



ACADEMIC
PRESS

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Quaternary Research 60 (2003) 356–367

QUATERNARY
RESEARCH

www.elsevier.com/locate/yqres

A statistical approach to evaluating distance metrics and analog assignments for pollen records

Daniel G. Gavin,^{a,*} W. Wyatt Oswald,^b Eugene R. Wahl,^c and John W. Williams^d

^a Department of Plant Biology, University of Illinois, Urbana, IL 61801, USA

^b College of Forest Resources, University of Washington, Seattle, WA 98195, USA

^c Environmental and Societal Impacts Group, National Center for Atmospheric Research, Boulder, CO 80301, USA

^d National Center for Ecological Analysis and Synthesis, University of California Santa Barbara, Santa Barbara, CA 93101, USA

Received 14 October 2002

Abstract

The modern analog technique typically uses a distance metric to determine the dissimilarity between fossil and modern biological assemblages. Despite this quantitative approach, interpretation of distance metrics is usually qualitative and rules for selection of analogs tend to be *ad hoc*. We present a statistical tool, the receiver operating characteristic (ROC) curve, which provides a framework for identifying analogs from distance metrics. If modern assemblages are placed into groups (e.g., biomes), this method can (1) evaluate the ability of different distance metrics to distinguish among groups, (2) objectively identify thresholds of the distance metric for determining analogs, and (3) compute a likelihood ratio and a Bayesian probability that a modern group is an analog for an unknown (fossil) assemblage. Applied to a set of 1689 modern pollen assemblages from eastern North America classified into eight biomes, ROC analysis confirmed that the squared-chord distance (SCD) outperforms most other distance metrics. The optimal threshold increased when more dissimilar biomes were compared. The probability of an analog vs no-analog result (a likelihood ratio) increased sharply when SCD decreased below the optimal threshold, indicating a nonlinear relationship between SCD and the probability of analog. Probabilities of analog computed for a postglacial pollen record at Tannersville Bog (Pennsylvania, USA) identified transitions between biomes and periods of no analog.

© 2003 Elsevier Science (USA). Published by Elsevier Inc. All rights reserved.

Keywords: Bayesian statistics; eastern North America; modern analog technique; pollen analysis; ROC analysis; squared-chord distance; test accuracy; vegetation type

Introduction

Environmental reconstruction using the modern analog technique (MAT) is accomplished by matching fossil biological assemblages to recently deposited (modern) biological assemblages for which environmental properties are known (e.g., Birks and Gordon, 1985). The relatedness of fossil and modern assemblages is usually measured using a distance metric that rescales multidimensional species assemblages into a single measure of dissimilarity (Guiot, 1990; Overpeck et al., 1985; Prell, 1985; Prentice, 1980). The distance-metric method is widely used among paleoecologists and paleoceanographers because the method is

intuitive, calculations are straightforward, and it requires no assumptions of statistical distributions. However, the interpretation of distance metrics remains largely subjective.

There are two major difficulties in interpreting distance metrics used in the MAT. First, there are no *a priori* criteria to determine whether two assemblages are the “same” (analogs) or “different” (no-analogs) based on the distance (i.e., dissimilarity) between the samples. Most studies have applied a threshold value of the distance metric to determine whether a modern assemblage is an analog. Such thresholds are often obtained, in the case of pollen analysis, by comparing distances between paired modern assemblages within and between vegetation types to demonstrate distances representative of analog or no-analog conditions (Anderson et al., 1989; Davis, 1995; Overpeck et al., 1985). Selection of threshold values is scale-dependent, depending on the eco-

* Corresponding author.

E-mail address: dgavin@life.uiuc.edu (D.G. Gavin).

logical resolution of the vegetation classification and number of variables (taxa) used in computing distances (Calcote, 1998; Sawada et al., 2001). Various methods have also been used to aid interpretation of modern analogs, such as using external information to constrain the modern analog selections and interpreting past climate from a weighted mean of climate observed at the analog sites (Davis et al., 2000; Guiot et al., 1993; Pflaumann et al., 1996; Waelbroeck et al., 1998). Second, the distance metric may not vary linearly with the actual probability that the fossil and modern samples are from the same group. If the distance metric could be transformed to a “probability of analog,” the results of modern analog studies would be more easily interpretable (Liu and Lam, 1985; Robertson et al., 1999).

The goal of this paper is to introduce a statistical tool, the receiver operating characteristic (ROC) curve, to aid the interpretation of distances between fossil and modern assemblages where the modern samples are classified in groups (e.g., vegetation types). ROC analysis provides a framework for assessing the accuracy of diagnostic tests when there are two alternative conditions (e.g., for the MAT: analog vs no-analog). ROC analysis was developed in the field of signal detection and is widely used as a diagnostic tool in medicine to determine the ability of a test to distinguish diseased and nondiseased cases (reviews in Henderson, 1993; Metz, 1978; Zweig and Campbell, 1993). The ROC method has also been used in a wide range of nonmedical applications, from weather forecasting to lie detection tests (Swets, 1988). When applied to the MAT, the ROC method can assess (1) the relative performance of different distance metrics in distinguishing between analog and no-analog cases, (2) the decision threshold for identifying analogs, and (3) using Bayesian statistics, the probability that a modern group is an analog of a fossil assemblage. We explored the first two objectives elsewhere using small data sets (Oswald, in press; Wahl, in press). In this study, we focus on the third objective, determining the probability of analog, using 1689 modern pollen assemblages from eastern North America and a postglacial pollen record from Tannersville Bog, Pennsylvania.

ROC analysis

Background

Low values of a distance metric suggest a high probability that two pollen assemblages are from the same vegetation type (i.e., analogs). Thus, the distance metric (d) is used to distinguish between analog (A+) and no-analog (A-) cases. ROC analysis is conducted on the distributions of d for known analog and no-analog cases (Fig. 1a). If the histograms for analog and no-analog cases have little overlap, d is largely effective at discriminating between analog and no-analog (Fig. 1a). However, such tests are rarely 100% accurate and typically a trade-off exists in the choice

of decision thresholds. For any decision threshold d' there are a corresponding true positive fraction (in Fig. 1a, the fraction of actually positive observations less than d') and true negative fraction (fraction of actually negative observations greater than d'). The true positive fraction (TPF, also referred to as sensitivity) is an estimate of the probability that the distance metric will be below the corresponding decision threshold given an analog case, $\Pr(d < d' | A+)$. The true negative fraction (TNF, also referred to as specificity) is an estimate of the probability that the distance metric will be greater than the threshold given a no-analog (A-) case, $\Pr(d > d' | A-)$. The complement of the TNF is the false positive fraction (FPF); a false positive is also known as a type I error. The complement of the TPF is the false negative fraction (FNF); a false negative is also known as a type II error.

Selection of an optimal decision threshold requires assessment of the trade-off between maximizing TPF or TNF. If equal importance is placed on maximizing TPF and TNF (other weightings could be chosen by the analyst; Zweig and Campbell, 1993), then an overall index for selecting an optimum threshold would be $\text{TPF} + \text{TNF} - 1$, with a possible range from 0 (no ability to discriminate at d') to 1 (perfect discrimination at d'). The optimum d' would be where this index was a maximum (Wahl, in press; Youden, 1950). Decreasing d' below this optimum would decrease TPF faster than the increase in TNF, and increasing d' above this optimum would decrease TNF faster than the increase in TPF (Fig. 1a).

Over a continuum of decision thresholds, TPF and TNF vary inversely in a way that depends on the amount of overlap between the sampled A+ and A- populations. A plot of TPF versus FPF is termed an ROC curve and shows the continuous range of TPF and FPF for all possible decision thresholds (Fig. 1b). The area under the ROC curve, AUC, is a measure of the overall ability of d at discriminating between A+ and A-. If a low d indicates analog (A+) cases (Fig. 1a), then AUC is equivalent to $\Pr(d_{A+} < d_{A-})$, where d_{A+} is a randomly selected case from A+ and d_{A-} is a randomly selected case from A-. Because AUC is not dependent on any particular decision threshold, it is a global measure of the diagnostic performance of a test (Metz, 1978). AUC ranges between 0.5 (ROC curve is a diagonal line; no discrimination by d because the A+ and A- distributions are identical) and 1 (ROC curve follows the left and upper borders of the ROC graph; perfect discrimination by d because the A+ and A- distributions are completely separated). A wide range of test accuracy can be demonstrated by the value of AUC (Figs. 1b and 1e).

Confidence intervals for estimates of AUC (\hat{AUC}) may be used to test the hypotheses that d can discriminate cases better than pure chance ($\hat{AUC} > 0.5$) or whether one test method is significantly better than another method ($\hat{AUC}_1 > \hat{AUC}_2$). Confidence intervals for \hat{AUC} may be computed by a parametric curve fit to the ROC curve (Metz, 1986), by the nonparametric standard error of the Wilcoxon rank-sum

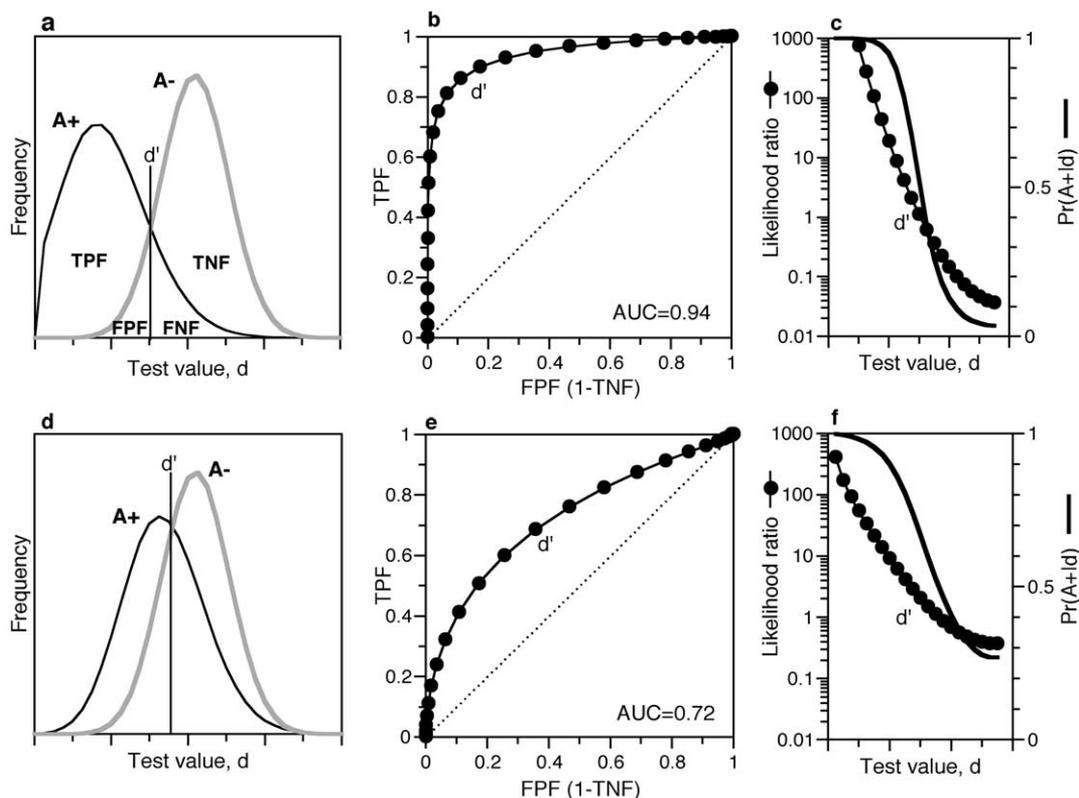


Fig. 1. Schematic showing the steps of ROC analysis. (a) Frequency distributions of test values from actually positive (black curve) and actually negative (gray curve) cases. Using an arbitrarily chosen threshold d' , each observation is truly or falsely classified as positive or negative. The proportion of actually positive cases correctly classified as true is the true positive fraction (TPF). Other fractions shown are the FPF (false positive), TNF (true negative), and FNF (false negative). (b) The corresponding ROC curve shows the relationship between FPF and TPF over the entire range of possible test values. In this example, d' is chosen where TPF-FPF is a maximum. AUC is the area under the curve, an index for the overall accuracy of a test. (c) Likelihood ratio of a positive test result calculated as the slope of the ROC curve over the continuum of test values and posterior probabilities computed from the likelihood ratio and a prior probability of 0.5. (d)–(f) Frequency distributions, ROC curve, and likelihood ratios for a situation where d is less powerful at discriminating positive and negative cases.

statistic (a statistic equivalent to \hat{AUC} ; Hanley and McNeil, 1982), or by resampling methods (Mossman, 1995). In this study we use the Wilcoxon rank-sum statistic because the distributions of A+ and A- frequently appear skewed (breaking the assumption of the parametric methods) and resampling methods were excessively complicated to perform.

Bayesian inference from ROC curves

In assessing a sample with unknown affiliation, the probabilities described by TPF and FPF are not useful to test hypotheses, and resorting to a decision threshold fails to make use of all the information provided by the test result. A more ideal, threshold-independent, measure is the probability of an A+ result given a certain test result, $\Pr(A+|d)$ (Liu and Lam, 1985). ROC analysis provides a way to assess this probability using Bayesian methods. This topic has been concisely reviewed in the medical literature on ROC curves (Henderson, 1993; Zweig and Campbell, 1993); we present a brief summary below.

The slope of the ROC curve at the portion of the curve that encompasses d is a function of the frequency of A+ and A- results with similar d . The slope of the ROC curve may be expressed as $\Pr(d|A+)/\Pr(d|A-)$, or the ratio of the probability of obtaining a certain d in the A+ population over the probability of obtaining d in the A- population. This slope is thus a likelihood ratio of obtaining an A+ result relative to an A- result (Fig. 1c). Note that the value of d where the likelihood ratio equals 1 (the point on the ROC curve tangent to a 1:1 line nearest the upper-left corner) is equivalent to the optimal decision threshold assuming equal priority of maximizing TNF and TPF (Fig. 1a). In Bayesian analysis, this likelihood ratio may be combined with a prior (pre-test) probability for a “positive” result to obtain a posterior (post-test) probability (Hilborn and Mangel, 1997).

Bayes' theorem, in its simplest form, allows the calculation of a conditional probability based on observed simple probabilities from the population,

$$\Pr(A+|d) = [\Pr(d|A+) \Pr(A+)]/\Pr(d) \quad (1)$$

where $\Pr(A+)$ is the probability of an A+ result for the population as a whole (prior probability) and $\Pr(d)$ and $\Pr(d|A+)$ are the probabilities of obtaining a test result d in the population as a whole or in the A+ population, respectively.

A simpler expression of Bayes' theorem is the odds–likelihood ratio form,

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds} \quad (2)$$

where the likelihood ratio in this case is the slope of the ROC curve at the point of the observed test result (d). Prior odds, in this sense, is also a likelihood ratio of the probability of an A+ to an A– result, calculated as $\Pr(A+)/[1 - \Pr(A+)]$. For example, given a prior probability of 0.05, prior odds are $0.05/(1 - 0.05) = 0.0526$. Given a likelihood ratio of 25, posterior odds = 1.315. Posterior odds are converted back to the posterior probability as $\text{odds}/(1 + \text{odds})$, yielding $\Pr(A+|d) = 0.57$.

There are two issues that must be addressed in calculating posterior probabilities. First, there may be few observations in the vicinity of the test result d , making the slope unreliable or incongruous with the ROC curve in general. To deal with this issue, the likelihood ratio may be based on the TPF and FPF observations within a window of d that encompasses a sufficient number of observations. Likelihood ratios also may be computed from a smooth ROC curve based on a parametric curve fit to the observed TPF and FPF values (Metz, 1986). Second, as with many applications of Bayesian statistics, the choice of prior probabilities is difficult and may bear heavily on the posterior probability (Hilborn and Mangle, 1997). Priors may be based on no assumed prior knowledge, on the prevailing knowledge before the study was conducted, or on previous results from the same study.

Application of ROC analysis to the modern analog technique

Rationale

To assess whether two pollen assemblages are from the same vegetation, the distance between the two pollen assemblages (palynological, not geographical, distances computed using a distance metric) may be analyzed using ROC analysis. We assume that distances between modern assemblages within vegetation types describe analog situations and the distances between assemblages in different vegetation types describe no-analog situations (Anderson et al., 1989; Davis, 1995; Wahl, in press). We propose a specific framework for computing the distances within and among vegetation types. Assume a modern pollen assemblage data set representing only two vegetation types (a and b), where type a is to be assessed as an analog based on the distance from fossil pollen assemblages to the nearest analog in type a . Calibration of the distance metric is then based on near-

est-neighbor distances within type a for the “positive” cases (i.e., distances from each pollen assemblage in type a to the most similar assemblage in type a) and nearest-neighbor distances from pollen assemblages in type b to type a for the “negative” cases (Fig. 2a). Distances are similarly calculated to calibrate the distance metric for type b (Fig. 2d). Note that this approach to calibrating distance metrics uses smaller distances for the distribution of actually positive and actually negative cases compared to using all pairwise comparisons. This approach should be more appropriate than using all pairwise comparisons if application of the modern analog technique uses only the nearest analog from fossil assemblages (Overpeck et al., 1985). Also note that the nearest-neighbor distances are not symmetric (Figs. 2b vs 2e), so the ROC curves differ depending on which group is being assessed as the analog (Figs. 2c vs 2f).

Methods

We applied ROC analysis to modern pollen assemblages from eastern North America (Fig. 3). Data were compiled from holdings at the global pollen database (<http://www.ngdc.noaa.gov/paleo/gpd.html>) and Brown University. Sites were classified into one of ten vegetation types derived from the IGBP DISCover land cover classification (Loveland et al., 2000) based on the modal land cover type in a 20×20 km square around each site (Williams and Jackson, 2003). We did not use sites classified as cropland, cropland/natural vegetation mosaics, or pasture, eliminating much of the agricultural area of the Midwest and southern Piedmont, leaving 1689 samples in eight biomes (Fig. 3). Dissimilarity values were calculated from 25 pollen types common in eastern North America (Table 1; Williams et al., 2001). All analyses could be run using spreadsheets and public-domain software for analog analysis (ANALOG, Schweitzer, 1999) and ROC curve analysis (ROCKIT, Metz, 1998).

To examine how well different distance metrics perform at distinguishing biomes, we examined ROC curves based on nine distance metrics (Prentice, 1980; Overpeck et al., 1985) (Table 2). These distance metrics represent three classes regarding how much weight is applied to rare pollen types. Equal-weight metrics (Canberra, Gower's, and standardized Euclidean) standardize the pollen types so that each affects the distance value equally. Unweighted metrics (Manhattan, squared cosine- θ , and Euclidean) do not scale pollen abundance in any way. Signal-to-noise metrics (squared chord, information index, and squared χ^2) moderately increase the contribution of rare pollen types (Prentice, 1980). To evaluate the different metrics, we compared two particularly well-represented biomes (Deciduous Broadleaf Forest and Southern Evergreen/Mixed Forest) by using only nearest-neighbor distances within the Deciduous Broadleaf Forest and from the Southern Evergreen/Mixed Forest to the Deciduous Broadleaf Forest, following the strategy depicted in Fig. 2. For each distance metric, we constructed ROC

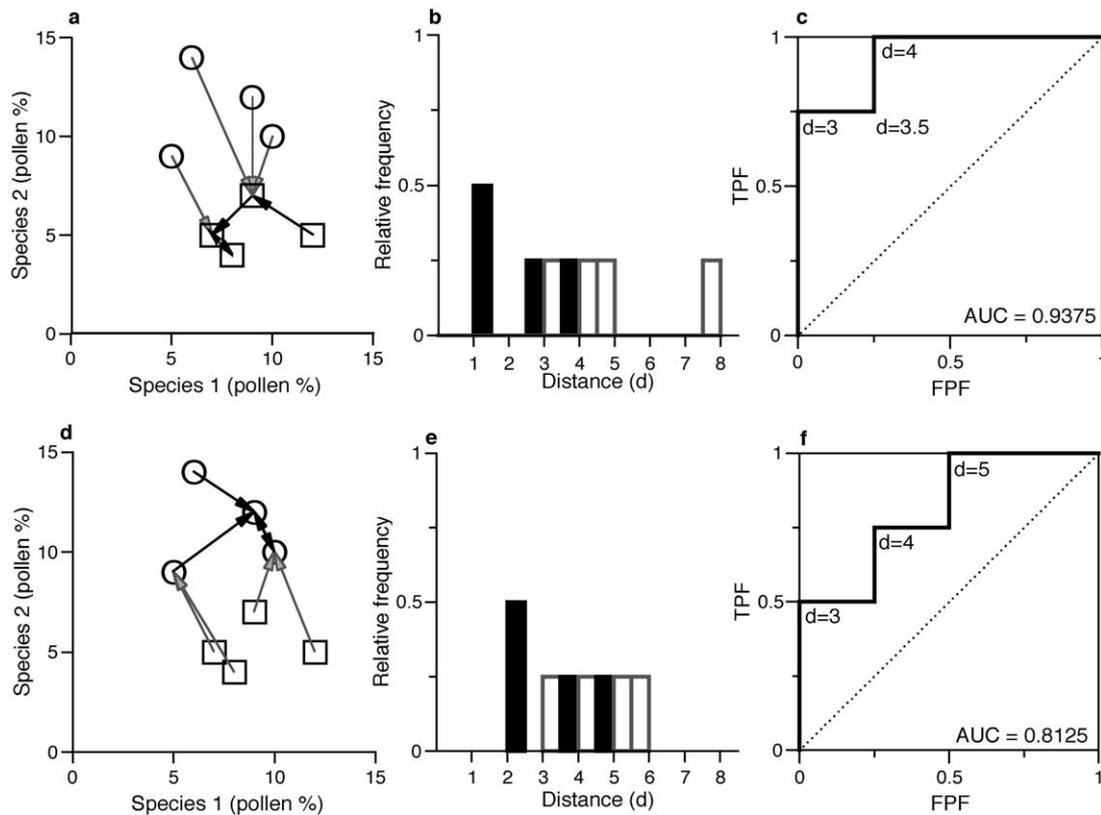


Fig. 2. Schematic showing how ROC analysis may be applied to the modern analog technique using modern pollen assemblages grouped into vegetation types. (a) Assemblages are shown with only two taxa for ease of presentation. Distances from assemblages to nearest neighbors in the same type (squares) represent distances typical of an analog situation (black arrows). Distances from assemblages in an alternative type (circles) to the nearest neighbor represent distances typical of no-analog situation (gray arrows). (b) Histograms of the two samples of distances (black codes for the analog situation, gray for the no-analog situation). (c) ROC curve based on the values in (b). The stair-step form of the ROC curve results from the relative ranking of distances representing analog and no-analog cases. Graphs (d) through (e) show the same methods applied to calibrating the distance metric for the other vegetation type (circles). Note that different within-vegetation type variance and the juxtaposition of assemblages in different vegetation types yield different ROC curves depending on which type is being assessed for analogs.

curves and computed the \hat{AUC} and its 95% confidence interval.

We chose the squared-chord distance (SCD) metric for more detailed analyses among all eight biomes because it performed well in the present analysis and has been identified in previous work as the best general metric for the MAT (Overpeck et al., 1985). We computed ROC curves for all 56 possible pairwise comparisons of biomes. For each biome, the relative frequencies of distances from the seven other biomes were averaged into a single histogram, which weights each biome equally irrespective of sample size. The likelihood ratios calculated from this average histogram represent, for a particular distance, the ratio of the probability that a fossil assemblage is from that biome vs. the probability it is from one of the seven other biomes. This likelihood ratio can then be used in conjunction with the prior probability to calculate a probability of analog for that biome (Eq. (2)).

We then used the likelihood ratio estimates for each biome to compute posterior probabilities of analog for fossil pollen assemblages from a detailed and well-dated pollen

record currently in the Deciduous Broadleaf Forest but close to the transition to the Southern Evergreen/Mixed Forest biome (Tannersville Bog, PA; Watts, 1979). The nearest analog from each fossil sample to each biome was determined using the SCD. We then computed posterior probabilities of analog for each fossil sample to each biome using the biome-specific likelihood ratios and prior probabilities fixed at 0.125 ($1/\#$ of biomes) following eq. (2). The posterior probabilities for each biome are for tests of hypotheses that a fossil assemblage is more likely an analog to that biome than to the other seven biomes. Thus, posterior probabilities do not necessarily sum to 1 over the eight biomes and may reflect the occurrence of no-analog vegetation types.

Our ability to discriminate among biomes that are potential analogs may be weakened by the inclusion of biomes with very low probability of analog. To better analyze the Tannersville Bog pollen record, we computed a second probability of analog using only biomes that in the first analysis (using eight biomes) had a probability of analog >0.05 . In this second analysis, the pollen record was broken

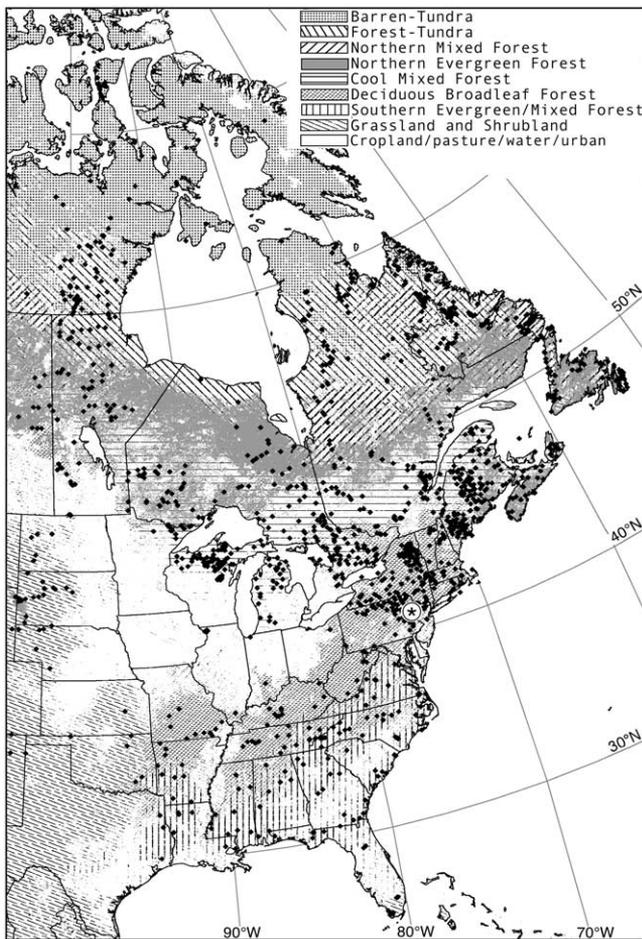


Fig. 3. Locations of 1689 modern pollen assemblages from eastern North America and the land cover classification (1-km resolution) modified the DISCover IGBP biome type (Loveland et al., 2000). Barren-Tundra, Forest-Tundra, and Grasslands and Shrublands represent several merged biome types. Mixed forest is split among three latitudinal zones (Northern, cool, and Southern), Evergreen Forest is split between two latitudinal zones (Northern and Southern), and the Southern Mixed and Evergreen Forest types are merged. The starred site is Tannersville Bog.

into five time periods, each with two to six biomes that were potential analogs. Likelihood ratios were recalculated using only the subset of biomes for each time period and new probabilities of analog were calculated using a prior probability of $1/\#$ of biomes.

Results

All metrics can distinguish between the Deciduous Broad-leaf Forest and the Southern Evergreen/Mixed Forest biomes better than chance alone ($AUC > 0.5$; Fig. 4). The three signal-to-noise metrics (squared chord, squared χ^2 , and information) and one unweighted metric (Euclidean) result in the best discrimination of the two biomes ($AUC =$ ca. 0.78). These metrics perform significantly better than two equal-weight metrics (Canberra and Gower's, $AUC =$ 0.641 and 0.621, respectively) and one unweighted metric

(squared $\cos-\theta$, $AUC = 0.631$). The standardized Euclidean ($AUC = 0.690$) and Manhattan ($AUC = 0.735$) metrics perform worse, though not at a statistically significant level, than the signal-to-noise metrics.

Overall, pollen assemblages from the deciduous broad-leaf forest are distinct from those from the seven other biomes when each biome is weighted equally ($AUC = 0.914$, using the SCD metric; Fig. 5). Separation of modern pollen assemblages among biomes increases with increasing geographic distance. There is nearly complete separation between pollen assemblages from the Tundra–Barren and Deciduous Broadleaf Forest biomes ($AUC = 0.994$), whereas pollen assemblages from the Cool Mixed Forest and Deciduous Broadleaf Forest biomes overlap strongly ($AUC = 0.732$). The optimal decision threshold (maximum TPF-FPF) also increases with increasing geographic distance between biomes, with values of 0.07 for adjacent biomes and as high as 0.15 for very dissimilar biomes.

The decision thresholds of the SCD vary substantially among biomes when each biome is compared to all other biomes combined (Fig. 6). Some biomes are tightly defined palynologically, as demonstrated by low within-biome distances and low decision thresholds (e.g., 0.06 for Forest–Tundra), while some biomes have more diverse pollen assemblages and high decision thresholds (e.g., 0.14 for Southern Evergreen/Mixed Forest). Despite this range in decision thresholds, many biomes were equally distinct from other biomes ($AUC =$ ca. 0.90), implying approximately equal power for SCD to distinguish among biomes. For all biomes, the likelihood ratio for the probability of analog vs no-analog decreases nonlinearly with increasing SCD (Fig. 6b). Likelihood ratios are more sensitive to SCD below the optimal decision threshold ($LR < 1$) than above it ($LR > 1$). However, each biome differs in the magnitude of the likelihood ratio at small SCD. Likelihood ratios are more easily interpreted when translated into posterior probabilities. The relationship between SCD and posterior probability of analog also is nonlinear, with much greater sensitivity to small changes at low SCD vs high SCD (Fig. 7).

Table 1
Pollen types used in the distance metric calculations

<i>Abies</i>	<i>Ostrya/Carpinus</i>
<i>Acer</i>	<i>Picea</i>
<i>Alnus</i>	<i>Pinus</i>
<i>Betula</i>	<i>Platanus</i>
<i>Carya</i>	Poaceae
<i>Corylus</i>	<i>Populus</i>
Cupressaceae/Taxaceae	Prairie Forbs*
Cyperaceae	<i>Quercus</i>
<i>Fagus</i>	<i>Salix</i>
<i>Fraxinus</i>	<i>Tilia</i>
<i>Juglans</i>	<i>Tsuga</i>
<i>Larix</i>	<i>Ulmus</i>
<i>Liquidambar</i>	

* Prairie Forbs = sum of Chenopodiaceae/Amaranthaceae and Asteraceae (excluding Ambrosia).

Table 2

Nine distance metrics for determining dissimilarity between two pollen assemblages (from Prentice, 1980; Overpeck et al., 1985)

Distance metric		Formula
Equal-weight metrics	Canberra distance	$d_{ij} = \sum_k \frac{ p_{ik} - p_{jk} }{p_{ik} + p_{jk}}$
	Standardized Euclidean distance	$d_{ij} = \sqrt{\sum_k \frac{(p_{ik} - p_{jk})^2}{s_k}}$
	Gower's distance	$d_{ij} = \sqrt{2 \sum_k \frac{ p_{ik} - p_{jk} }{R_k}}$
Unweighted metrics	Manhattan distance	$d_{ij} = \sum_k p_{ik} - p_{jk} $
	Euclidean distance	$d_{ij} = \sqrt{\sum_k (p_{ik} - p_{jk})^2}$
	Squared cos- θ distance	$d_{ij} = \sum_k \left(\frac{p_{ik}}{\sqrt{\sum_k p_{ik}^2}} - \frac{p_{jk}}{\sqrt{\sum_k p_{jk}^2}} \right)^2$
Signal-to-noise metrics	Squared-chord distance	$d_{ij} = \sum_k (\sqrt{p_{ik}} - \sqrt{p_{jk}})^2$
	Squared χ^2 distance	$d_{ij} = \sum_k \frac{(p_{ij} - p_{jk})^2}{p_{ik} + p_{jk}}$
	Information statistic	$d_{ij} = \sum_k \left(p_{ik} \ln \frac{2p_{ik}}{p_{ik} + p_{jk}} + p_{jk} \ln \frac{2p_{jk}}{p_{ik} + p_{jk}} \right)$

Note. p_{ik} = the proportion of pollen type k in sample i , R_k = the range of proportions for pollen type k over all samples, and s_k = the standard deviation of proportions of pollen type k over all samples.

As expected, the relationship between SCD and posterior probability depends strongly on the prior probability (Fig. 7; Eq. (2)).

The pollen record from the Tannersville Bog shows a fluctuating probability of analog over the past >16,000 years (Fig. 8). The initial analysis of all eight biomes using a constant prior probability of 0.125 resulted in a time series of fairly low probabilities of analog (rarely >0.2). However, the analysis based on subsets of the most likely biomes resulted in higher probabilities (often >0.4) and time series that appeared more sensitive to changes in the pollen record. The difference between the two analyses is likely due to a combination of higher prior probabilities and more sensitive relationships between SCD and LR (not shown) when using subsets of potential analog biomes. We focus below only on the results based on the analysis of subsets of biomes.

Prior to 16,000 cal yr B.P., the lowest SCD values (ca. 0.1) are to the Northern evergreen forest, resulting in a moderate probability of analog (ca. 0.2). SCD values decrease sharply, and probability of analog increases, for Northern Evergreen Forest at a peak in *Picea* pollen at 16,000 cal yr B.P. Between ca. 16,000 and 13,500 cal yr B.P., cool mixed forest has the greatest probability of analog (ca. 0.2). A period of very low probability of analog for all biomes occurs between 13,500 and 12,500 cal yr B.P. At 12,500 cal yr B.P., concurrent with an increase in *Pinus* pollen and decrease in *Betula*, *Picea*, and *Alnus* pollen, SCD decreases to <0.1 and probability of analog increases to ca. 0.25 for three biomes (Cool Mixed Forest, Deciduous

Broadleaf Forest, and Southern Evergreen/Mixed Forest). At 10,000 cal yr B.P. SCD to Cool Mixed Forest increases to 0.2 and the corresponding probability of analog decreases to <0.05, shortly followed by a similar decrease in probability of analog to Southern Evergreen/Mixed Forest. This period of low probability of analog continues until an increase in *Carya* pollen at ca. 6000 cal yr B.P. when probabilities increase to ca. 0.45 for Deciduous Broadleaf Forest. During a period with low *Tsuga* pollen (5000 to 3500 cal yr B.P.) Southern Evergreen/Mixed Forest has a slightly greater probability of analog than Deciduous Broadleaf Forest. For the remainder of the record both Deciduous Broadleaf Forest and Southern Evergreen/Mixed Forest have SCD values of ca. 0.10–0.15 and higher probabilities of analog (0.4–0.5). The probability of analog to the Deciduous Broadleaf Forest increases to 1.0 for the uppermost sample (which was part of the modern data set).

Discussion

ROC analysis provides a formal means of assessing previously ambiguous steps of the MAT. It can be used to compare the performance of different distance metrics, evaluate the tradeoffs of using different decision thresholds, and to compute posterior probabilities of analog for fossil assemblages. We discuss the effectiveness of the ROC method when applied to modern pollen assemblages from eastern North America.

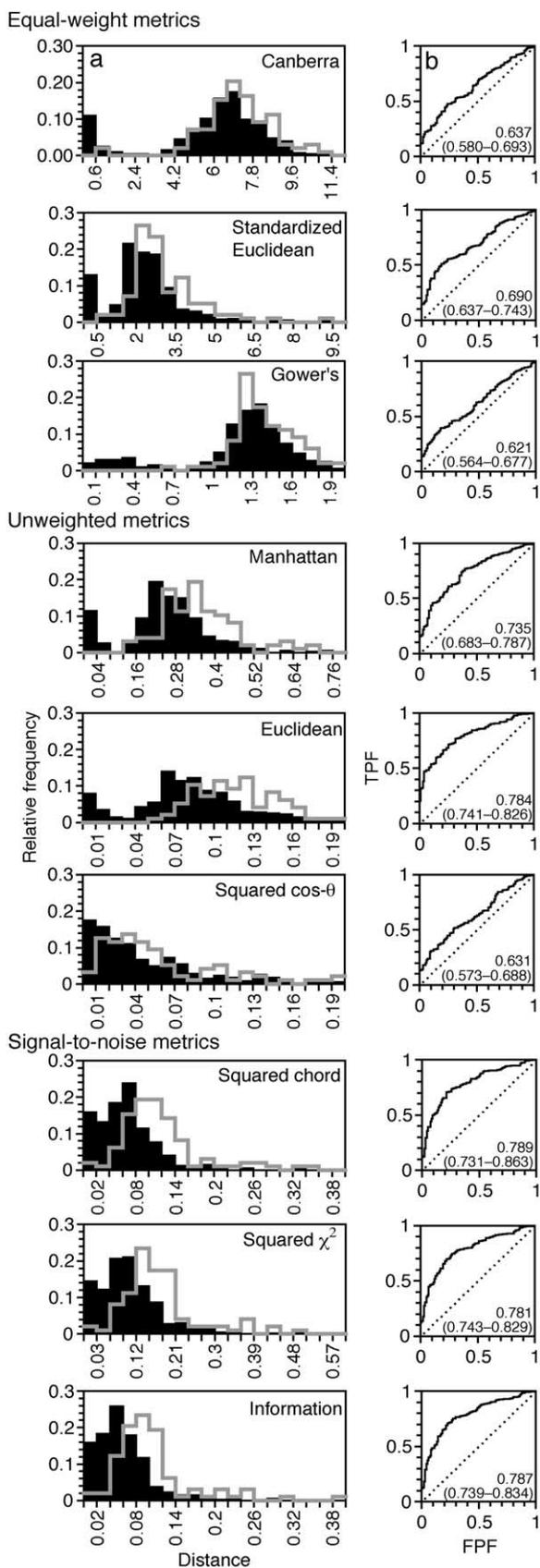


Fig. 4. Accuracy of nine different distance metrics at distinguishing between pollen assemblages from Deciduous Broadleaf and Southern Ever-

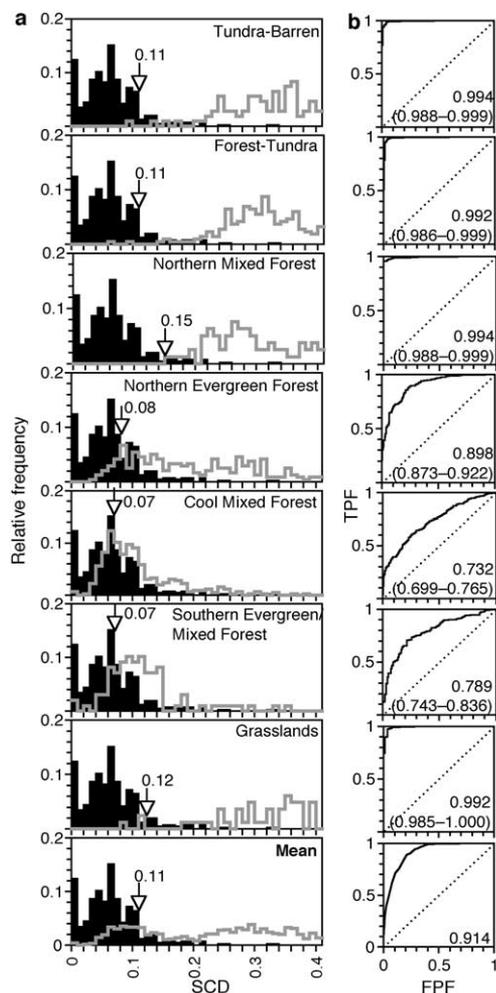


Fig. 5. Accuracy of the squared-chord distance (SCD) at distinguishing Deciduous Broadleaf Forest pollen assemblages from pollen assemblages in seven other biomes. (a) Histograms of the nearest-neighbor distances between pollen assemblages within the deciduous broadleaf forest (solid bars) and the nearest-neighbor distances from each pollen assemblage in each of the seven other biomes to pollen assemblages in the Deciduous Broadleaf Forest (gray line). Relative frequencies are used to account for unequal sample sizes in different biomes. Arrows indicate the optimal decision threshold for SCD (jointly maximizing the true positive and true negative fractions). The average of the histograms from each of the seven biomes is shown in the bottom panel. (b) ROC curves for each comparison. AUC and its 95% confidence interval determined using the Wilcoxon rank-sum statistic is shown in the lower right corner.

green/Mixed Forest. (a) Histograms of the nearest-neighbor distances between pollen assemblages within Deciduous Broadleaf Forest (solid bars) and the nearest-neighbor distances from each pollen assemblage in the Southern evergreen/mixed forest to pollen assemblages in the deciduous broadleaf forest (gray line). Relative frequencies are used to account for unequal sample sizes in different biomes. All within-biome comparisons are shown, but a large proportion of between-biome comparisons (gray line) occur beyond the range shown. (b) ROC curves for each set of histograms, with AUC and its 95% confidence interval determined using the Wilcoxon rank-sum statistic shown in the lower right corner.

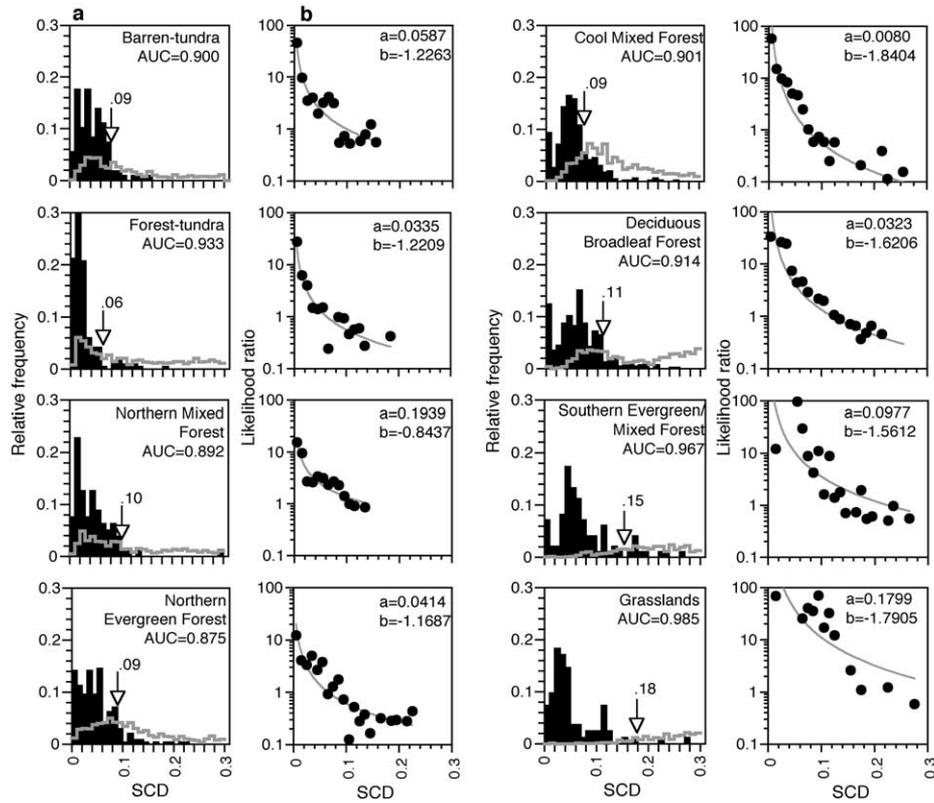


Fig. 6. Accuracy of the squared-chord distance (SCD) at distinguishing pollen assemblages from each of eight North American biomes from pollen assemblages in every other biome. (a) Histograms of the nearest-neighbor distances between pollen assemblages within each biome (solid bars) and the average of the histograms of nearest-neighbor distances between pollen assemblages from the seven other biomes to that biome (gray line; see Fig. 5). (b) Likelihood ratio for an analog vs no-analog result computed as the ratio of relative frequencies of within and between-biome comparisons over a range of SCD values. Values for a and b are parameters of the fitted curve: $LR = a*(SCD)^b$. Arrows in part (a) indicate the optimal threshold for SCD (jointly maximizing the true positive and true negative fractions), which roughly corresponds to the point in part (b) where $LR = 1$.

Statistical comparison of distance metrics

Some distance metrics have greater power to discriminate among vegetation types than others. Although our ranking of distance metrics agrees with a previous study using a similar data set (Overpeck et al., 1985), it is now possible to attach levels of significance to these findings and assess data sets where other distance metrics are more appropriate. For example, Oswald (in press) used ROC analysis to show that an equal-weight metric (Canberra) was better than the SCD for discriminating between pollen assemblages from contrasting arctic tundra communities where the vegetation types differed only in terms of rare taxa.

Decision thresholds for assessing modern analogs

Our analysis of modern pollen assemblages from eastern North American biomes show that decision thresholds for the SCD vary at different spatial scales and among biomes. The biome-by-biome comparison (Fig. 5) shows that thresholds are scale-dependent; i.e., thresholds decreased from 0.15 to 0.07 when increasingly similar biomes were con-

trasted. Studies that compare vegetation types at finer resolution would require smaller decision thresholds (e.g., 0.05 for forest stand types; Calcote, 1998). The accuracy of the distance metric (as measured by AUC) also decreases when biomes are more similar, so that the decision threshold and accuracy tend to be positively correlated (Fig. 5). In addition, the overall decision threshold (each biome compared to the average of the other biomes) varies among biomes (Fig. 6). These differences are attributable to differences in within-biome variation of pollen assemblages as well as the separation of pollen assemblages among biomes.

We identified lower optimal decision thresholds (maximum TPF-FPF) for most biomes (ca. 0.08; Fig. 6) than reported by Overpeck et al. (1985) (0.15) using a similar data set. This discrepancy is due to our calibration of the SCD using the nearest neighbor distances within and between biomes because analyses of the fossil records are often based only on distance to the nearest modern analog. In contrast, Overpeck et al. (1985) used essentially all paired comparisons among modern assemblages. Including more than just nearest-neighbor comparisons increases the variance in the distributions used in the ROC analysis, and thus increases the decision threshold. In addition, Overpeck

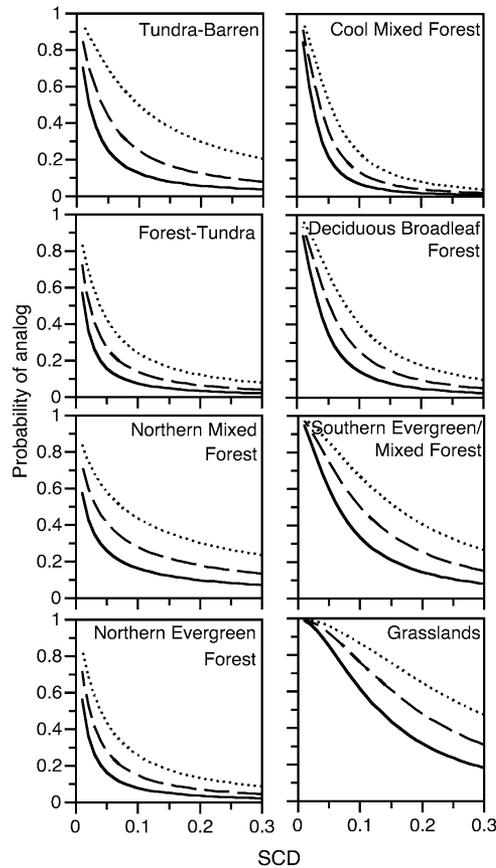


Fig. 7. Relationship between squared-chord distance (SCD) and posterior probability of analog, computed using the likelihood ratios from Fig. 6 and a range of prior probabilities: 0.125 (solid line), 0.25 (dashed line), and 0.5 (dotted line).

et al. (1985) used visual observation of distances to determine the value of decision thresholds, which does not allow rigorous examination of how threshold values interact with the false positive and false negative fractions.

The loss in vegetation discrimination can be asymmetric above and below the optimal threshold that maximizes the difference between TPF and FPF. Initial work on this issue identified a more rapid loss in discrimination at thresholds below the optimal d' than above the optimal d' when the variance of the actually positive distribution is less than the variance of the actually negative distribution (Wahl, in press). This situation is consistently noted in modern pollen surface sample sets from which similar distributions are reported (Anderson et al., 1989; Davis, 1995; Wahl, in press) and is true for the data used in this analysis (Figs. 5 and 6). This asymmetry results in a faster increase in false negatives when decreasing the threshold below the optimal d' relative to the corresponding increase in false positives when increasing the threshold above the optimal d' . ROC-based analysis offers a way to explicitly examine the effects of this asymmetry and thus make rigorous choices concerning the tradeoffs involved when choosing a threshold other than the optimal d' : increasing false negatives (sacrificing

sensitivity) and decreasing false positives (boosting specificity) when d' is set below the optimum; and decreasing false negatives (boosting sensitivity) and increasing false positives (sacrificing specificity) when d' is set above the optimum.

Likelihood ratios and the probability of analog

The North American pollen data demonstrated a highly nonlinear relationship between the likelihood ratio (probability of analog vs no-analog) and the distance metric (Fig. 6). Minor changes in the SCD have a greater effect on the probability of analog at low than at high SCD values. In fact, the relationship between the SCD and posterior probability for a given prior probability shows posterior probabilities are a power function (steeper-than-exponential) of the SCD (Fig. 7). Using likelihood ratios in a Bayesian analysis, therefore, is a means to transform the SCD values into an interpretable probability of analog that accommodates the nonlinear relationship between SCDs and posterior probabilities. This Bayesian approach to environmental reconstruction has already been advocated for different applications (Robertson et al., 1999; Toivonen et al., 2001).

Our use of Bayesian analysis with the MAT applied to the Tannersville Bog pollen record highlights several important considerations (Fig. 8). The initial analysis of the pollen record, using all eight biomes and low prior probability of 0.125, showed only minor changes in the probability of analog for several biomes despite major changes in pollen assemblages. The constant low prior probability used in this analysis effectively muted the response of the probability of analog to changes in SCD. Our approach to this problem was to limit the number of potential biomes in the analysis, allowing the use of higher prior probabilities. Limiting the number of biomes in the analysis resulted in a probability of analog that was more sensitive to changes in the pollen record (compare dashed and solid lines in Fig. 8). Another approach would be to set priors based on the posteriors of the preceding sample. However, this approach would cause the posterior probability (probability of analog) to decrease asymptotically to 0 if the likelihood ratio is consistently less than 1, or increase asymptotically to 1 if the likelihood ratio is consistently greater than 1. Very small or large prior probabilities would prevent any rapid changes in vegetation from registering as rapid changes in the probability of analog. Our decision to use equal and constant prior probabilities was chosen to make the time series of probability of analog easier to interpret.

Decisions and tradeoffs inherent to the MAT and ROC analysis

The nearest-neighbor framework for constructing ROC curves (Fig. 2) is very flexible but also may be overly affected by single outlier samples. A modern assemblage that falls very near a different group, or is misclassified,

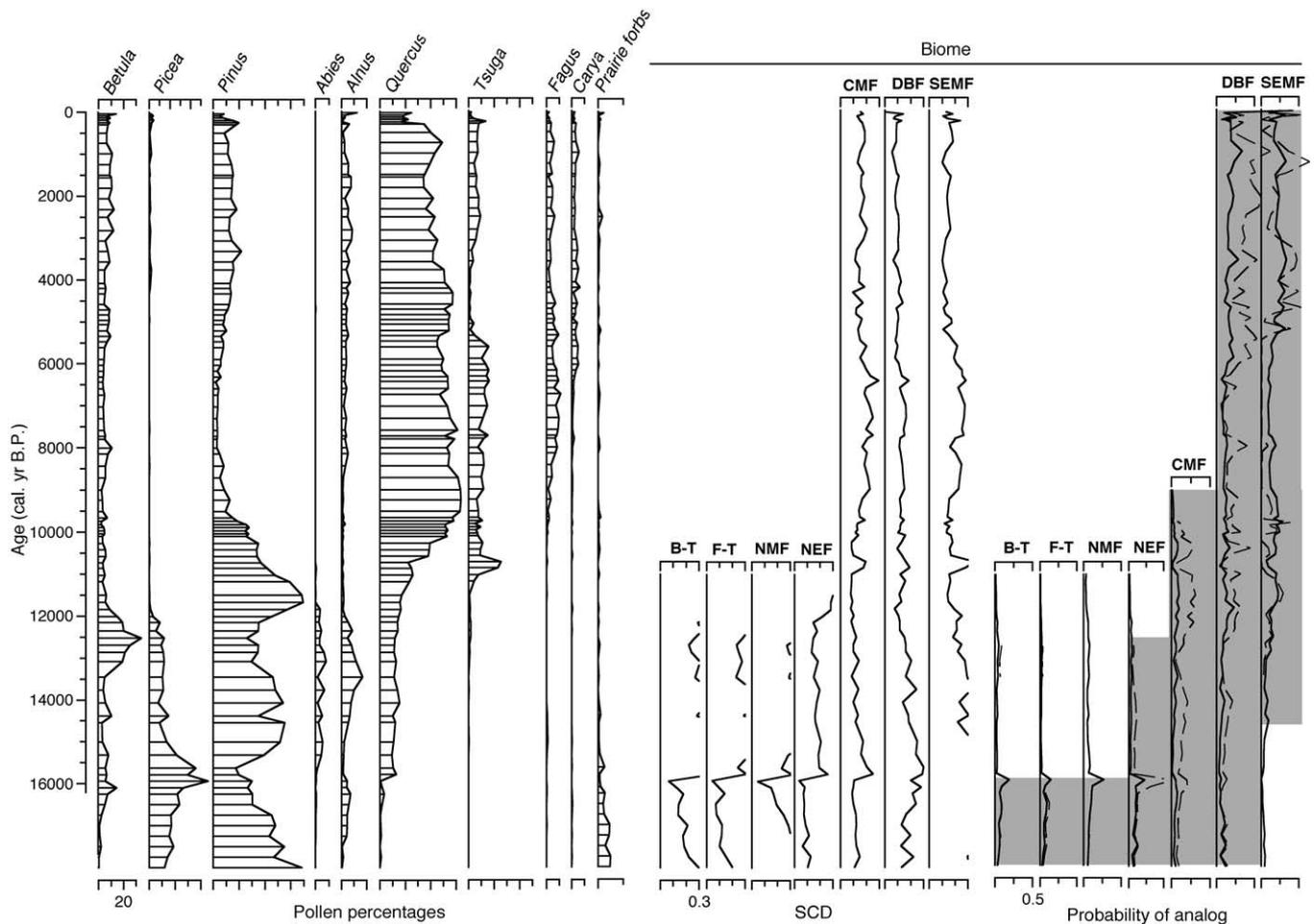


Fig. 8. Selected pollen types and analog results from Tannersville Bog, PA (Watts, 1979). Squared-chord distances (SCD) to the nearest modern assemblages in each of seven biomes are shown, where $SCD < 0.3$. Bayesian posterior probability of analog of fossil pollen assemblages was calculated using the relationship between likelihood ratios and the SCD to the nearest modern analog in each biome. Solid lines show the initial analysis using all eight biomes (Fig. 6) and a constant prior probability of 0.125. Dashed lines show a second analysis including only biomes in the shaded portions and using higher prior probabilities ($1/\#$ of biomes). All posterior probabilities for the Grassland biome (not presented) were < 0.001 . The chronology is based on linear interpolation between calibrated radiocarbon dates (Stuiver et al., 1998).

could have a large effect on the resulting ROC curve, especially with small sample sizes. Outliers could be assessed using multivariate techniques such as principal component analysis. The effect of outliers could also be minimized by averaging distances from each modern sample to the nearest 3–10 neighbors rather than the nearest single neighbor, thus decreasing the effect of the largest outliers. Using this approach in the current analysis, we found no significant differences between ROC curves constructed from the average of the nearest three neighbors and the nearest single neighbor, suggesting little influence of single outliers in this large data set.

A second difficulty of ROC analysis is that it requires modern samples be classified into groups. This approach is also necessary for other methods of environmental reconstruction, including biomization (Prentice et al., 1996) and linear discriminant analysis (Liu and Lam, 1985). There will naturally be ambiguities in classifying modern communities

into discrete groups, such as the eight biomes used in this study (Fig. 3). However, this prerequisite should not affect analysis of fossil records with substantial periods that are transitional between biomes. Such periods may be represented by simultaneous moderate probabilities of analog to both biomes, assuming that the two biomes are sufficiently palynologically distinct (e.g., the late Holocene in Fig. 8).

Conclusions

ROC analysis resolves previously ambiguous steps of the modern analog technique: the choice of decision thresholds, the discriminating power of various distance metrics, and a formalized assessment of the probability that a pollen assemblage is from any given biome. ROC analysis requires no parametric assumptions. Applied to modern pollen assemblages from eastern North America, ROC analysis is

successful at identifying the SCD as an appropriate distance metric, at determining appropriate thresholds for the SCD, and at estimating the (Bayesian) probability of analog of fossil assemblages to major biomes. Future refinements to the ROC method applied to the MAT include (1) computing confidence intervals for likelihood ratios to more critically evaluate the probability of analog, (2) applying the method to finer-scale ecological classifications, and (3) explicit comparisons of the method with other quantitative methods of classifying fossil pollen assemblages.

Acknowledgments

This paper is part of D. Gavin's postdoctoral research, supported by a Packard Fellowship in Science and Engineering to F. S. Hu. We thank T. Webb III and an anonymous reviewer for comments on the manuscript.

References

- Anderson, P.M., Bartlein, P.M., Brubaker, L.B., Gajewski, K., Ritchie, J.C., 1989. Modern analogs of late-Quaternary pollen spectra from the western interior of North America. *Journal of Biogeography* 16, 573–596.
- Birks, H.J.B., Gordon, A.D., 1985. *Numerical Methods in Quaternary Pollen Analysis*. Academic Press, London.
- Calcote, R., 1998. Identifying forest stand types using pollen from forest hollows. *The Holocene* 8, 423–432.
- Davis, O.K., 1995. Climate and vegetation patterns in surface samples from arid western USA: application to Holocene climatic reconstructions. *Palynology* 19, 95–117.
- Davis, M., Douglas, C., Calcote, R., Cole, K.L., Winkler, M.G., Flakne, R., 2000. Holocene climate in the western Great Lakes national parks and lakeshores: implications for future climate change. *Conservation Biology* 14, 968–983.
- Guiot, J., 1990. Methodology of the last climatic cycle reconstruction in France from pollen data. *Palaeogeography Palaeoclimatology Palaeoecology* 80, 49–69.
- Guiot, J., Harrison, S.P., Prentice, I.C., 1993. Reconstruction of Holocene precipitation patterns in Europe using pollen and lake-level data. *Quaternary Research* 40, 139–149.
- Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.
- Henderson, A.R., 1993. Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. *Annals of Clinical Biochemistry* 30, 521–539.
- Hilborn, R., Mangel, M., 1997. *The Ecological Detective*. Princeton University Press, Princeton, NJ.
- Liu, K., Lam, N.S., 1985. Paleovegetational reconstruction based on modern and fossil pollen data: an application of discriminant analysis. *Annals of the Association of American Geographers* 75, 115–130.
- Loveland, T.R., Reed, B.C., Brown, J.F., Ohlen, D.O., Zhu, J., Yang, L., Merchant, J.W., 2000. Development of a global land cover characteristics database and IGBP DISCover from 1-km AVHRR data. *International Journal of Remote Sensing* 21, 1303–1330.
- Metz, C.E., 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 8, 283–298.
- Metz, C.E., 1986. Statistical analysis of ROC data in evaluating diagnostic performance, in: Herbert, D.E., Myers, R.H. (Eds.), *Multiple Regression Analysis: Application in the Health Sciences*. American Institute of Physics, Washington, DC, pp. 365–384.
- Metz C.E. 1998. ROCKIT User's Guide. University of Chicago. Retrieved from <http://www-radiology.uchicago.edu/krl/toppage11.htm>, January 2003.
- Mossman, D. 1995. Resampling techniques in the analysis of non-binomial ROC data. *Medical Decision Making* 15, 358–366.
- Oswald, W.W., Brubaker, L.B., Hu, F.S., Gavin, D.G. Pollen-vegetation calibration for tundra communities in the Arctic Foothills, northern Alaska. *Journal of Ecology*, in press.
- Overpeck, J.T., Webb III, T., Prentice, I.C., 1985. Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. *Quaternary Research* 23, 87–108.
- Pflaumann, U., Duprat, J., Pujol, C., Labeyrie, L.D., 1996. SIMMAX: A modern analog technique to deduce Atlantic sea surface temperatures from planktonic foraminifera in deep-sea sediments. *Paleoceanography* 11, 15–35.
- Prell, W.L., 1985. The Stability of Low-Latitude Sea-Surface Temperatures: An Evaluation of the CLIMAP Reconstruction with Emphasis on the Positive SST Anomalies, Rep. TR 025. U.S. Department of Energy, Washington, D.C.
- Prentice, I.C., 1980. Multidimensional scaling as a research tool in Quaternary palynology: a review of theory and methods. *Review of Palaeobotany and Palynology* 31, 71–104.
- Prentice, I.C., Guiot, J., Huntley, B., Jolly, D., Cheddadi, R., 1996. Reconstructing biomes from palaeoecological data: a general method and its application to European pollen data at 0 and 6 ka. *Climate Dynamics* 12, 185–194.
- Robertson, I., Lucy, D., Baxter, L., Pollard, A.M., Aykroyd, R.G., Barker, A.C., Carter, A.H.C., Switsur, V.R., Waterhouse, J.S., 1999. A kernel-based Bayesian approach to climatic reconstruction. *The Holocene* 9, 495–500.
- Sawada, M., Viau, A., Gajewski, K., 2001. Critical thresholds of dissimilarity in the modern analog technique (MAT) for quantitative paleoclimate reconstruction, in: Chylek, P., Lesins, G. (Eds.), *1st Annual Conference on Global Warming and the Next Ice Age*. Dalhousie University, Halifax, Nova Scotia, Canada, pp. 149–152 (Compilers).
- Schweitzer, P.N., 1999. ANALOG: A Program for Estimating Paleoclimate Parameters Using the Method of Modern Analogs U.S. Geological Survey Open-File Report 94–645. United States Geological Survey, Reston, VA.
- Stuiver, M., Reimer, P.J., Bard, E., Beck, J.W., Burr, G.S., Hughen, K.A., Kromer, B., McCormac, G., Van der Plicht, J., Spurk, M., 1998. INTCAL98 radiocarbon age calibration, 24,000–0 cal BP. *Radiocarbon* 40, 1041–1083.
- Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. *Science* 240, 1285–1293.
- Toivonen, H.T.T., Mannila, H., Korhola, A., Olander, H., 2001. Applying Bayesian statistics to organism-based environmental reconstruction. *Ecological Applications* 11, 618–630.
- Waelbroeck, C., Labeyrie, L., Duplessy, J.-C., Guiot, J., Labracherie, M., Leclaire, H., Duprat, J., 1998. Improving past sea surface temperature estimates based on planktonic fossil faunas. *Paleoceanography* 13, 272–283.
- Wahl, E.R. A general framework for determining cutoff values to select pollen analogs with dissimilarity metrics in the modern analog technique. *Review of Palaeobotany and Palynology*, in press.
- Watts, W.A., 1979. Late Quaternary vegetation of central Appalachia and the New Jersey coastal plain. *Ecological Monographs* 49, 427–469.
- Williams, J.W., Shuman, B.N., Webb III, T., 2001. Dissimilarity analyses of late-Quaternary vegetation and climate in eastern North America. *Ecology* 82, 3346–3362.
- Williams, J.W., Jackson, S.T. (2003). Palynological and AVHRR observations of modern vegetational gradients in eastern North America. *The Holocene* 13, 485–497.
- Youden, W., 1950. Index rating for diagnostic tests. *Cancer* 3, 32–35.
- Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 39, 561–577.