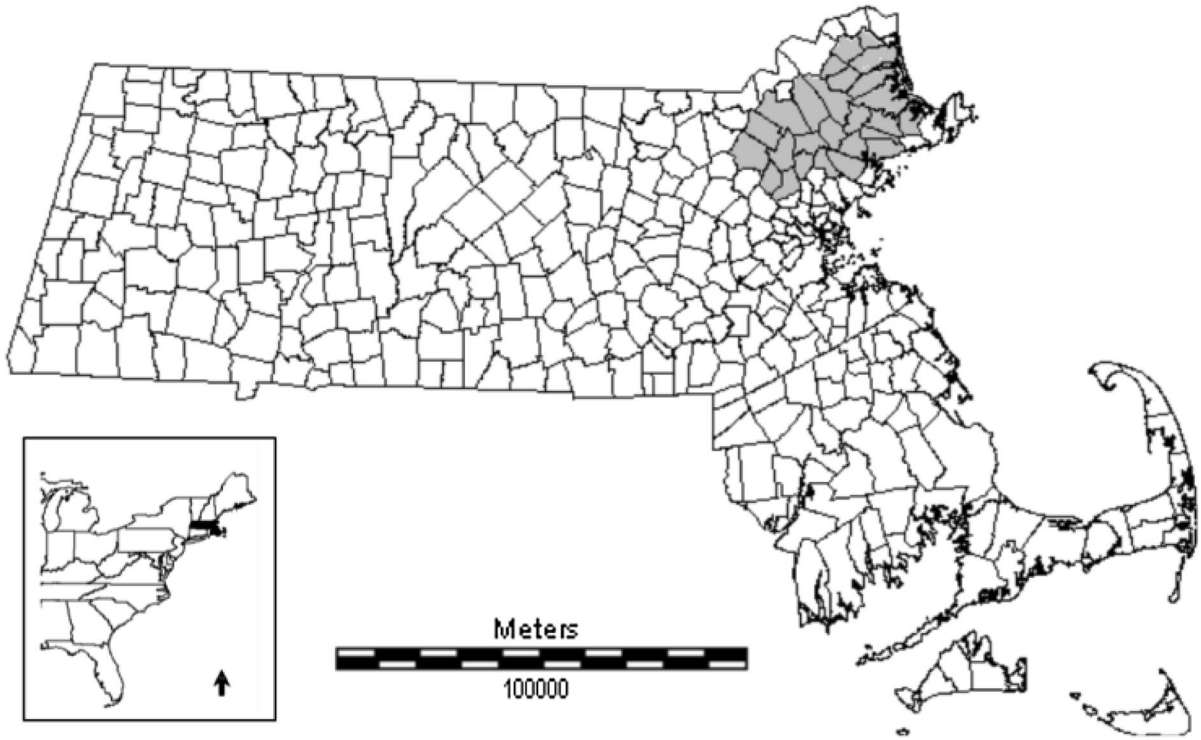

Fig. 2 PIE study area consisting of 26 towns in Northeastern Massachusetts

We obtained spatially-explicit data concerning protection status, landuse, and elevation from the State of Massachusetts (MASSGIS 2010), and then processed the maps using the IDRISI GIS software (Eastman 2012). Figure 3 displays the changes in the Built category during 1985–1999. The pixels that were in the Non-built category at 1985 are candidates for the gain of Built during 1985–1999. This region of Non-built of 1985 is called the candidate region. Maps of unprotected areas, distance to Built at 1985, and slope were used as independent variables to generate three different index maps for post-1985 gain of Built.

Figure 4 shows three maps of indices based on (1) protection status, (2) proximity to 1985 Built, and (3) a logistic regression with topographic slope. We call the three indices (1) unprotected, (2) proximity, and (3) logistic. Index values range from 0 to 1, with larger values indicating a higher priority for post-1985 gain of Built. The unprotected index was obtained by reclassifying information on protection status into protected and unprotected areas, which are represented respectively by the white areas with a value of 0.0 and grey areas with a value of 0.5 (Fig. 4a). The proximity index was developed by assigning index values according to the distance to Built areas at 1985, where smaller index values are assigned to larger distances by using a linear transformation where Built areas of 1985 have a value of 1 and the farthest distance from a Built pixel has a value of 0 (Fig. 4b). The logistic index is obtained by using logistic regression with Built at 1985 as the dependent variable and topographic slope as the independent variable (Fig. 4c). The logistic regression assumes a monotonic relationship and indicates that the stock of Built at 1985 is concentrated on flatter slopes. A naïve index is one that the researcher could make based on an extremely simplistic idea that a reader can understand easily. For our case study, the unprotected and proximity maps are naïve indices because they are based on a single simple idea without any calibration procedure. Naïve indices serve as appropriate baselines to assess other indices that more elaborate procedures create, such as our logistic index. The Methods section describes how each of the three indices are compared to a Boolean variable of the gain of Built from 1985 to 1999, while the Built at 1985 is eliminated from the comparison.

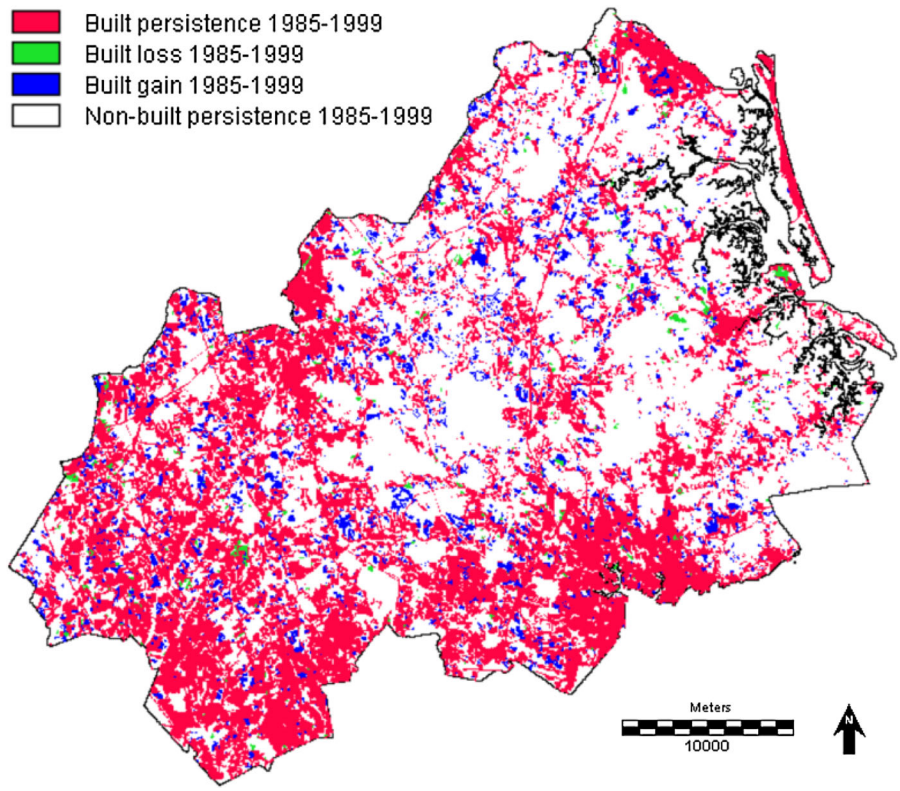

Fig. 3 Changes concerning Built from 1985 to 1999 in the PIE. The candidate region is the union of Built gain and Non-built persistence, which are respectively the presence and absence categories of the reference Boolean variable

It is common that many observations have the same tied index value. Ties introduce stair-shapes in the plot of the cumulative distribution function (CDF) for an index. Figure 5 shows CDFs that plot the percent of the candidate region that is at or below a particular index value versus the index value. The CDF for the unprotected index indicates that 30 % of the candidate region has an index value of 0.0, indicating protected, and the remaining 70 % has a higher index value of 0.5, indicating unprotected. The CDFs for both the proximity and logistic indices indicate that 30 % of the candidate region has an index value less than 0.87. The median index values for the proximity and logistic indices are respectively 0.92 and 0.93.

Methods

To highlight important thresholds

The ROC curve is obtained by slicing in succession an index at various thresholds. This process produces a series of diagnoses that are compared to the reference Boolean variable. The number of diagnoses matches the number of threshold points on the ROC curve. The maximum possible number of unique threshold points on an ROC curve is one plus the number of unique values of the index, where the one additional point derives from the last threshold where all observations are diagnosed as presence. If there are many different index values, then it is common to use an automated procedure to select a number of thresholds that is less than the number of unique index values.

We used three thresholds at 1.00, 0.25, and less than 0 for the unprotected index, because the unprotected index has two unique index values at 0.5 and 0.0. We used only eight thresholds for the proximity and logistic indices, because we want to show later how the number of thresholds influences the AUC. Five of the thresholds were at 1.00, 0.75, 0.50, 0.25, and less than 0. We selected three additional thresholds strategically. We selected a threshold at 8.0 % of the candidate region, which we call the observed quantity threshold, because the Boolean variable shows that 8.0 % of the candidate region gained Built during 1985–1999. We selected a threshold at 8.5 % of the candidate region

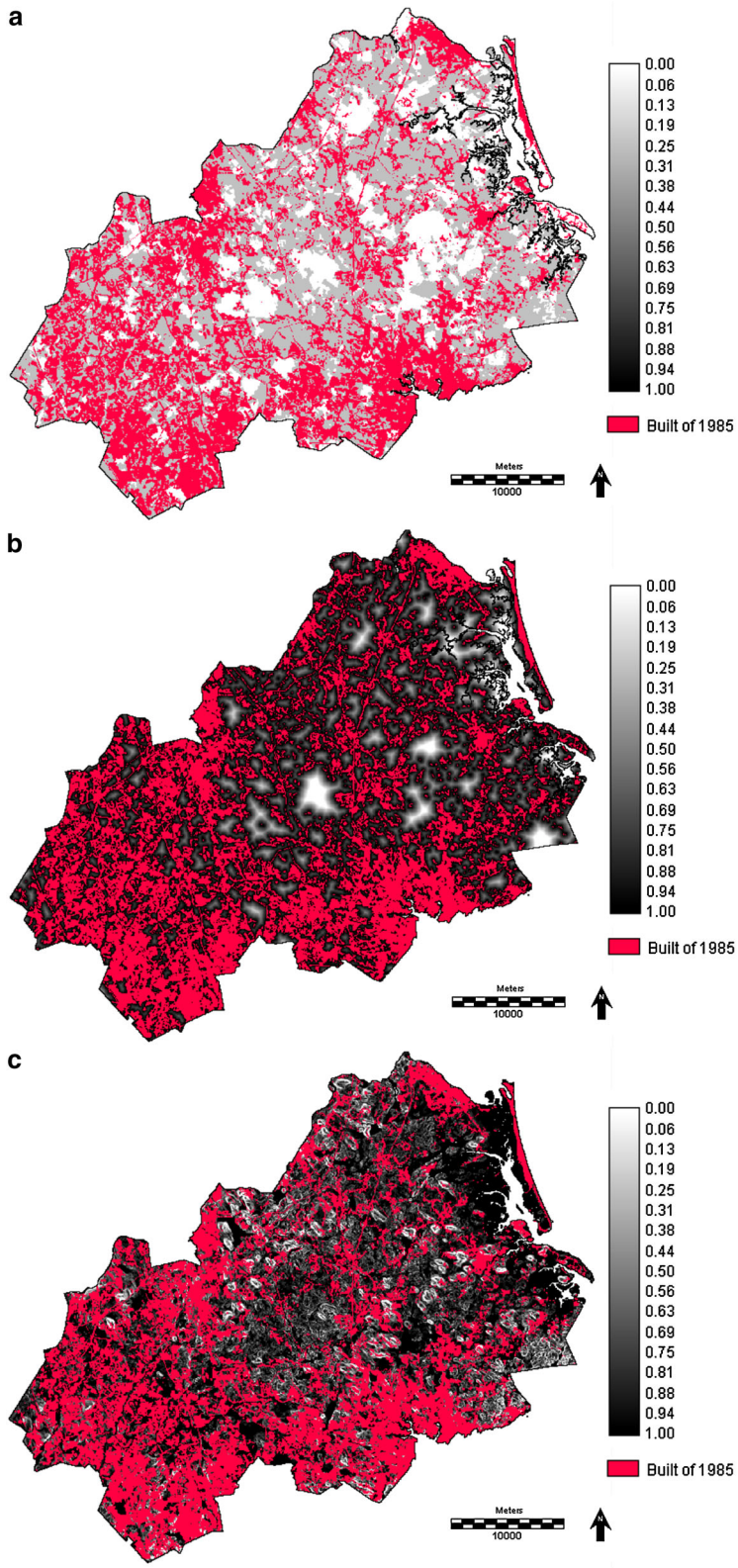

Fig. 4 Maps of indices: **a** unprotected, **b** proximity, and **c** logistic

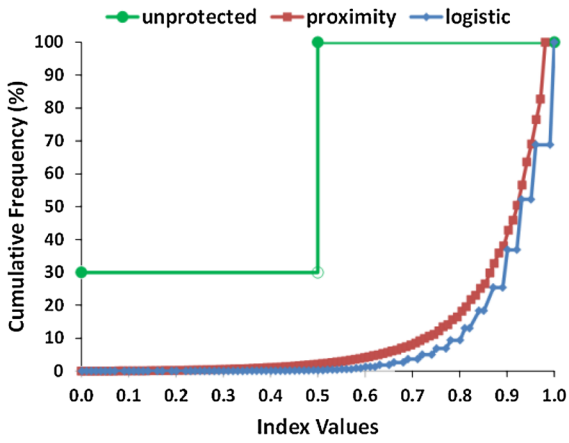

Fig. 5 CDFs for the unprotected, proximity and logistic indices. The *vertical axis* is the percent of the candidate region that has an index value less than or equal to the corresponding value on the *horizontal axis*. The unprotected CDF shows that 30 % of the candidate region has an index value of 0.0, and the remainder of the candidate region has a value of 0.5. Most of the candidate region has index values that lie in the range [0.9, 1.0] for the proximity and logistic indices. The stair shape in the logistic CDF shows that there are fewer unique values in the logistic index compared to the proximity index

because extrapolation from 1971 to 1985 would project that 8.5 % of the candidate region would become Built during 1985–1999. Lastly, we selected a threshold at 70 % of the candidate region because that is the amount of unprotected land in the candidate region. This last threshold corresponds to an index value of 0.87 for the proximity and logistic indices.

Figure 6 presents three ROC curves, with three thresholds for the unprotected index and eight thresholds each for the proximity and logistic indices. We refer to these three ROC curves respectively as unprotected3, proximity8, and logistic8. Figure 6a labels the threshold points to show the index values. Figure 6b shows the same ROC curves, but with labels that indicate the percent of the candidate region that has an index value greater than the threshold. Thresholds that have higher values reside closer to the origin of the curve and portray fewer observations diagnosed as gain of Built. As the value for the threshold becomes smaller, the percent of the candidate region that has an index value greater than the threshold increases. Each pair of successive threshold values defines a bin of observations that relates to a straight-line segment on the ROC curve. Solid segments indicate that the bin contains observations that have

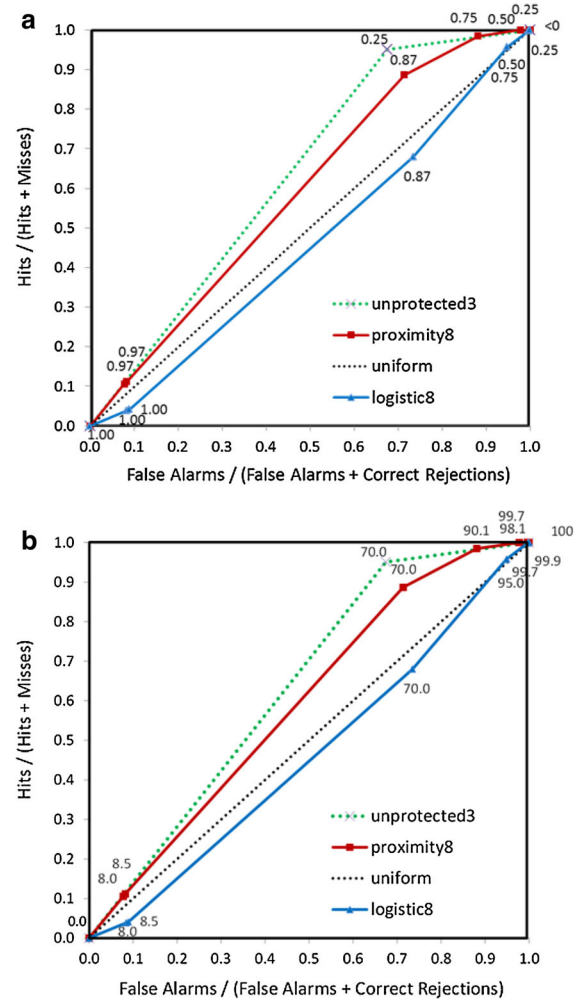

Fig. 6 ROC curves for the uniform, unprotected, proximity and logistic indices. The uniform ROC curve has two thresholds, the unprotected ROC curve has three thresholds, while the proximity and logistic ROC curves have eight thresholds. The labels show **a** the index values at the thresholds and **b** the percent of the candidate region that has an index value greater than the threshold. A *dotted line* segment indicates that all observations within the bin have the same index value

more than one index value, such as in the proximity8 and logistic8 curves. Dotted segments indicate the bin contains observations that are tied with a single index value, such as in the unprotected3 curve.

To interpret the shape of the ROC curve

The shape of the ROC curve informs the user how the index values are associated with the Boolean feature. The ROC curve always begins at the origin (0, 0),

which corresponds to the initial high threshold, for which none of the observations lie above the threshold. Other thresholds are obtained by lowering in succession the threshold, so that more observations lie above the threshold (Fig. 6b). The ROC curve always ends at the point (1, 1), which corresponds to the lowest threshold, for which all observations lie above the threshold. The start of the curve near the origin pertains to thresholds that capture the very highest ranking index values, while the end of the curve near the upper right corner pertains to thresholds that capture the very lowest ranking index values (Fig. 6a). If the presence of the Boolean feature is concentrated on the high ranking index values, then the start of the ROC curve lies close to the vertical axis. Similarly, if the absence of the Boolean feature is concentrated on the low ranking index values, then the end of the ROC curve lies near the horizontal line defined by $Y = 1$. Thus, the shape of the curve between the point (0, 0) and the observed quantity threshold addresses the question "Is the Boolean feature concentrated on high ranked index values?" while the shape of the curve between the observed quantity threshold and the point (1, 1) addresses the question "Is the Boolean feature concentrated on low ranked index values?"

The slope of the line segment that joins two successive thresholds relates to the density of the feature within the bin that the two thresholds bound. A vertical line segment between two successive thresholds indicates that all the observations within the bin are presence, while a horizontal line segment indicates that all the observations within the bin are absence in the Boolean variable. By design, the slopes of segments are never negative and steeper slopes mean denser occurrence of the Boolean feature within the bin.

A perfect index is one for which the highest ranking index values are allocated where the Boolean variable shows presence of the feature. Thus many different indices could conform to this definition of perfect. For a perfect index, the ROC curve corresponds to a sequence of points that starts at (0, 0), proceeds up the vertical axis to the point (0, 1), and then goes right to the point (1, 1). A perfect index has an AUC of 1.

It is common for investigators to establish a baseline for comparison by considering a random index. However, the randomness means that every random index gives a slightly different result. We find it more helpful to establish a baseline by envisioning a uniform index for which all observations are tied with the same index value, thus portrays all observations as equally appropriate. The uniform index does not distinguish any particular observation from any other observation in terms of the index value, thus the uniform index serves as a straight-forward baseline with respect to other indices. The ROC curve for the uniform index follows the diagonal line from the lower left corner (0, 0) to the upper right corner (1, 1). A uniform index has an AUC of 0.5. We recommend using the "uniform" terminology to replace the "no better than random" terminology that is popular in the ecology literature.

To bound the uncertainty of AUC

Selection of the thresholds can influence the shape of the ROC curve, hence its AUC. Theoretically, the largest number of unique thresholds on the curve equals one plus the number of unique index values. In practice, an index might contain a very large number of unique index values, which can make it challenging to write an algorithm to select thresholds such that each bin contains observations that have the same index value, especially when the number of observations is very large. Therefore researchers often select thresholds by taking equal intervals of index values, such as [0, 0.25), [0.25, 0.50), [0.50, 0.75), [0.75, 1.00]. Ties among observations that have the same index value can influence the ROC curve, depending on the relationship between the tied index values and the selected threshold values. Consequently, we examine the uncertainty in the ROC curve that derives from the selection of the thresholds and discuss two cases: (1) an ROC curve for which each bin contains a unique index value, and (2) an ROC curve for which some bins contain various index values.

Each bin contains a unique index value

If each and every bin contains tied observations that have the same unique index value, then the ROC curve is completely certain. In this case, Eq. 3 uses a trapezoidal rule to define exactly the area under the curve, denoted AUCtrapezoidal. Most literature that we have read equates the AUC with the AUCtrapezoidal.

$$\text{AUCtrapezoidal} = \sum_{t=1}^{T-1} \{[X_{t+1} - X_t][Y_t + (Y_{t+1} - Y_t)/2]\} \quad (3)$$

where T = number of thresholds ≥ 2.

Some bins contain various index values

If a single bin contains various index values, then there is uncertainty concerning the ROC curve, because additional thresholds could theoretically further distinguish among the various index values within each bin (Fawcett 2006). For this situation, we use two additional ROC curves called ROClower and ROCupper. ROClower is obtained by joining successive threshold points, first by a horizontal line extending to the right and then by a vertical line extending up. This process produces a stair shaped curve that lies below the ROCtrapezoidal curve. ROCupper is obtained by joining successive threshold points, first by a vertical line extending up and then by a horizontal line extending to the right. This creates a second stair shaped curve that lies above the ROCtrapezoidal curve. From these three curves, i.e. ROClower, ROCtrapezoidal and ROCupper, three AUCvalues are derived respectively called: AUClower, AUCtrapezoidal, and AUCupper. Equations 4 and 5 describe calculations for the AUC of ROClower and ROCupper respectively.

$$\text{AUClower} = \sum_{t=1}^{T-1} \{S_t[X_{t+1} - X_t][Y_t + (Y_{t+1} - Y_t)/2] + [1 - S_t][X_{t+1} - X_t]Y_t\} \quad (4)$$

$$\text{AUCupper} = \sum_{t=1}^{T-1} \{S_t[X_{t+1} - X_t][Y_t + (Y_{t+1} - Y_t)/2] + [1 - S_t][X_{t+1} - X_t]Y_{t+1}\} \quad (5)$$

where S_t is an indicator variable such that $S_t = 1$ when all the observations have the same tied index value within the bin that has a lower threshold of t, and $S_t = 0$ when observations have various non-tied index values within the bin that has a lower threshold of t.

Figure 7 illustrates how the ROClower and ROCupper form rectangles around segments of the ROCtrapezoidal. ROClower and ROCupper constitute the widest possible range for hypothetical ROC curves that could exist if each bin were to contain a unique index value.

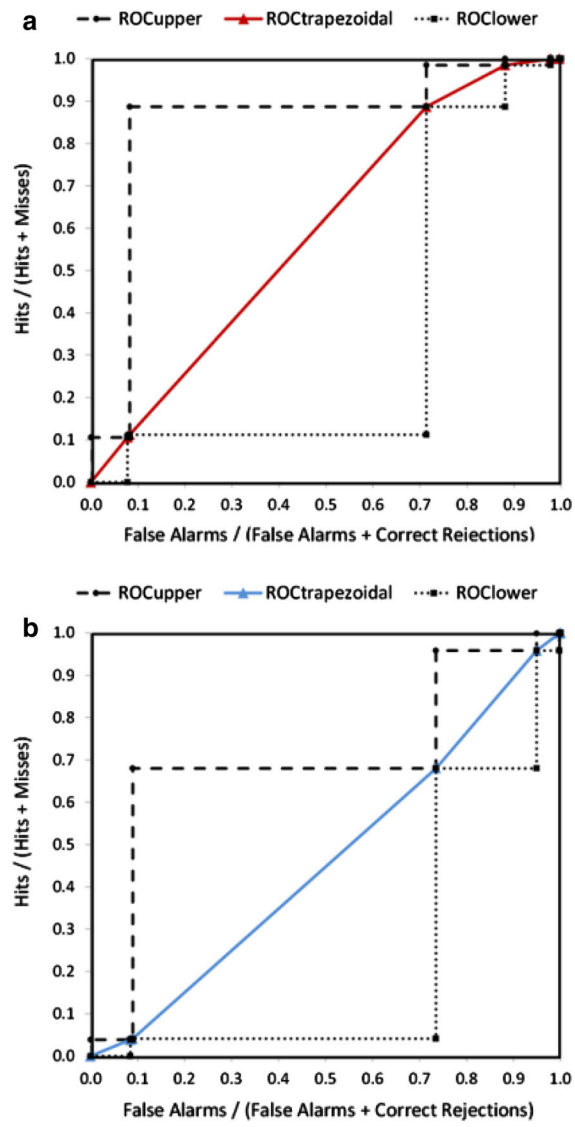

Fig. 7 Uncertainty in ROC curves due to threshold selection for the **a** proximity index and **b** logistic index

Equations 4 and 5 are more general than Eq. 3, because Eqs. 4 and 5 explicitly account for the existence of various index values within each bin. When thresholds are selected such that each bin contains a single unique index value, then the AUClower, AUCtrapezoidal, and AUCupper are identical because the ROClower, ROCtrapezoidal, and ROCupper curves coincide. The dotted lines in Fig. 6 show that this is the case for the unprotected and uniform indices, thus their ROC curves and AUCs are completely certain.

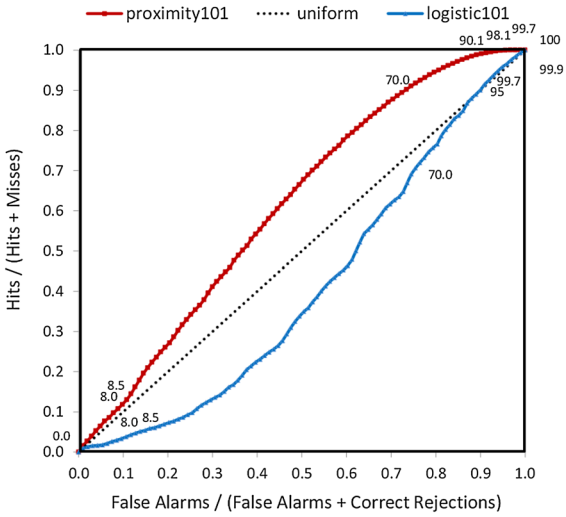

Fig. 8 ROC curves showing 101 thresholds that capture an approximately equal number of observations in each bin for the proximity and logistic indices. The labels on the points show the percent of the candidate region that has an index value greater than the threshold, as in Fig. 6b

In addition to user defined thresholds in Fig. 7, we generated two additional ROC curves for the proximity and logistic indices based on equal interval thresholds. Figure 8 shows such curves using 101 thresholds that define 100 bins, where each bin contains approximately one percent of the candidate region. These two additional ROC curves show how the uncertainty of the AUC shrinks when the number of thresholds increases from 8 to 101 thresholds for the logistic and proximity indices. Later sections of our manuscript refer to these ROC curves respectively as "proximity101" and "logistic101".

To map the ROC information

We propose a novel way to map the information contained in the ROC curve. This is done by calculating the density of the Boolean feature in each bin of the ROC curve. Bin t is defined by a lower threshold t and an upper threshold $t + 1$. Thus the number of bins is equal to the number of thresholds minus one. The map corresponding to each ROC curve shows the density of the presence of the Boolean feature in each bin. Equation 6 expresses the density of the presence of the Boolean feature in the bin that has a lower threshold of t, where Fig. 1 defines H_t and M_t respectively as Hits and Misses.

$$\text{Density of the Boolean feature in bin } t = \frac{(H_{t+1} - H_t)}{(H_{t+1} - H_t) + (M_{t+1} - M_t)} \quad (6)$$

Results

To highlight important thresholds

Figure 6 shows how the threshold labels are crucial for interpretation of the ROC. Figure 6b shows that the threshold at 70 for the unprotected index is above the corresponding thresholds at 70 for the proximity and logistic indices, thus the unprotected index is more accurate than the proximity and logistic indices at predicting the persistence of Non-built. The unprotected ROC curve does not have a threshold that corresponds to the observed quantity of gain of Built, thus the unprotected index is designed to detect where persistence of Non-built occurs rather than where gain of Built occurs. Figure 6b shows that the proximity and logistic curves have thresholds that correspond to the observed quantity of 8.0 and the predicted quantity of 8.5.

Figure 8 shows that it is possible to create many more thresholds for the proximity and logistic indices than for the unprotected index because the proximity and logistic indices have many more unique index values. Thus it is possible to select thresholds for the proximity and logistic indices so that we can test their abilities to predict both the gain of Built near (0, 0) and the persistence of Non-built near (1, 1).

To interpret the shape of the ROC curve

Figure 8 shows how the slopes of various segments of the ROC curve give important information. The proximity ROC curve lies just above the uniform ROC line between (0, 0) and the observed quantity threshold of 8.0 %. This means that the proximity index is slightly better than uniform at predicting the allocation of the gain of Built. The logistic ROC curve is closer to the horizontal axis than the uniform ROC line near (0, 0), which reveals that the highest logistic index values correspond to persistence of Non-built. The part of the proximity ROC curve that extends to the right of the observed quantity threshold resides between the uniform line and the horizontal line of

Y = 1, which means the proximity index is better than uniform at predicting the allocation of persistence of Non-built. Near (1, 1), the logistic ROC curve is on the uniform ROC line, which means that the logistic index is as good as uniform at predicting the persistence of Non-built.

The logistic index was created based on the relationship between topographic slope and the allocation of Built at 1985. At that point in history, Built was concentrated on the flatter slopes, hence the logistic regression reveals a negative relationship between topographic slope and concentration of Built at 1985. However, that relationship inverted during the interval 1985–1999, when humans concentrated the gain of Built on the steeper parts of the candidate region. Thus the AUC of less than 0.5 for the logistic index reveals that the land change process is non-stationary. This interpretation is more helpful than to state that the logistic index performed poorly. Proper interpretation gives insight into the land change process.

To bound the uncertainty of AUC

Figure 9 shows that AUCtrapezoidal equals both AUClower and AUCupper for the unprotected index, because each of the unprotected index's two bins contain observations that are tied with the same index value. Figure 9 shows that the AUCtrapezoidal is 0.60 for proximity8 and is 0.46 for logistic8, but this difference is not meaningful because Fig. 9 shows overlap in the range for AUClower and AUCupper. Specifically, Fig. 9 shows that the AUCs for both proximity8 and logistic8 could exist simultaneously anywhere between 0.34 and 0.70. Additional threshold points narrow the range of AUC for the proximity and logistic indices. There are clear differences among the AUC of 0.64 for unprotected3, 0.61 for proximity101, 0.5 for uniform, and 0.41 for logistic101.

Figure 9 shows that when the number of thresholds increases from 8 to 101, the AUCtrapezoidal increases for the proximity index and decreases for the logistic index. Figure 8 shows that the ROC curve for the proximity101 is concave down, like an umbrella, in which case a smaller number of thresholds produces a smaller AUCtrapezoidal. In contrast, the ROC curve for the logistic101 is concave up, like a bowl, in which case a smaller number of thresholds produces a larger AUCtrapezoidal.

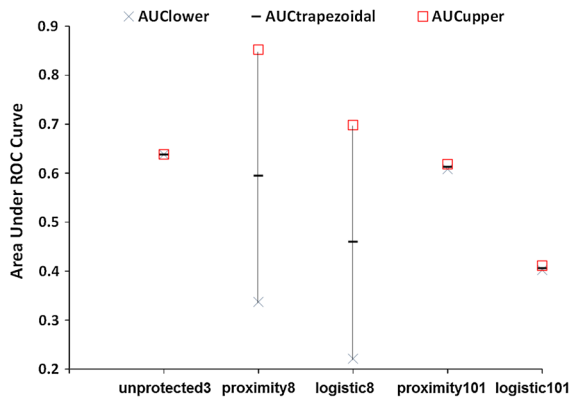

Fig. 9 Uncertainty in the AUC due to threshold selection for the unprotected, proximity, and logistic indices. The intervals [AUClower, AUCupper] overlap for proximity and logistic indices when there are 8 thresholds; however there is no overlap in the intervals for proximity and logistic indices when there are 101 thresholds. AUCtrapezoidal for proximity8 is less than proximity101, while AUCtrapezoidal for logistic8 is greater than for logistic101. AUC values for unprotected3, proximity101 and logistic101 are meaningfully different from 0.5, but the bounds for proximity8 and logistic 8 include 0.5

To map the ROC information

Figure 10 shows the density of the gain of Built for the proximity and logistic indices for the seven bins that are defined by eight thresholds. The threshold names of 1.00, 0.87, 0.75, 0.50, 0.25, and <0 denote the index value that defines the threshold. Threshold names of observed and predicted denote thresholds that match the observed and predicted quantity of the gain of Built. The uniform index has only one bin, for which the density of the gain of Built is 8.0 %. A perfectly correct index would have a density of 100 % for high ranking bins and a density of 0 % for low ranking bins. A perfectly incorrect index would have densities of 0 % for high ranking bins and densities of 100 % for low ranking bins. It is important to examine Fig. 10 to see whether the density of gain of Built decreases from the high ranking bins on the left to the low ranking bins on the right. Figure 10 indicates that the proximity index follows this pattern but the logistic index does not. The bins on the left part of Fig. 10 imply that the proximity index predicts the allocation of gain of Built better than the logistic index, while the bins in the right of Fig. 10 imply that the proximity index predicts allocation of persistence of Non-built better than the logistic index.

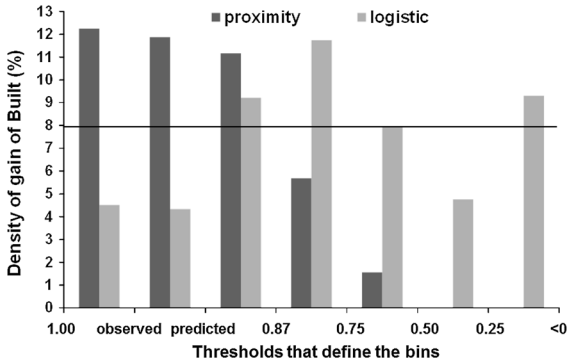

Fig. 10 Density of gain of Built expressed as a percentage in each bin. Seven bins lie between the eight thresholds that the horizontal axis shows. The *horizontal line* at 8 % corresponds to the density within the single uniform index bin. The observed and predicted thresholds have different index values on the proximity and logistic indices

Figure 11a shows a map of the density of gain of Built for the seven proximity index bins, while Fig. 11b shows the density for the seven logistic index bins. If the predictive ability of the index were as accurate as uniform, then all the bins would have the same density. If the predictive ability of the index were perfect, then each bin would have a density of either 0 % or 100 %. The proximity density map has more extreme values compared to the logistic density map.

Discussion

Four groups of recommendations

Many authors have criticized the use of the AUC as a single measurement of goodness of fit. The criticisms are valid because the AUC offers only one number to summarize the wealth of information that the ROC approach can reveal. In some situations, it can be helpful to have one summary metric, such as when comparing thousands of ROC curves in the context of Monte Carlo analysis. However, it can be frustrating for readers to see papers that report only the AUC in situations where a few important ROC curves are computed but not shown. Critics have advised against the use of the AUC, but that advice tells investigators what not to do and does not provide positive guidance. We endorse the use of the ROC, as long as scientists present the ROC in a helpful manner. This subsection offers four groups of recommendations, organized in the next four paragraphs.

First, scientists should highlight important threshold points on the ROC curve. This involves selecting and showing strategic thresholds that are important to the research question. One important threshold is the threshold at which the percent of observations above the threshold matches the prevalence, i.e. the percent of observations of presence in the reference Boolean variable. Figures 6, 7 and 8 show this threshold of the correct quantity of prevalence. In our manuscript's case study, it is relevant to include also the thresholds that correspond to the quantity of predicted gain of Built and the quantity of unprotected area. The CDF can be helpful in selecting the thresholds strategically (Fig. 5). Furthermore, scientists should label important threshold points on the ROC curve with index values and/or the percent of observations that have an index value greater than the threshold. It is not necessary to label all points, but the reader usually needs to know some important points as in Fig. 8. This is especially important when the index values are predicted probabilities, in which case it is important to know how the probabilities relate to the prevalence.

Second, scientists should show and interpret the shape of the ROC curve, especially near the curve's lower left and its upper right. The lower left indicates the association between the high ranking index values and the Boolean feature, while the upper right indicates the association between the low ranking index values and the Boolean feature. The interpretation should tell the reader the part of the curve that is most important for the applied research question. If the goal is to predict the presence of rare events, then the lower left of the ROC curve is likely to be most important. The upper right of the curve is usually more important than the lower left for ecological niche models that use presence only data, because the upper right considers the thresholds that capture all confirmed observations of the species (Peterson et al. 2008). The point at which the ROC curve meets the $Y = 1$ line can be important for ecological niche models that use data with many more pseudo absences than presences. When plotting the ROC curve, scientists should mark each threshold on the ROC curve with a symbol and draw a straight line segment between each pair of symbols. The number of symbols communicates the number of thresholds, which is essential for the reader to understand. The sharp corners at the threshold points on the ROC curve indicate important information concerning where the

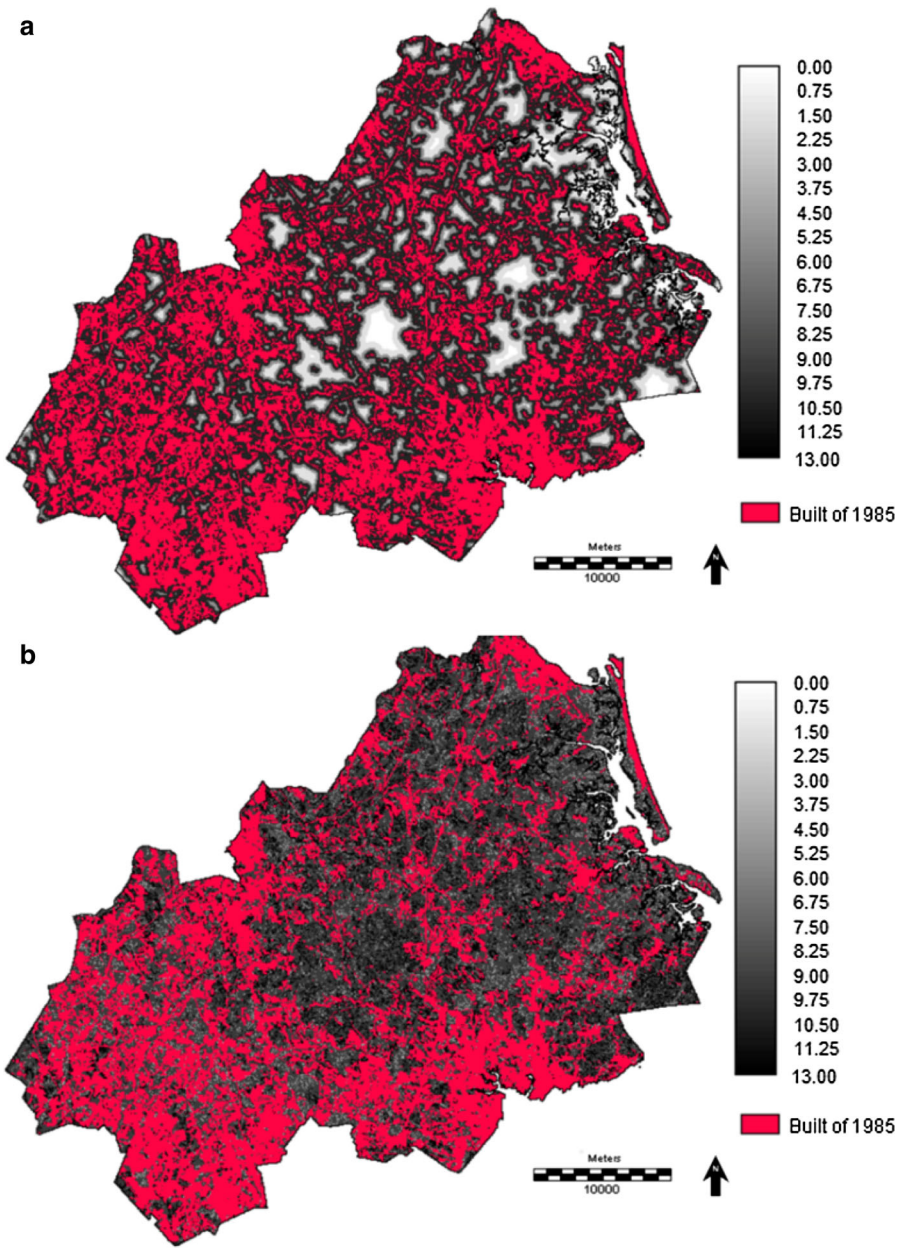

Fig. 11 Maps of density of presence expressed as a percent of gain of Built in seven bins of the ROC curve for the **a** proximity index and **b** logistic index

thresholds reside. Scientists should resist the temptation to smooth the ROC curve, which hides important information concerning the number of thresholds. Furthermore, it is helpful to use dotted line segments to denote bins of the ROC curve that have tied index values and solid line segments to denote bins that have various index values.

Third, scientists should define lower and upper bounds for the AUC and compare the uncertainty of the ROC curve for the index of primary interest to a baseline ROC curve for a naïve index, especially when considering whether two AUCs are meaningfully different. Our manuscript includes the unprotected and proximity indices as naïve indices, in addition to the conventional uniform index. Naïve indices serve as appropriate baselines for comparison versus alternative methods to create the index, such as logistic regression. The uniform ROC line offers one possible baseline, which is the expected ROC curve from a random index. However, randomness does not

necessarily serve as a helpful baseline, because investigators usually already know that the pattern is not random. A naïve index should be specific to the particular case study. Investigators should resist the temptation to use universal baselines and universal rules to crown particular AUC values as acceptable or good. Universal rules are not useful because universal rules are designed for neither a particular research question nor a particular study area. Ultimately, the goal should be to interpret the AUC in order to shed light on the phenomenon of interest, regardless of the value of AUC. In our case study of land change, an AUC of less than 0.5 revealed that the calibrated relationship between slope and Built reversed during the validation interval. For other situations, an AUC of less than 0.5 can reveal helpful information, because an AUC of less than 0.5 reveals that the presences in the validation data are systematically concentrated on lower ranking index values. There could be a variety of reasons for this, such as systematic differences between a model's calibration and validation data in terms of regions, time intervals, or sampling schemes. In any case, if authors follow our recommendations, then it is probably not even necessary to compute the AUC, because the AUC shows much less information than a clearly plotted ROC curve. If authors choose to publish the AUC, then they should report the AUClower, AUC trapezoidal, and AUCupper for both the index of interest and a naïve index. Investigators should use the AUClower and AUCupper to report whether the AUC of primary interest is meaningfully different than the AUC from a naïve index.

Fourth, scientists should examine the density of the presence in each bin, and then show a map of the densities for cases where the map shows important spatial information. The map might reveal spatial patterns that the ROC curve does not. If the patterns are relevant to the research question, then authors should interpret and publish the map.

Next steps

One of the next important steps is to produce software to compute the ROC according to the recommendations in this manuscript. Mas et al. (2013) offer a good start to compute ROC within the land change modeling system Dinamica. We plan to create an additional computer program to select the minimum number of optimal thresholds to estimate AUC to within a specified precision as defined by AUCupper minus AUC lower. Also, we must design software to compute the total operating characteristic (TOC). Pontius and Si (2014) have proposed the TOC to replace the ROC, because the TOC shows all four entries of each two-by-two contingency table for each threshold, therefore the TOC shows strictly more information than the ROC while using the same size of space in a figure. The information in the TOC can be used to construct the ROC, but not vice versa. Pontius and Si (2014) also show how to convert a sample ROC curve to an estimated population TOC curve for cases when the prevalence in the sample reference data is different than the prevalence in the population, such as when the reference data derive from presence only information, while the prevalence in the population is unknown but might be estimated (Elith et al. 2011). The TOC is a step in the direction of showing summary metrics that have the same units as the numbers in the contingency table (Fig. 1). It is desirable that graphical methods show information that allows readers to understand the numbers in the contingency tables as clearly as possible, which is what the TOC does. We have begun to replace other popular unitless metrics with alternatives that express results in terms of the units of the numbers in the contingency table. For example, kappa is a popular indicator of accuracy, but kappa is difficult to interpret for several reasons, one of which is that kappa is a unitless index, just as AUC is a unitless index. Pontius and Millones (2011) have proposed to replace kappa with components of quantity disagreement and allocation disagreement, which have the same units of as the numbers in the contingency tables, for example square kilometers or percent of the study area.

Conclusions

Several authors have warned against the use of the AUC of the ROC as a single measurement of accuracy. The warning tells what not to do, but the warning does not give guidance concerning how to proceed. Our manuscript presents four groups of recommendations to improve the use and interpretation of the ROC method of measurement.

First, our manuscript encourages researchers to select and to label strategically important threshold points on the ROC curve using the index values and/or

the percent of the observations that reside above the threshold. Researchers might benefit from using the CDF to select points that are relevant for the particular research question.

Second, our manuscript shows how to interpret the shape of the ROC curve. The slope of the curve near the origin shows whether high ranking index values are concentrated on the Boolean feature; and the slope of the curve near the upper right show whether low ranking index values are concentrated on the Boolean feature.

Third, we recommend computing AUClower and AUCupper to determine whether an AUC from one index is different than the AUC from a naïve index that represents some meaningful baseline. More thresholds tend to make AUClower closer to AUCupper. The maximum number of potentially meaningful thresholds is one plus the number of unique values of the index. If each bin contains a unique index value, then the ROClower and ROCupper are identical.

Fourth, we introduced the ROC map to visualize the content of the ROC curve by mapping the density of the feature's presence in each bin. We also show a plot of the distribution of density within each bin to show any trends between index bins and the Boolean feature.

These concepts have given insight to our case study concerning the prediction of the spatial allocation of gain of Built land in a suburban landscape. Results reveal a naïve index that distinguishes protected areas from unprotected areas is better than any of the other indices at predicting where the persistence of Non-built occurs. However, this unprotected index is not designed to predict the 8 % of the candidate region where the gain of Built occurs. If eight thresholds determine the ROC curve for the proximity and logistic indices, then their AUCs are indistinguishable from each other and from the unprotected and uniform baselines, given the uncertainty due to threshold selection. If 101 thresholds determine the ROC curve for the proximity index, then its AUC is certainly greater than 0.5 because the persistence of non-Built is concentrated relatively far from existing Built. If 101 thresholds determine the ROC curve for the logistic index, then its AUC is certainly less than 0.5 because the gain of Built is relatively sparse on the flatter of the candidate slopes, which indicates that the calibrated pattern of change is opposite the pattern of change during the validation interval.

We agree with many critics that the AUC should not be the single measurement of agreement between a Boolean variable and an index. The ROC offers many insights that are frequently overlooked and unreported. We hope our manuscript's insights will help researchers to gain a deeper appreciation of the information that the full ROC curve can communicate.

Acknowledgments The United States National Science Foundation (NSF) supported this work via three of its programs: LTER via grant OEC-1238212, Coupled Natural Human Systems via grant BCS-0709685, and Research Experiences for Undergraduates via grant 0849985. Any opinions, findings, conclusions, or recommendation expressed in our manuscript are those of the authors and do not necessarily reflect those of the NSF. Massachusetts' Office of Geographic Information (MassGIS) supplied data for this project. Clark Labs facilitated this work by creating the GIS software Idrisi®. Anonymous reviewers provided constructive feedback that improved our manuscript.

References

Boscolo D, Metzger JP (2009) Is bird incidence in Atlantic forest fragments influenced by landscape patterns at multiple scales? Landscape Ecol 24(7):907–918

Cortes C, Mohri M (2004) Confidence intervals for the area under the ROC curve. In: Saul LK, Weiss Y, Bottou L (eds) Advances in neural information processing systems. A Bratford Book, Cambridge, pp 305–312

Dlamini WM (2010) A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. Environ Model Softw 25(2):199–208

Eastman JR (2012) Idrisi Andes Guide to GIS and Remote Sensing. Clark Labs, Clark University, Worcester

Eastman JR, Van Fossen M, Solorzano LA (2005) Transition potential modeling for land cover change. In: Maguire DJ, Batty M, Goodchild MF (eds) GIS, spatial analysis, and modeling. ESRI Press, Redlands, CA, pp 357–385

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ (2011) A statistical explanation of MaxEnt for ecologists. Divers Distrib 17(1):43–57

Evans JS, Cushman SA (2009) Gradient modeling of conifer species using random forests. Landscape Ecol 24(5):673–683

Fang S, Gertner GZ, Anderson AB (2006) Prediction of multinomial probability of land use change using a bisection decomposition and logistic regression. Landscape Ecol 22(3):419–430

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24(01):38–49

Jolliffe IT, Stephenson DB (2003) Forecast verification a practitioner guide in Atmospheric Science. Wiley, Chichester

Lesschen JP, Verburg PH, Staal SJ (2005) Statistical methods for analysing the spatial dimension of changes in land use and farming systems—LUCC Report Series 7. The International Livestock Research Institute, Nairobi, Kenya and Wageningen University, The Netherlands

Lin Y-P, Chu H-J, Wu C-F, Verburg PH (2010) 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-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Glob Ecol Biogeogr 17(2):145–151

Mas J-F, Soares Filho B, Pontius RG, Farfán Gutiérrez M, Rodrigues H (2013) A suite of tools for ROC analysis of spatial models. ISPRS Int J Geo-Inf 2(3):869–887

MASSGIS (2010) Land use map and topographic information. Executive Office of Environmental Affairs, Boston, MA. http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/. Accessed 10 Nov 2010

Mcskassy SA, Provost F, Rosset S (2005) ROC confidence bands: An Empirical Study. Information Systems Working Papers Series:537-544

Mochizuki S, Murakami T (2011) Change in habitat selection by Japanese macaques (*Macaca fuscata*) and habitat fragmentation analysis using temporal remotely sensed data in Niigata Prefecture, Japan. Int J Appl Earth Obs Geoinf 13(4):562–571

Muñoz J, Felicísimo ÁM (2004) Comparison of statistical methods commonly used in predictive modelling. J Veg Sci 15(2):285–292

Peterson AT, Papeş M, Soberón J (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol Model 213(1):63–72

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190(3–4):231–259

Pontius RG, Millones M (2011) Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int J Remote Sens 32(15):4407–4429

Pontius R, Pacheco P (2004) Calibration and validation of a model of forest disturbance in the Western Ghats, India 1920–1990. GeoJournal 61(4):325–334

Pontius RG, Si K (2014) The Total Operating Characteristics to measure diagnostic ability for multiple thresholds. Int J Geogr Inf Sci. doi:10.1080/13658816.2013.862623

Rizkalla CE, Moore JE, Swihart RK (2008) Modeling patch occupancy: relative performance of ecologically scaled landscape indices. Landscape Ecol 24(1):77–88

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77

Saatchi S, Buermann W, ter Steege H, Mori S, Smith TB (2008) Modeling distribution of Amazonian tree species and diversity using remote sensing measurements. Remote Sens Environ 112(5):2000–2017

Swets JA (2010) Tulips to thresholds. Peninsula Publishing, Los Altos Hills

Taylor RS, Oldland JM, Clarke MF (2008) Edge geometry influences patch-level habitat use by an edge specialist in south-eastern Australia. Landscape Ecol 23(4):377–389

Verburg PH, de Nijs TCM, Ritsema van Eck J, Visser H, de Jong K (2004) A method to analyse neighbourhood characteristics of land use patterns. Comput Environ Urban Syst 28(6):667–690

Westphal MI, Field SA, Tyre AJ, Paton D, Possingham HP (2003) Effects of landscape pattern on bird species distribution in the Mt. Lofty Ranges, South Australia. Landscape Ecol 18(4):413–426