Ecological Applications, 6(1), 1996, pp. 132–139
© 1996 by the Ecological Society of America

THE STATISTICAL EVALUATION OF ECOLOGICAL INDICATORS[1]

PAUL A. MURTAUGH
Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

Abstract. Ecological indicators are easily measured surrogates for underlying properties or responses of a system that are difficult to measure accurately and reliably. Methods from signal detection theory can be used to assess the usefulness of such indicators, based on pilot data sets in which some "gold standard" of the underlying response has been measured. For responses that can naturally be dichotomized (e.g., absent vs. present, or acceptable vs. unacceptable), we can estimate an indicator's *sensitivity* (the probability of a positive indicator, given that the true response is positive) and *specificity* (the probability of a negative indicator, given that the true response is negative). These properties, together with the prevalence of the response in the population being studied, determine the indicator's *predictive value* (e.g., positive predictive value is the probability of a positive response, given that the indicator is positive).

Applications of this methodology are described for two examples: the use of satellite imagery to infer oceanic pigment concentrations, and the use of baseline levels of acid-neutralizing capacity (ANC) to anticipate acidification episodes in lakes and streams.

Key words: binary response; ecological indicator; lake and stream acidification; predictive value; receiver operating characteristic (ROC) curve; remote sensing; sensitivity; signal detection theory; specificity.

INTRODUCTION

Many ecological responses are complex and difficult to measure accurately and reliably. It is tempting to describe such responses in terms of surrogates that are more accessible and easier to measure. For example, increases in cyanobacteria often indicate lake eutrophication (e.g., see Edmondson 1979); sedimentary diatom remains reflect past trophic conditions in lakes (Agbeti 1992); the ratio of total dissolved solids to mean depth has been used to predict fish production in lakes (Ryder 1982); butterfly assemblages are indicators of topographic and moisture gradients (Kremen 1992); and satellite images can be used to estimate oceanic pigment concentrations (Sullivan et al. 1993). Substantial effort is being devoted to the identification and development of indicators of environmental quality for use in monitoring programs, or as response variables in ecological investigations (e.g., see National Research Council 1986, Noss 1990, Messer et al. 1991).

Often the ecological response of interest can be naturally dichotomized. For example, different levels of eutrophication in a lake might represent either acceptable or unacceptable water quality, or concentrations of phytoplankton in the ocean might be classified as blooms if they exceed a certain threshold. For such responses, we can compare the distribution of the indicator among units having a positive response to the distribution among units having a negative response, in order to get an idea of the ability of the indicator to discriminate between the two kinds of units. This approach could be used, for example, to evaluate the association between remote-sensing data and ground-based classifications of semiarid regions as either shrub dieback areas or non-dieback areas (Price et al. 1992), or to assess the usefulness of body size and coloration in distinguishing the sexes of individual birds, known from behavioral observations (Ainley et al. 1985).

I here describe methods from signal detection theory (e.g., see Green and Swets 1966, Swets 1988) that are useful for assessing the accuracy of an indicator in reflecting some underlying, dichotomous response. These methods have been widely used in medical applications—for example, in evaluating tests for conditions that are difficult to diagnose definitively—and many reviews reflect that emphasis (e.g., see Hanley 1989, Begg 1991). I show how these methods can be adapted for ecological applications, and suggest that this framework provides a rigorous standard that should help in the identification of useful surrogates for hard-to-measure ecological responses.

SENSITIVITY, SPECIFICITY, AND ROC CURVES

Theory

Let Y be a random variable taking on the values 0 and 1 for a negative and positive response, respectively, and X be a continuous random variable for the indicator, or marker, of interest. For example, $Y = 1$ might denote unacceptable water quality in a lake, and X might be the lake's Secchi transparency. Suppose we plan eventually to use the marker, X, in the following way: if X takes on a value greater than some cutoff, c, then we will guess that $Y = 1$ for that unit, and, if $X \leq c$, we will guess that $Y = 0$. (If X and Y are negatively as-

[1] Manuscript received 5 July 1994; revised and accepted 3 January 1995.

sociated, as in the lake example, then the new marker, $-X$, can be used as described above).

Assume that the distribution of X depends on the value of Y. In particular, define

$$F(c) \equiv P(X \leq c \mid Y = 0); \qquad \text{and}$$
$$G(c) \equiv P(X \leq c \mid Y = 1). \tag{1}$$

Fig. 1 shows a hypothetical example of forms of the densities of F and G. For an indicator of pollution, for example, F and G might be the distributions of the marker variable in unstressed and stressed environments, respectively (Patil 1991).

Note that there are two situations in which the marker value correctly reflects the underlying value of Y: (1) when $X \leq c$ for an item having $Y = 0$, and (2) when $X > c$ for an item having $Y = 1$. The accuracy of the marker can be summarized in terms of the frequencies of these two situations:

Sensitivity

$$\begin{aligned} &\equiv H(c) \\ &= P(\text{positive marker} \mid \text{true response is positive}) \\ &= P(X > c \mid Y = 1) = 1 - G(c); \end{aligned} \tag{2}$$

Specificity

$$\begin{aligned} &\equiv P(\text{negative marker} \mid \text{true response is negative}) \\ &= P(X \leq c \mid Y = 0) = F(c). \end{aligned} \tag{3}$$

As the cutoff, c, increases (dashed vertical line in Fig. 1 moves to the right), sensitivity decreases and specificity increases. A plot of $H(c)$ vs. $F(c)$—or, more traditionally, $H(c)$ vs. $1 - F(c)$—joining points corresponding to all possible values of c is called a "receiver operating characteristic" (ROC) curve. The ROC curve corresponding to the hypothetical marker depicted in Fig. 1 is shown in Fig. 2. The stronger the association between the marker and the response, the more bowed to the upper right the ROC curve is. The ROC curve expected for a marker having *no* association with the response would be a diagonal line connecting the points (0,1) and (1,0)—that is, $H(c) + F(c) = 1$ for all c.

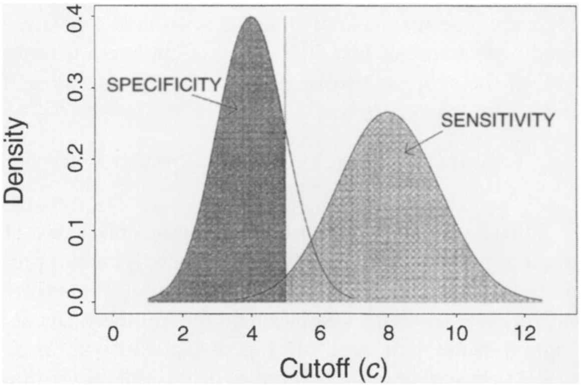


FIG. 1. Hypothetical distributions of indicator values, X, when $Y = 0$ (left) and when $Y = 1$ (right).

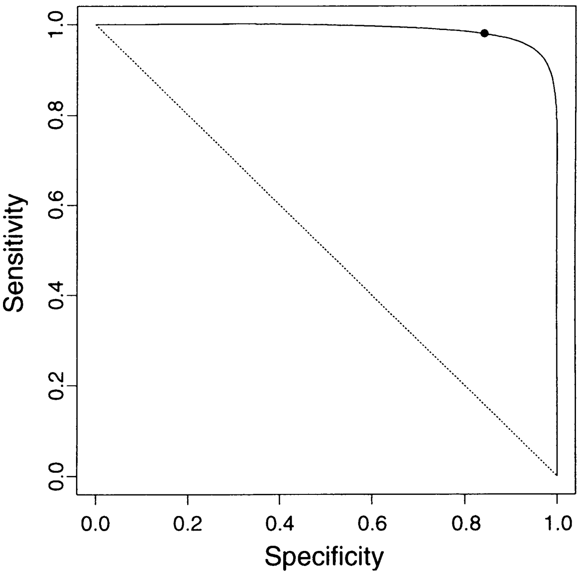
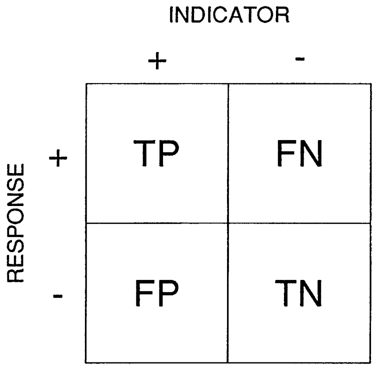


FIG. 2. Receiver operating characteristic (ROC) curve for the hypothetical indicator (marker) shown in Fig. 1. The point on the curve indicates the specificity (0.84) and sensitivity (0.98) corresponding to a marker cutoff of 5 (cf. Fig. 1). The dashed line represents the curve expected for a marker having no association with the response.

Sometimes the response Y is a dichotomization of an inherently continuous random variable. For example, we might define a stressed lake as one having a total chlorophyll concentration exceeding some threshold value. In this case, sensitivity and specificity can be thought of as double integrals of the conditional density of the indicator, given the value of the continuous response (see the Appendix). Only when we are interested in the distribution of the indicator over a *range* of values of the continuous response does it make sense to collapse that response into the binary version used by the ROC approach. This is often the case in environmental monitoring and management, when concern or intervention is triggered only when the ecological indicator exceeds some usual, or "normal," value.

Estimation

Suppose we have marker measurements on a set of N units, and the true response for each unit is known from measurement of a "gold standard" (i.e., the most accurate available measurement of the response). Typically in such studies we try to collect equal-sized random samples of positive and negative units. Then, for any marker cutoff c, each of the N units can be classified into one of four categories, based on its marker value and true response (Fig. 3). If TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively, then we can estimate sensitivity and specificity as



		INDICATOR	
		+	-
RESPONSE	+	TP	FN
	-	FP	TN

FIG. 3. Table for classifying subjects according to their values of X and Y, for a given marker cutoff, c. TP = true positive; TN = true negative; FP = false positive; and FN = false negative.

$$\hat{H}(c) = \frac{\mathrm{TP}(c)}{\mathrm{TP}(c) + \mathrm{FN}(c)}; \tag{4}$$

$$\hat{F}(c) = \frac{\mathrm{TN}(c)}{\mathrm{TN}(c) + \mathrm{FP}(c)}. \tag{5}$$

The variances of $\hat{H}(c)$ and $\hat{F}(c)$ can be estimated from the theory of binomial random variables (e.g., see Metz 1978):

$$\widehat{\mathrm{Var}}\, \hat{H}(c) = \frac{\hat{H}(c)[1 - \hat{H}(c)]}{\mathrm{TP}(c) + \mathrm{FN}(c)}; \tag{6}$$

$$\widehat{\mathrm{Var}}\, \hat{F}(c) = \frac{\hat{F}(c)[1 - \hat{F}(c)]}{\mathrm{TN}(c) + \mathrm{FP}(c)}. \tag{7}$$

Because they are based on separate samples of known positive and negative units, $\hat{H}(c)$ and $\hat{F}(c)$ are statistically independent.

The above estimation methods give the ROC approach a nonparametric flavor, in the sense that the parameters of the underlying distributions of the marker need not be estimated in order to obtain estimates of sensitivity and specificity. Sensitivity, specificity, and their variances can also be estimated under the assumption that the underlying marker distributions, conditional on the response, are Gaussian (e.g., see Greenhouse and Mantel 1950, McNeil and Hanley 1984).

A suitable nonparametric test of the overall association between the marker and response is a Wilcoxon rank-sum test comparing marker values between units with positive and negative responses. A useful summary of the overall accuracy of the marker is the area under the ROC curve, which is expected to be ≈0.5 for a non-informative marker and 1 for a perfect marker (Bamber 1975, Swets 1988). Statistical methodology has been developed to compare these areas for two markers measured under both paired and unpaired designs (Wieand et al. 1989).

Predictive value

Sensitivity and specificity express the probability of observing a particular range of marker or indicator values, X, given a particular value of the underlying response, Y. In future applications of the indicator, however, one will likely want to *predict* the value of Y, based on an observed value of the indicator. Bayes' theorem can be used to express the positive predictive value (PPV) and negative predictive value (NPV) of a marker in terms of its sensitivity and specificity and the overall prevalence of the condition it is supposed to indicate:

$$\begin{aligned}\mathrm{PPV} &= P(Y = 1 \mid X > c)\\ &= [P(X > c \mid Y = 1)P(Y = 1)]\\ &\quad \div [P(X > c \mid Y = 1)P(Y = 1)\\ &\qquad + P(X > c \mid Y = 0)P(Y = 0)]\\ &= \frac{H(c)\cdot p_y}{H(c)\cdot p_y + [1 - F(c)]\cdot(1 - p_y)}; \end{aligned}\tag{8}$$

$$\begin{aligned}\mathrm{NPV} &= P(Y = 0 \mid X \le c)\\ &= \frac{F(c)\cdot(1 - p_y)}{F(c)\cdot(1 - p_y) + [1 - H(c)]\cdot p_y}, \end{aligned}\tag{9}$$

where $p_y \equiv P(Y = 1)$ is the underlying prevalence of the response of interest (e.g., the frequency of lakes with unacceptable water quality in the study region). If the sampling scheme is designed to produce equal numbers of positive and negative units, an estimate of p_y cannot be obtained directly from those numbers, but could be based on previous studies or surveys of the target population. In some applications, the prevalence can be estimated from the sensitivity, specificity, and proportion of units having positive markers in the observed data (Rogan and Gladen 1978).

The dependence of predictive value on the prevalence of the response in the population being studied has important implications for the practical value of an indicator or marker (Gastwirth 1987). For example, a marker that has 99% sensitivity and 99% specificity, but that is used in a population having only a 1% prevalence of the response of interest, will have a positive predictive value of just 50%—that is, on average, only half of the subjects testing positive for the marker will in fact be true positives.

EXAMPLE 1: REMOTE SENSING OF OCEANIC PIGMENTS

Sullivan et al. (1993) compared concentrations of plant pigments (chlorophyll *a* and phaeopigments) predicted from satellite imagery with concentrations from in situ measurements taken around the Southern Ocean. Figs. 4–6 use data read off Fig. 2A of Sullivan et al. (1993), based on the "global processing" algorithm for converting water-leaving radiances into pigment concentrations. These data are circumglobal means of thousands of satellite and in situ observations, calculated for each degree of latitude between 30° and 65° S. The operating characteristics estimated from this analysis, therefore, are not expected to be the same as those for satellite images used to predict pigment concentrations on much smaller scales—for example, in individual parcels of water at specific locations. For a review of accuracy assessment in remote sensing, see Congalton (1991).

There is a strong association between the pigment concentrations measured in situ and those predicted from the satellite imagery (Fig. 4). If we suppose that in practice the satellite data will be used to identify areas of "high" pigment concentration—say, greater than 0.5 mg/m^3—then we can calculate the sensitivity and specificity of the satellite data in indicating high vs. low pigment concentration. Fig. 5 shows the distributions of satellite scores for units (i.e., aggregates of data for individual degrees of latitude) having low and high pigment concentrations, and illustrates how sensitivity and specificity are calculated for a particular satellite score cutoff (0.284 in this illustration).

In this example, there are 12 units for which the in situ pigment concentration exceeds 0.5 mg/m^3. Of those, 10 units have a satellite score exceeding a cutoff of 0.284, so the estimated sensitivity at that cutoff is 10/12 = 0.83. That is, we estimate that 83% of high-pigment units will have satellite scores exceeding 0.284. Ninety-five percent confidence limits for the true sensitivity, based on the exact binomial distribution (e.g., see Rosner 1990:172), are (0.52, 0.98).

Of the 22 units for which the in situ pigment concentration is $\leq$0.5 mg/m^3, 16 units have a satellite score less than or equal to the cutoff of 0.284. The estimated specificity is therefore 16/22 = 0.73, with 95% confidence limits of (0.50, 0.89). That is, we estimate that 73% of low-pigment units will have satellite scores not exceeding 0.284.

When these calculations are done for cutoffs corresponding to all of the observed values of the satellite score, the ROC curve shown in Fig. 6 results. The curve is positively bowed away from the diagonal, suggesting an informative marker. A Wilcoxon rank-sum test comparing the values of the satellite score between the low and high pigment groups confirms that there is a statistically significant association between satellite score and pigment group (P = 0.0001).

If we use as an estimate of the prevalence of high pigment concentration (p_y) the observed prevalence in the data at hand (12/34 = 0.35) and a marker cutoff of 0.284, then we estimate a positive predictive value of 0.62 (Eq. 8) and a negative predictive value of 0.89 (Eq. 9) for the satellite scores. Using p_y estimated from the observed data in this way is equivalent to calculating the positive predictive value as the proportion of positive-testing units having Y = 1, and negative predictive value as the proportion of negative-testing units having Y = 0. This is a sensible approach only if the sampling scheme used to generate the data gives unbiased estimates of the frequencies of units with positive and negative responses, which may be roughly true in this example, given the circumglobal sampling and averaging used to generate the data.

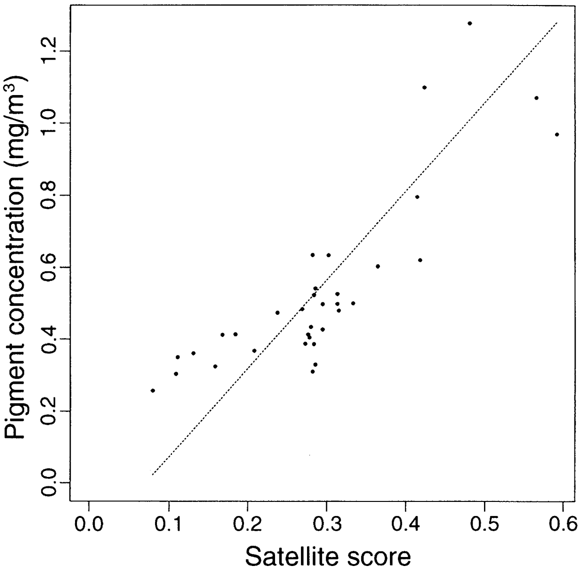


FIG. 4. Scatterplot of in situ pigment concentration vs. concentration predicted from the "global processing" algorithm applied to satellite data, from Fig. 2A of Sullivan et al. (1993). The dashed line is the linear fit presented by Sullivan et al. (1993), from a regression of satellite score against in situ concentration.

EXAMPLE 2: DETECTION OF ACIDIFICATION EPISODES IN LAKES AND STREAMS

These data, compiled by the U.S. Environmental Protection Agency, pertain to episodic acidification of lakes and streams in the northeastern United States. The indicator or marker of interest is "initial" acid-neutralizing capacity (ANC) in μmol$_c$/L—that is, the baseline or low-flow level of ANC in the lake or stream water. This relatively easily measured marker is hoped to be predictive of the depression of ANC (increase in acidity) that occurs during brief, hard-to-capture runoff events. The latter response is summarized as "minimum" ANC. The scientists are not interested in predicting minimum ANC exactly, but rather would like to be able to guess whether minimum ANC will drop below zero, signifying an acidic episode in the lake or stream. The data considered here consist of measurements of initial and minimum ANC in 87 lakes and streams in the mid-Appalachian region of Pennsylvania and the Adirondack Mountains of New York State, from Wigington et al. (1990, 1993).

Fig. 7 shows a scatterplot of minimum ANC vs. initial ANC, and Fig. 8 shows an ROC curve for initial ANC as an indicator of acidification episodes (minimum ANC < 0). For example, if we impose a cutoff of 40 μmol$_c$/L—i.e., if we "guess" that a lake or stream

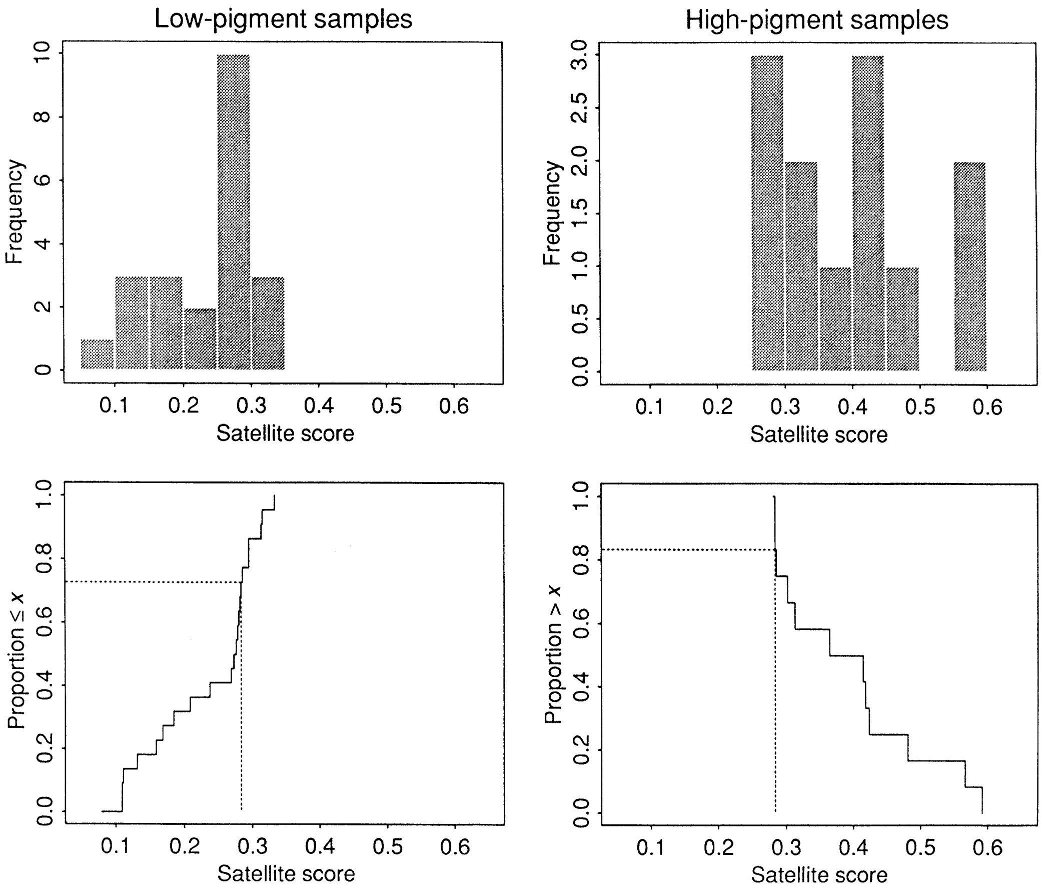


FIG. 5. Distributions of satellite scores for units with in situ pigment concentration ≤0.5 mg/m^3 (left graphs) and for units with in situ concentration >0.5 mg/m^3 (right graphs). Bottom left: proportion of low-pigment units with satellite score not exceeding the cutoff (i.e., the estimated *specificity*) vs. satellite score cutoff. Bottom right: proportion of high-pigment units with satellite score exceeding the cutoff (i.e., the estimated *sensitivity*) vs. satellite score cutoff. Dashed lines indicate the specificity and sensitivity corresponding to a cutoff of 0.284.

with initial ANC <40 μmol$_c$/L will have a negative value of minimum ANC—we find that 41 of 52 lakes or streams with negative minimum ANC are correctly categorized. Thus, the estimated sensitivity is 41/52 = 0.79 (exact 95% confidence interval, 0.65 to 0.89). Twenty-nine of 35 lakes or streams *not* experiencing acidification episodes have initial ANC ≥40 μmol$_c$/L, so the estimated specificity is 29/35 = 0.83 (exact 95% confidence interval, 0.66 to 0.93).

Fig. 9 shows the predictive value of initial ANC as a function of the prevalence of lakes and streams with acidification episodes (minimum ANC < 0), for a cutoff of 40 μmol$_c$/L, calculated according to Eqs. 8 and 9. Positive predictive value (the probability that the minimum ANC is negative, given that initial ANC is <40 μmol$_c$/L) increases with the prevalence of lakes and streams with acidification episodes, and negative predictive value (the probability that the minimum ANC exceeds zero, given that initial ANC is >40 μmol$_c$/L) decreases with increasing prevalence. For example, for a prevalence of 0.2, the positive predictive value is 0.53, and the negative predictive value is 0.94.

DISCUSSION

The concepts of sensitivity, specificity, and predictive value provide a convenient framework for evaluating the usefulness of an easily measured indicator (marker) in reflecting some harder-to-assess underlying response. As pointed out by Patil (1991), a good indicator will be *sensitive* to the underlying condition of interest, and it will be insensitive to other extraneous conditions, i.e., it will be *specific* to the condition of interest.

An advantage of the ROC (receiver operating characteristic) approach is that it is nonparametric, i.e., it is free of assumptions about the mathematical relationship between response and indicator. For instance, if we were to use a linear regression approach in the two examples developed above, we would need to decide on an appropriate parametric model for the apparently curvilinear relationships between the responses and indicators (Figs. 4 and 7), perhaps transforming the response to reduce heterogeneity of variance (Fig. 7). In any case, the details of our inference

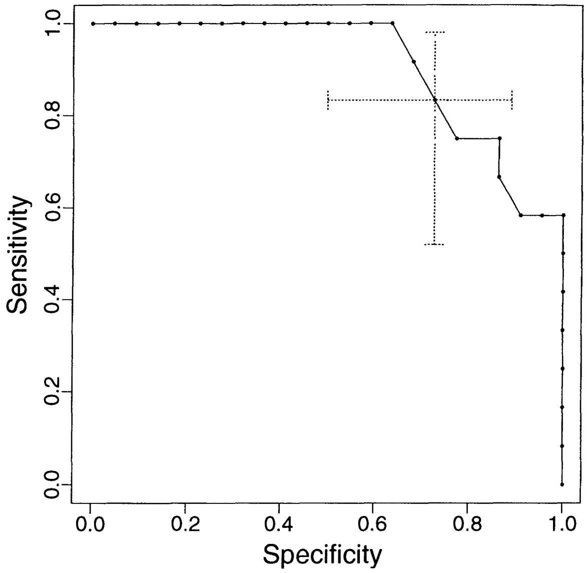


FIG. 6. Estimated ROC curve for satellite score as an indicator of in situ pigment concentration (less than or greater than 0.5 mg/m^3). Error bars are exact 95% confidence intervals for the sensitivity and specificity corresponding to a cutoff of 0.284.

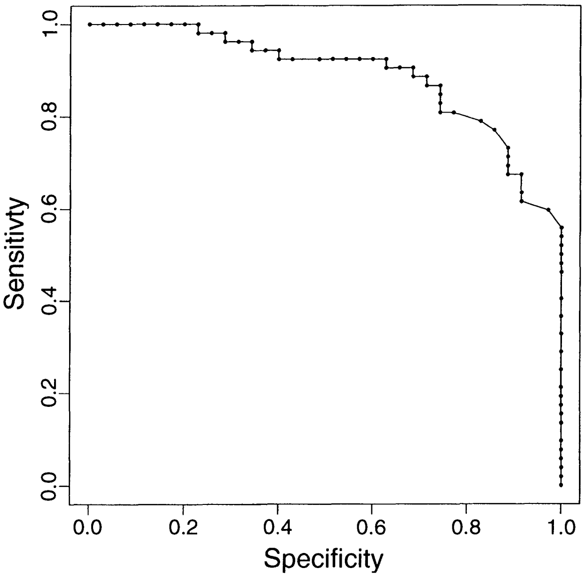


FIG. 8. ROC curve for initial ANC as an indicator of lakes and streams with acidification episodes (minimum ANC < 0).

would depend on the model chosen. A single ROC curve, on the other hand, is obtained for *any* monotonic transformation of the indicator variable.

The ROC approach is best suited for responses that are naturally dichotomous, e.g., the occurrence or nonoccurrence of acidification episodes in Example 2, or the classification of a region as a shrub dieback area or non-dieback area (Price et al. 1992). Collapsing an inherently continuous response into two values will likely sacrifice useful information. If the intent is merely to estimate the indicator value corresponding to a particular response value (or vice versa), regression modeling is the suitable approach. Even for responses that are usually measured on a continuous scale, however, there is often a threshold value above which some kind of action or intervention will be undertaken (Patil 1991), in which case the ROC approach provides a useful summary of the value of a potential indicator.

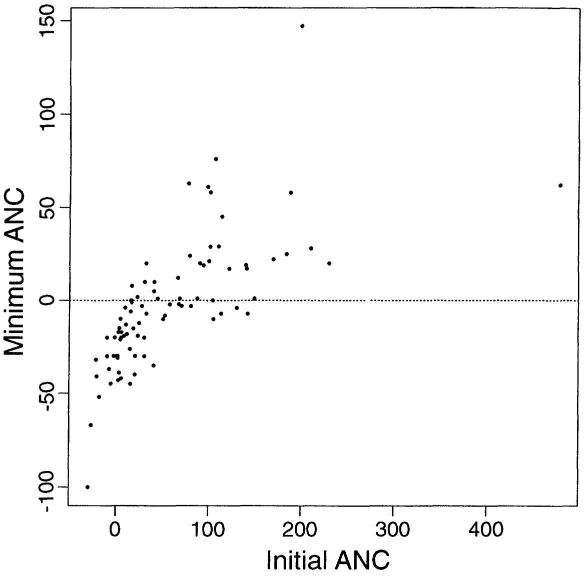


FIG. 7. Minimum vs. initial ANC (acid-neutralizing capacity, in μmol$_c$/L) for 87 lakes and streams in the northeastern United States. Horizontal line shows the threshold below which a lake or stream suffers an acidification episode.

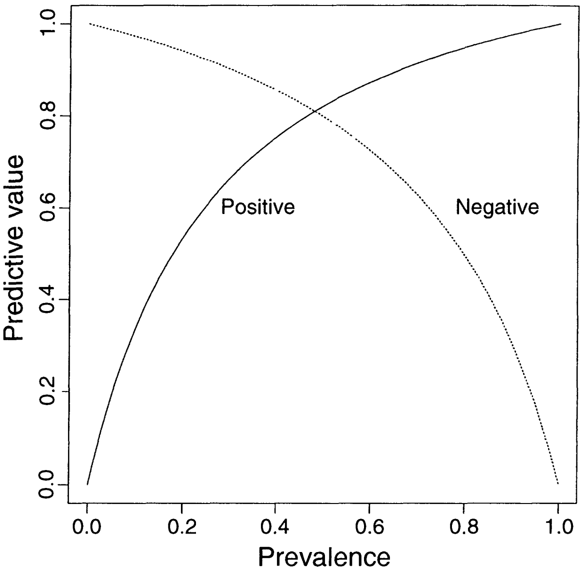


FIG. 9. Positive and negative predictive value of initial ANC, as a function of prevalence of lakes and streams with acidification episodes, for a cutoff of 40 μmol$_c$/L.

The predictive value of an indicator depends on the prevalence of the underlying response in the population being studied, as well as on the indicator's sensitivity and specificity. In environmental monitoring, for example, where observed values of the indicator lead to predictions or guesses about the existence of some condition (e.g., an unacceptable level of pollution), the success rate of the predictions will depend critically on the prevalence of the condition in the environments being monitored. If the condition is rare, then the positive predictive value of the indicator will generally be low, even for indicators that are quite sensitive and specific.

The methodology described here provides information on which decisions about indicators can be based, but it does not unambiguously identify the best indicators or the best ways to use the available indicators. Those judgements must incorporate the relative costs of the different kinds of mistakes that can be made in applications of the indicator: failure to detect a condition where one exists, or apparent detection of a nonexistent condition (Metz 1978). For example, if the event is such that the cost of overlooking it is extraordinary (e.g., an acute pollution episode), we would probably opt for an indicator, or a cutoff for a particular indicator, that yields high sensitivity at the cost of some specificity. The marriage of cost–benefit considerations with the statistical statements available from the ROC analysis—a problem in the domain of decision analysis (e.g., see Raiffa 1970)—should lead to improved protocols for the selection and use of ecological indicators.

An important caveat in the use of any indicator is that an association between the indicator and the response is not necessarily causal, and it does not imply that interventions leading to changes in the indicator will necessarily have any effect on the response. For example, there is increasing evidence that the use of surrogate markers for human disease (e.g., the decline in the number of CD4 cells in AIDS progression) can lead to misleading inferences in clinical trials of therapeutic agents, due to imperfect associations between the markers and key clinical endpoints such as death (Nowak 1994). Indicators must be screened rigorously and quantitatively before they are put forth as meaningful surrogates for the responses we are really interested in.

Acknowledgments

This paper is dedicated to W. T. Edmondson, in appreciation of his unfailing and enthusiastic support through the perambulations of my career. This work was partially supported by an Environmental Protection Agency cooperative agreement with W. Scott Overton (Award Number CR-816721-02-3). I thank J. Van Sickle for making the lake and stream data available and for helpful discussions; D. Pyke and C. Ribic for directing me to other examples of studies for which this methodology could be useful; and two anonymous referees for constructive criticisms.

Literature Cited

Agbeti, M. D. 1992. Relationship between diatom assemblages and trophic variables: a comparison of old and new approaches. Canadian Journal of Fisheries and Aquatic Sciences **49**:1171–1175.

Ainley, D. B., L. B. Spear, and R. C. Wood. 1985. Sexual color and size variation in the South Polar Skua. Condor **87**:427–428.

Bamber, D. 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology **12**:387–415.

Begg, C. B. 1991. Advances in statistical methodology for diagnostic medicine in the 1980's. Statistics in Medicine **10**:1887–1895.

Congalton, R. G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment **37**:35–46.

Edmondson, W. T. 1979. Lake Washington and the predictability of limnological events. Archiv für Hydrobiologie Beiheft Ergebnisse der Limnologie **13**:234–241.

Gastwirth, J. L. 1987. The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data. Statistical Science **2**:213–238.

Green, D. M., and J. A. Swets. 1966. Signal detection theory and psychophysics. John Wiley & Sons, New York, New York, USA.

Greenhouse, S. W., and N. Mantel. 1950. The evaluation of diagnostic tests. Biometrics **6**:399–412.

Hanley, J. A. 1989. Receiver operating characteristic methodology: the state of the art. CRC Critical Reviews in Diagnostic Imaging **29**:307–335.

Kremen, C. 1992. Assessing the indicator properties of species assemblages for natural areas monitoring. Ecological Applications **2**:203–217.

McNeil, B. J., and J. A. Hanley. 1984. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Medical Decision Making **4**:137–150.

Messer, J. J., R. A. Linthurst, and W. S. Overton. 1991. An EPA program for monitoring ecological status and trends. Environmental Monitoring and Assessment **17**:67–78.

Metz, C. E. 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine **8**:283–298.

National Research Council. 1986. Indicator species and biological monitoring. Pages 81–87 *in* National Research Council. Ecological knowledge and environmental problem-solving: concepts and case studies. National Academy Press, Washington, D.C., USA.

Noss, R. F. 1990. Indicators for monitoring biodiversity: a hierarchical approach. Conservation Biology **4**:355–364.

Nowak, R. 1994. Problems in clinical trials go far beyond misconduct. Science **264**:1538–1541.

Patil, G. P. 1991. Encountered data, statistical ecology, environmental statistics, and weighted distribution methods. Environmetrics **2**:377–423.

Price, K. P., D. A. Pyke, and L. Mendes. 1992. Shrub dieback in a semiarid ecosystem: the integration of remote sensing and geographic information systems for detecting vegetation change. Photogrammetric Engineering & Remote Sensing **58**:455–463.

Raiffa, H. 1970. Decision analysis. Introductory lectures on choices under uncertainty. Addison-Wesley, Reading, Massachusetts, USA.

Rogan, W. J., and B. Gladen. 1978. Estimating prevalence from the results of a screening test. Epidemiology **107**:71–76.

Rosner, B. 1990. Fundamentals of biostatistics. PWS-KENT Publishing, Boston, Massachusetts, USA.

Ryder, R. A. 1982. The morphoedaphic index—use, abuse, and fundamental concepts. Transactions of the American Fisheries Society **111**:154–164.

Sullivan, C. W., K. R. Arrigo, C. R. McClain, J. C. Comiso, and J. Firestone. 1993. Distributions of phytoplankton blooms in the Southern Ocean. Science **262**:1832–1837.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. Science **240**:1285–1293.

Wieand, S., M. H. Gail, B. R. James, and K. L. James. 1989. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika **76**:585–592.

Wigington, P. J., Jr., J. P. Baker, D. R. DeWalle, W. A. Kretser, P. S. Murdoch, H. A. Simonin, J. Van Sickle, M. K. McDowell, D. V. Peck, and W. R. Barchet. 1993. Episodic acidification of streams in the Northeastern United States: chemical and biological results of the Episodic Response Project. EPA/600/R-63/190. United States Environmental Protection Agency, Corvallis, Oregon, USA.

Wigington, P. J., Jr., T. D. Davies, M. Tranter, and K. N. Eshleman. 1990. Episodic acidification of surface waters due to acidic deposition. NAPAP SOS/T 12. *In* Acidic deposition: state of science and technology. Volume II. National Acid Precipitation Assessment Program, Washington, D.C., USA.

APPENDIX

Assume the binary response of interest, Y^*, is a dichotomization of some underlying continuous response, Y, such that

$$Y^* = \begin{cases} 1 & \text{if } Y > c_y \\ 0 & \text{if } Y \leq c_y, \end{cases}$$

where c_y is some response cutoff (e.g., a threshold between acceptable and unacceptable values of Y). If c_x is the indicator or marker cutoff above which we will "guess" that $Y^* = 1$, then we can write the sensitivity, H, and specificity, F, as functions of both cutoffs:

$$H(c_x, c_y) = P(X > c_x \mid Y^* = 1) = P(X > c_x \mid Y > c_y)$$
$$= \int_{c_y}^{\infty} \int_{c_x}^{\infty} f_{X|Y}(u \mid v)\, du\, dv;$$

$$F(c_x, c_y) = P(X \leq c_x \mid Y^* = 0) = P(X \leq c_x \mid Y \leq c_y)$$
$$= \int_{-\infty}^{c_y} \int_{-\infty}^{c_x} f_{X|Y}(u \mid v)\, du\, dv,$$

where $f_{X|Y}(\cdot)$ is the conditional density of the marker variable, given a value of the response.

This formulation allows us to explore the effect of the response cutoff, c_y, on the form of the receiver operating characteristic (ROC) curve [$H(c_x, c_y)$ vs. $F(c_x, c_y)$ for varying c_x]. For example, Fig. A1 shows the ROC *surface* for satellite score as an indicator of oceanic chlorophyll concentration (Example 1), when "high" chlorophyll is variously defined as in situ concentration exceeding values from 0.40 to 0.60 mg/m^3. This sort of graph illustrates the consequences of the choice of response cutoff on the ROC curve, but, it must be emphasized, it is *not* a tool for selecting that cutoff. The judgement of what constitutes a "high" response must be based on the subject matter of the problem at hand, not on the appearance of the ROC curve calculated for that cutoff.

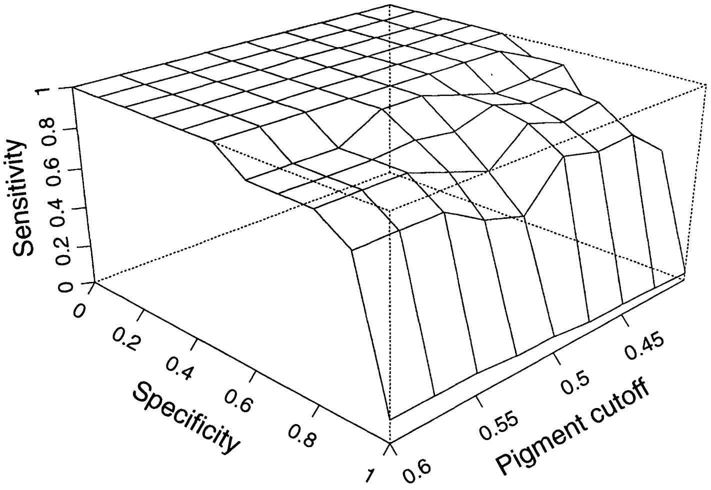


FIG. A1. Estimated ROC (receiver operating characteristic) surface for satellite score as an indicator of in situ pigment concentration, for various pigment cutoffs (i.e., values defining elevated concentrations).