Classification Trees as an Alternative to Linear Discriminant Analysis

Marc R. Feldesman*

Department of Anthropology, Portland State University, Portland, Oregon 97207

KEY WORDS multivariate analysis; binary recursive trees; morphometrics; rpart; open source software; missing values imputation; discriminant analysis

ABSTRACT Linear discriminant analysis (LDA) is frequently used for classification/prediction problems in physical anthropology, but it is unusual to find examples where researchers consider the statistical limitations and assumptions required for this technique. In these instances, it is difficult to know whether the predictions are reliable. This paper considers a nonparametric alternative to predictive LDA: binary, recursive (or classification) trees. This approach has the advantage that data transformation is unnecessary, cases with missing predictor variables do not require special treatment, prediction success is not dependent on data meeting normality conditions or covariance homogeneity, and variable selection is intrinsic to the methodology. Here I compare the efficacy of classification trees with LDA, using typical morphometric data. With data from modern hominoids, the results show that both techniques perform nearly equally. With complete data sets, LDA may be a better choice, as is shown in this example, but with missing observations, classification trees perform outstandingly well, whereas commercial discriminant analysis programs do not predict classifications for cases with incompletely measured predictor variables and generally are not designed to address the problem of missing data. Testing of data prior to analysis is necessary, and classification trees are recommended either as a replacement for LDA or as a supplement whenever data do not meet relevant assumptions. It is highly recommended as an alternative to LDA whenever the data set contains important cases with missing predictor variables. Am J Phys Anthropol 119:257–275, 2002. © 2002 Wiley-Liss, Inc.

Physical anthropologists commonly use linear discriminant analysis (LDA) for prediction and classification, but rarely discuss its limitations, which are quite-well documented in any basic textbook of multivariate statistics (e.g., Flury, 1997; Johnson and Wichern, 1998; McLachlan, 1992; Mardia et al., 1979; Morrison, 1990). A few researchers in physical anthropology and allied fields (e.g., Campbell, 1984a–c; Corruccini, 1978; Feldesman, 1997; Kowalski, 1972; Schaafsma and van Vark, 1977, 1979; van Vark, 1976, 1995) explicitly commented on the assumptions required to undertake discriminant analysis, canonical variates analysis, or multiple discriminant analysis. These papers have not been cited often, and some researchers who use LDA continue to do so without mentioning any of the crucial assumptions or constraints (e.g., Aiello et al., 1999; Holliday, 2000).

In most applications of LDA, researchers assume either tacitly or explicitly that the data are multivariate normal and have homogeneous group covariance matrices. In addition, the method depends on correct assignment of training cases to groups, and works most efficiently if the smallest group has significantly more cases than variables and when groups are approximately equally sized. In practice, while LDA is only marginally affected by nonnormality, outliers, covariance heterogeneity, and disparate, unequal, and/or frequently small sample sizes make classification results unstable under resampling or cross-validation. These issues rarely get addressed in reports where LDA is used for classification or prediction. To make matters more complicated, the most frequently used implementations of LDA (i.e., SAS, SPSS, and SYSTAT) are designed only for complete suites of measurements on each individual, and delete any case missing one or more predictor variables. To overcome this, a few researchers have used multiple regression to predict missing values, or have substituted means for missing values.[1] These approaches were discredited by Schafer (1997) and Schimert et al. (2000), who noted that deleting cases throws away important informa-

Grant sponsor: Portland State University Faculty Development; Grant sponsor: Intel Foundation.

*Correspondence to: Marc R. Feldesman, Department of Anthropology, Portland State University, P.O. Box 751, Portland, OR 97207. E-mail: feldesmanm@pdx.edu

Received 5 March 2000; accepted 22 January 2002.

DOI 10.1002/ajpa.10102
Published online in Wiley InterScience (www.interscience.wiley.com).

[1] SPSS has this option available.

tion; replacing a missing value with a mean preserves sample means but biases the estimated variances and covariances toward zero, while using multiple regression imputation biases the correlations away from zero. The effect of all these approaches is to distort information in the remaining data, leading to potentially misleading answers, or creating situations where it is difficult to discern whether the methods work or fail.

Consequently, there are only limited parametric options available for researchers whose interest is classification of cases with incompletely observed predictor variables.[2] Researchers must either 1) analyze only cases with complete data sets, 2) form multiple (and sometimes overlapping) subsets of cases, each with slightly different measurement suites analyzed sequentially, or 3) find, understand, and apply a data imputation algorithm designed both to "fill in" missing values and to estimate confidence bounds on the resulting statistical estimates. The first reduces the comparative sample size and introduces subtle biases, the second reduces the number of variables in any single analysis (e.g., Stringer, 1974; Kidder et al., 1992) and can inflate significance levels when there are multiple comparisons, and the third is problematic unless the number of missing values is relatively small, and troublesome unless great care is taken to avoid matrix singularities resulting from the bias introduced by "filling in" missing values (for a discussion of imputation bias, see Breiman et al., 1984; Schafer, 1997).

When the research goal is classification, what are the alternatives to LDA when the data fail to meet its requisite assumptions? Alternatives include quadratic discriminant analysis, multinomial logistic regression, flexible discriminants, mixture discriminant analysis, robust discriminant analysis, and neural networks. Hastie et al. (2001) and Ripley (1996) reviewed the benefits and pitfalls of these techniques. The present study considers a nonparametric alternative to LDA. This technique is known as binary recursive partitioning or, more commonly, as "classification trees" (Breiman et al., 1984; Venables and Ripley, 1997, 1999; Therneau and Atkinson, 1997; Steinberg and Colla, 1997). My review found no application of nonparametric alternatives to LDA for classification and prediction problems in physical anthropology.

In this paper, I compare the predictive accuracy of nonparametric classification trees with LDA. I also examine the results obtained with classification trees that are applied to cases with incompletely observed predictor variables; commercial LDA software typically deletes such cases. While the idea of a classification tree may be familiar to physical anthropologists accustomed to making taxonomic decisions, classification tree algorithms are not ordinarily used for taxonomic purposes, and the statistical methodology that underpins them may be novel. Below, I discuss the methodology used in one major classification tree algorithm through an extended example.

Classification trees, binary recursive partitioning, or tree-structured analyses have been around since the 1960s, but computational requirements limited their use until recently. Breiman et al. (1984) were responsible for bringing classification trees into the mainstream of applied statistics and, in the process, for developing their essential theoretical properties.

There are now many algorithms for formulating such trees (e.g., Lim et al., 2000, who compared the performance of 33 such algorithms). Of these, however, the binary recursive partitioning algorithm developed by Breiman et al. (1984), henceforth called the BFOS algorithm, remains the best-known, most dependable, and most thoroughly tested (Lim et al., 2000).

Steinberg and Colla (1997) enumerated the general technical advantages of the BFOS classification tree algorithm over parametric techniques like LDA, quadratic discriminant analysis, and multinomial logistic regression. The primary advantages are: 1) it is nonparametric, which makes questions of the appropriate distributional form moot; 2) it requires no advance variable selection, because variables are automatically selected for their efficacy in reducing classification errors; variables making little or no contribution to classification success are not used; 3) it is robust to outliers, which rarely define split points that correctly classify a significant number of cases (if they did, they wouldn't be outliers); 4) its results are invariant to monotone transformations of independent variables (e.g., logarithmic transformation will not change the tree structure); 5) it can use any combination of categorical and continuous predictor variables (e.g., height, weight, sex, age, hair color, or marital status); 6) it handles missing values in predictor variables by developing splitting rules based on alternate measurements (surrogates) that exhibit strong concordance with the primary splitting variable at any given point on the tree; and 7) cases with unknown and unknowable response (or classification) variables (e.g., fossils) can be placed in their/its own group and participate in tree construction, which contrasts with LDA, where groups with one or only a few cases must be excluded from calculation of the LDA because it is not possible to compute a meaningful covariance matrix on classes that small.[3]

[2] Note the usage here. This makes no reference to cases with unknown (missing) response or classification variables. I do not consider this problem in the current investigation (but see footnote 3 for a brief discussion).

[3] The solution to the "unknown and unknowable group incumbency" problem is intrinsic to binary tree methodology, as is the simultaneous classification of cases with missing predictor variables. The current paper does not deal with the former, but the principle is simple. Briefly, fossils are assigned to taxa. The relevant question is the relationship between the fossil taxa and the other groups for

Predictive or classificatory LDA is based on a significant set of assumptions, many of which are violated by typical morphometric data. While failure to meet some or all the assumptions is not a fatal flaw, it is important to identify the extent of the departures, and to consider alternative techniques when deviations are significant. As a result, I consider whether such classification questions can be answered effectively with nonparametric binary trees.

This paper explores the efficacy of binary recursive classification trees in two circumstances: first, as an alternative to predictive LDA when data do not meet the necessary assumptions; and second, when missing data reduce the size of the data set presented to a standard packaged LDA routine.

MATERIALS AND METHODS

Data for the present inquiry consisted of 10 measurements primarily from the distal humerus of 237 modern hominoids: 86 *Pan gorilla*, 114 *Pan troglodytes*, 23 *Pongo pygmaeus*, and 24 *Homo sapiens*. Sex, taxonomic affiliation, and complete measurement suites were recorded for all specimens. All nonhumans were wild-shot adults curated in five Western European museums; the human materials were mixed ethnicity and sex (autopsy) skeletons that make up Portland State University Anthropology Department's osteological teaching collection. The measurements, described in Feldesman (1976), include: LATSUPRI (length of the lateral supracondylar crest), MEDEPICO (medial epicondyle expansion), PDHTCAPI (proximo-distal height of capitulum), MLHTCAPI (medio-lateral breadth of capitulum), APHTTROC (antero-posterior height of trochlea), ANTARTBR (anterior articular breadth), OLECRDEP (olecranon fossa depth), HUMLENGT (maximum humerus length), and BIEPI (biepicondylar breadth). Previous morphometric investigations (Feldesman, 1974, 1976, 1982, 1986) demonstrated that these measurements are very effective in sorting hominoid humeri on the basis of habitual locomotor behavior.

Recent studies (Schilling, 1997; Feldesman, unpublished findings) showed that the current data set is nonnormal, has grossly unequal within-taxon covariance structures (partly resulting from sample size differences, partly from gross physical size differences), has subtle (but "normal") outliers, and suffers from gross inequalities in sample mixture proportions. These characteristics make the data challenging to use reliably with discriminant analysis, for it is difficult to disentangle the results of the analysis from the problems that afflict the data themselves.

As a baseline for comparing the BFOS algorithm, I first run and report the results of a standard canonical linear discriminant analysis using S-Plus 2000's discrim function (Mathsoft, 1999). I ignored issues of univariate or multivariate normality, covariance heterogeneity, or sample mixture. Since binary, recursive partitioning uses a cross-validation scheme not found in any commercial LDA package, I wrote an S-Plus function (based on Venables and Ripley, 1999) for discrim that implements 10-fold cross-validation in place of ordinary leave-one-out (or jackknife) cross-validation. This allowed me to compare cross-validated LDA classification statistics directly with those from classification trees.

Typically, morphometricians develop prediction equations with LDA and use the classification accuracy statistics to validate the prediction equations: high classification accuracy equates with "valid" prediction models. However, if the data appear to violate one or more crucial assumptions (e.g., neither univariate nor multivariate normality, unequal covariance structures, significant outliers, disparate sample mixture proportions, or missing predictor variables), the classification statistics (both resubstitution and cross-validated) may be untrustworthy, which in turn makes the equations suspect. In these circumstances, it is advisable to compare the LDA results with alternative methods before drawing conclusions about their validity.

For the classification tree analysis, I chose Therneau and Atkinson's (1997) implementation of the BFOS algorithm (rpart) because it conforms closely to the specifications laid out in Breiman et al. (1984). Specifically, I used both Windows and Linux versions of rpart in conjunction with the GNU-S open source statistical package called R (Ihaka and Gentleman, 1996). Both R and rpart are available free for most operating systems at http://www.r-project.org.

As noted earlier, the concept of classification trees does not appear to be well-known to most physical anthropologists. Thus, before presenting my results, I walk the reader through the major rpart calculations for a relatively simple example.

Binary recursive partitioning: an example

When Fisher (1936) developed the LDA, he demonstrated it with data from the taxonomic survey by Anderson (1935) of irises. Anderson (1935) measured four attributes (sepal length, sepal width, petal length, and petal width) on 50 specimens each from three different species of irises: *Iris setosa*, *Iris*

which group incumbency has been established by a host of consistent information additional to skeletal morphology. Since binary trees are not computed from covariance matrices, the number of cases per group is irrelevant; groups having few cases are involved in tree construction from the outset, unlike the situation in LDA (but for an alternative parametric approach involving covariance matrices, LDA, unsupervised classification, and principal components analysis, see van Vark, 1995). The major challenge for any technique that involves "pregrouped" data is how to assign group prior probabilities and misclassification costs. Breiman et al. (1984) discussed this extensively, and the present investigation offers some clues about how prior probabilities affect results. Misallocations under binary trees yield important insights about how fossils are related to extant groups in the study. Similarly, fossils that are "correctly" assigned (i.e., remain in a coherent group) reveal equally significant information about intergroup relationships.

versicolor, and *Iris virginica*. Fisher (1936) used LDA to test a genetic hypothesis that *I. versicolor* was a hybrid two-thirds of the way between the other two species. The iris data of Anderson (1935) are commonly supplied as test data for any modern LDA program, and are nicely suited to demonstrate the BFOS classification tree algorithm.

For simplicity, I used a random half of the data of Anderson (1935). The sample drew unequally from the three species; however I treated the prior probabilities of each group as equal to mimic the real-life problem of unknown population mixtures and opportunistic sampling. This mirrors the typical paleoanthropological investigation, and assumes that a random draw from this sample could be assigned to any one of the 3 species with an equal likelihood. Thus:

$$\pi_{\text{setosa}} = \pi_{\text{versicolor}} = \pi_{\text{virginica}} = 0.3333,$$

where π denotes the prior probability of any group. Such priors mean that 67% of the cases would be misassigned by chance alone. We can express this expectation by a "loss" function that measures the heterogeneity of the sample as a function of the prior probabilities:

$$\text{loss} = 1 - \sum_{\text{i}} \pi_{\text{i}}^2, \text{ where i} = 1 \ldots \text{number of groups.} \quad (1)$$

The expected loss in the *Iris* example can therefore be calculated as $1 - 0.33^2 - 0.33^2 - 0.33^2 = 0.67$, following Eq. (1). A modified version of Eq. (1) utilizing "Bayesian probabilities" (see below; often symbolized by I(Node)) is called the Gini diversity index (Therneau and Atkinson, 1997; Venables and Ripley, 1999). This index measures the "impurity" (i.e., heterogeneity) in the node and $\text{I(Node)} \in [0,1]$. Breiman et al. (1984) recommended that binary splits be chosen to minimize the Gini diversity index.

Binary, recursive trees are easily visualized by examining box plots for each iris measurement across the three species. Obvious division points are evident in the boxplots depicted in Figure 1. *Iris setosa* can be clearly distinguished from the other species by either petal length or petal width.

The BFOS algorithm examines all possible univariate divisions of the data. Discounting duplicates, it iterates over a maximum of 300 possible branch points (75 cases × 4 measurements) to find the one that produces the greatest increase in classification accuracy. For each potential split, cases are moved to the left or right descendant based on the answer to a "yes/no" question, e.g., "Is Petal Length <2.5?" Since the response is unambiguous except if Petal Length is missing, a case must go either left ("yes") or right ("no"). To find the "best" split, the BFOS method calculates the Gini diversity index for each possible bifurcation. From this, it computes the "improvement" in classification accuracy resulting from the proposed decision rule. The winning split combines the greatest reduction in the Gini diversity index with the greatest increase (improvement) in the number of correct assignments. Once the algorithm locates the "optimum" division point, the sample is split at the midpoint between the actual "best" value, and the closest (but larger) recorded value of the same variable. Appendix 1 and Figure 2 depict the actual rpart tree, in two different formats, both confirming that all *setosa* cases are sent to the left when Petal Length <2.5, while non-*setosa* go right with Petal Length ≥ 2.5.

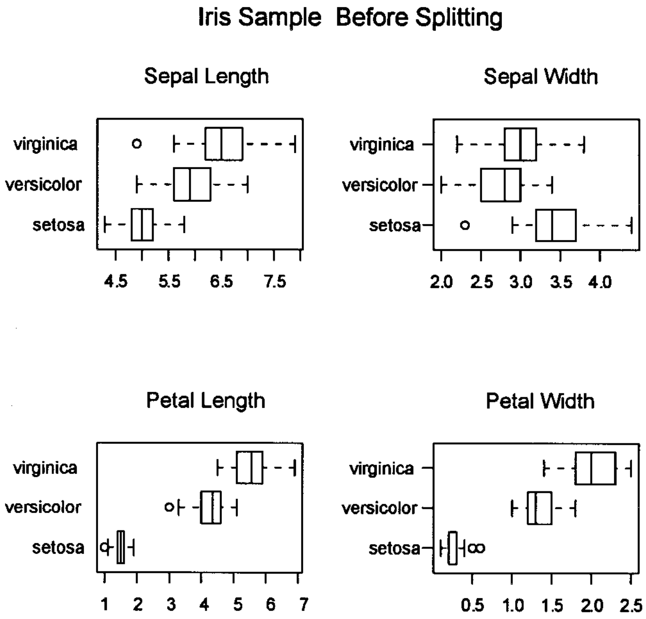

Fig. 1. Box plots of *Iris* species at root node. Note that *Iris setosa* can be distinguished easily from other species on basis of Petal Length or Petal Width.

These calculations (and others) can be expressed as general formulae (see Breiman et al., 1984; Therneau and Atkinson, 1997), but are considerably easier to understand in the context of a specific worked example. I do this below for the BFOS computations relevant to the *Iris* binary tree depicted in Figure 2.

The calculation of the Gini diversity index and the classification "improvement" are essential starting points and depend on information from the problem statement. We are given:

$$\pi_{\text{setosa}} = \pi_{\text{versicolor}} = \pi_{\text{virginica}} = 0.3333$$

$$\text{N} = 75$$

$$\text{n}_{\text{setosa}} = 28; \ \text{n}_{\text{versicolor}} = 22; \ \text{n}_{\text{virginica}} = 25.$$

From this we can compute the Gini diversity index (or impurity) for the unsplit root node:

$$\text{I(Root)} = 1 - \pi^2_{setosa} - \pi^2_{versicolor} - \pi^2_{virginica} = 0.67,$$

from Eq. (1) above.

Subsequent calculations require access to both the condensed summary of the numerical information (Appendix 1), and the actual classification tree (Fig. 2) with branches denoted by the splitting criterion. Our immediate interest is in nodes 2 and 3, which

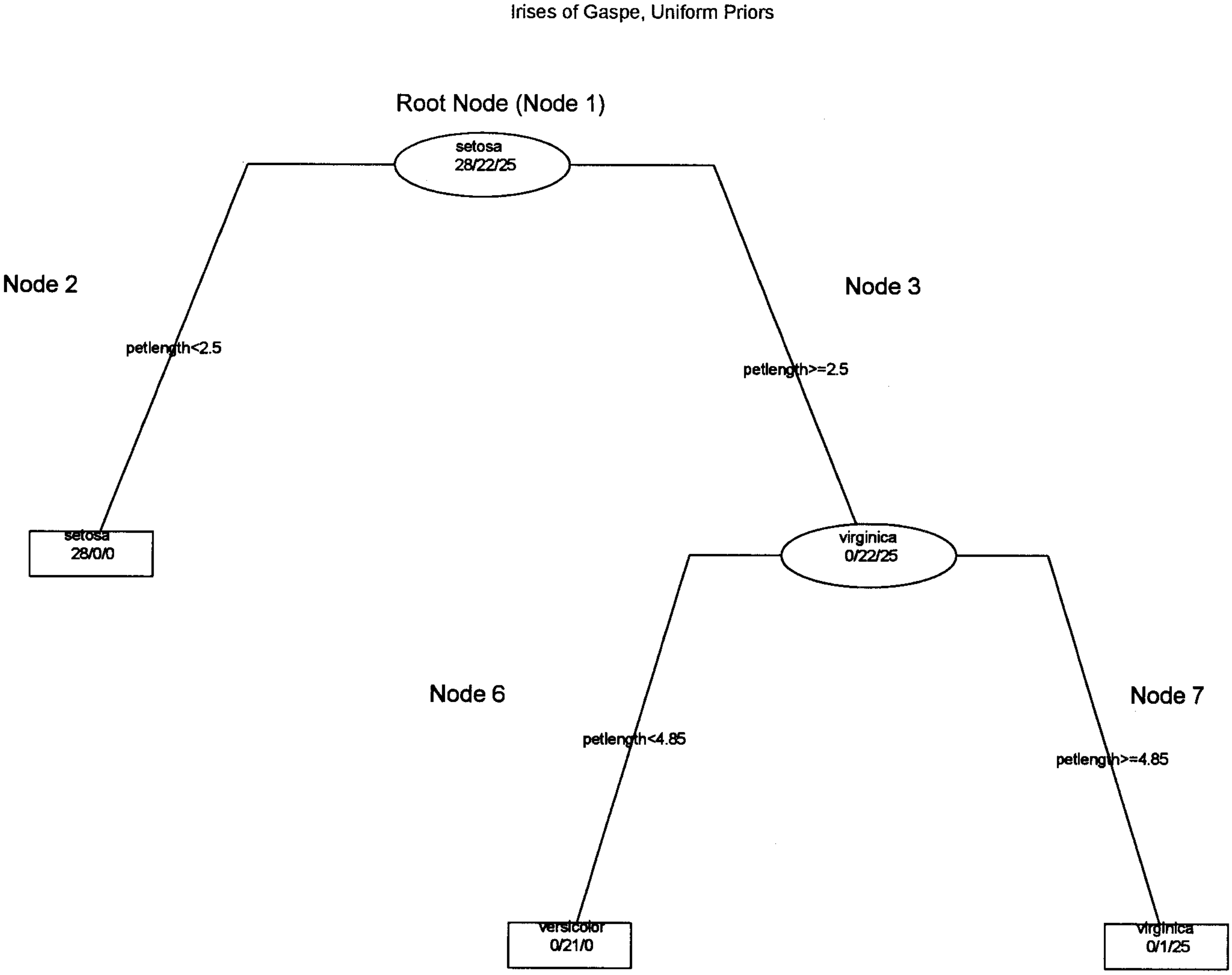

Fig. 2. Complete classification tree of *Iris* species generated from rpart. Terminal nodes are symbolized by rectangles; nonterminal nodes, by ovals. Splitting criteria specified on each branch. Numbers below taxon represent number of cases assigned to each taxon within the node. The convention in rpart is to number the "root node" as 1 and splits as n and n + 1; node numbers do not reflect an actual count of the number of splits. Thus, the split of node 1 results in nodes 2 and 3. The split of node 3 results in nodes 6 and 7; node 6 splits to 12 and 13; node 103 represents the right child of the split involving node 51, etc.

result from dividing the root node. The summary illustrated in Appendix 1 contains information essential to understand the tree construction. Figure 2 can be drawn entirely from the details of Appendix 1.

More extensive output is available in the full rpart object. It appears here as Appendix 2. This output can be copious for a large tree, but is crucial for understanding the internals of the algorithm. Most of the information in Appendix 2 derives from elementary probability theory.

"Petal Length <2.5" splits the root node. Cases meeting this criterion move left to node 2; the rest move right to node 3. The information reported in Appendix 2 can be replicated from the following computations.

The conditional probability of a case being assigned to node 2, given that it is classified as *Iris setosa* is: $P(2 \mid setosa) = 28/28 = 1.0$.

This is the proportion of the *I. setosa* specimens reaching node 2. Since this node consists entirely of *I. setosa*, there are no additional conditional probabilities to compute. The principle for calculating conditional probabilities in impure nodes is, however, quite straightforward, as we will see for node 3.

Rpart computes two separate node probabilities ("Bayesian probabilities" and "altered priors") but only reports one. Both are needed at various points in tree construction. The current rpart version (3.1.1) prints Bayesian probabilities, which are used to calculate the Gini diversity index. "Altered priors" are node probabilities that have been scaled both for initial prior probabilities and by the proportion of the original cases that actually reach the descendant. These are not printed, but are important for computing the "expected loss" (probability of misclassification, adjusted for prior probabilities). This calculation is illustrated below for node 2:

$$p_{\text{setosa(altered)}} = (\pi_{\text{setosa}} \times (P(2|\text{setosa})))/(n_2/N) = (0.33 \times 1.0)/(28/75) = 0.8929 \quad (2)$$

Since node 2 consists of only one species, it is "pure" (homogeneous), cannot be split further, and has an expected loss of 0.0. Nodes that include more than one group have misclassifications and a nonzero expected loss, calculated as:

$$L(\text{node}) = \sum_{i=1..n} p_{i(\text{altered})} - \max(p_{i(\text{altered})}) \quad (3)$$

Rpart also calculates unconditional probabilities for all nodes, i.e., the likelihood that any case ends up in a particular node. This can be calculated for node 2 from the total probability theorem as:

$$P(2) = \pi_{\text{setosa}} \times P(2|\text{setosa}) + \pi_{\text{versicolor}} \times P(2|\text{versicolor}) + \pi_{\text{virginica}} \times P(2|\text{virginica}).$$

Substituting all the previously calculated and given values, $P(2) = 0.3333$.

From Bayes' Theorem, we can also calculate the probability of an *Iris setosa* given that the specimen is in Node 2. Bayesian probabilities are those in parentheses in Appendix 1, listed as "probabilities" within each node in Appendix 2, and are used to calculate the Gini Diversity Index. The highest Bayesian probability at each node is also used to determine the node classification, assuming all descendant nodes were pruned away:

$$P(\text{setosa}|2) = \pi_{\text{setosa}} \times P(2|\text{setosa})/P(2).$$

Not surprisingly, $P(\text{setosa}\,|\,2)$ is 1.0, since all node 2 cases are *I. setosa*.

This yields enough information to determine node heterogeneity. Recall that the algorithm cycles through every possible binary split, making all these calculations for each one, before choosing the "winning" variable and its value. At each, an impurity measure is calculated to assess the extent to which the node is heterogeneous. The winning split has the lowest summed impurities across both descendant nodes. node 2 impurity is calculated as:

$$\text{Impurity}(2) = I(2) = 1 - P(\text{setosa}\,|\,2)^2 = 0.0.$$

The impurity measure for node 3 (and all subsequent nodes) is computed exactly as for node 2, but heterogeneity increases the number of individual calculations. For node 3, these are:

$$P(3\,|\,\text{setosa}) = 0/28 = 0.0 \text{ (altered prior} = 0.0)$$

$$P(3\,|\,\text{versicolor}) = 22/22 = 1.0 \text{(altered prior} = (0.33 \times 1.0)/(47/75) = 0.5319)$$

$$P(3\,|\,\text{virginica}) = 25/25 = 1.0 \text{(altered prior} = (0.33 \times 1.0)/(47/75) = 0.5319)$$

By using the "altered priors" formulation (2), the expected misclassification proportion ("expected loss") for node 3 is:

$$L(3) = (0.5319 + 0.5319) - 0.5319 = 0.5319,$$

which shows in Appendix 2 as "expected loss." It is scaled to node size in Appendix 1 (i.e., 47 × L(3) = 25).

The unconditional probability of Node 3 is:

$$P(3) = \pi_{\text{setosa}} \times P(3\,|\,\text{setosa}) + \pi_{\text{versicolor}} \times P(3\,|\,\text{versicolor}) + \pi_{\text{virginica}} \times P(3\,|\,\text{virginica}).$$

Making the relevant substitutions, $P(3) = 0.67$. This leads to the Bayesian probabilities of each species given their assignment to node 3:

$$P(\text{setosa}\,|\,3) = \pi_{\text{setosa}} \times P(3\,|\,\text{setosa})/P(3) = 0.0$$

$$P(\text{versicolor}\,|\,3) = \pi_{\text{versicolor}} \times P(3\,|\,\text{versicolor})/P(3) = 0.5$$

$$P(\text{virginica}\,|\,3) = \pi_{\text{virginica}} \times P(3\,|\,\text{virginica})/P(3) = 0.5$$

Finally, node 3 impurity is computed as:

$$I(3) = 1 - P(\text{setosa}\,|\,3)^2 - P(\text{versicolor}\,|\,3)^2 - P(\text{virginica}\,|\,3)^2 = 0.5$$

These calculations are preliminary to the computation of overall "improvement." "Improvement" is always measured relative to the parent node. Since nodes 2 and 3 are children of node 1 (the root), relative improvement (RI) is compared to the root node

$$RI = I(R) - P(2) \times I(2) - P(3) \times I(3),$$

Substituting all computed values (0.67 − 0.33 × 0.0 − 0.67 × 0.5), RI = 0.33. The absolute improvement (25) is scaled to overall sample size. This indicates that the maximum increase in correct classifications over the root node by splitting at petal length <2.5.

Because node 3 is 50% impure and includes more than 5 cases,[4] rpart splits it again. Nodes 6 and 7 represent the bifurcation of node 3. Figure 3 depicts the boxplots for cases specifically assigned to node 3, and again shows petal length to be the probable binary "splitter" separating the remaining *Iris* species.

To move from node 3 to nodes 6 and 7 requires the same computational sequence as described above. The principal difference is that "improvement" is measured relative to node 3 instead of to the root node.

To complete the example, the key computations are:

$$P(6\,|\,\text{setosa}) = 0/28 = 0.0 \text{ (altered prior } = \ 0.0)$$

$$P(6\,|\,\text{versicolor}) = 21/22 = 0.9546 \text{ (altered prior} = (0.33 \times 21/22)/(21/75) = 1.136$$

[4] This defaults to 5, but is under user control. In circumstances where there are a priori groups with fewer than 5 cases (a fossil taxon, perhaps) the user might want to set this to a lower value.

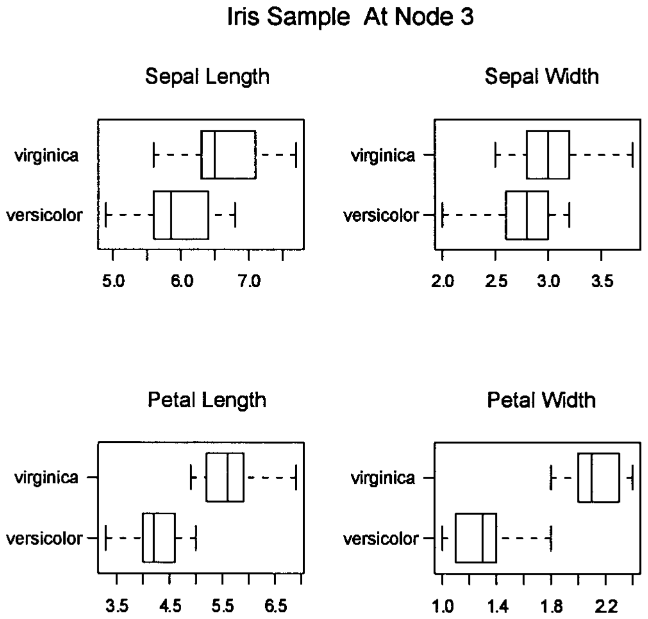

Fig. 3. Box plots of *Iris* species after splitting root node and removing all *Iris setosa* specimens.

$$P(6 \mid \text{virginica}) = 0.25 = 0.0 \text{ (altered prior} = 0.0)$$

$$P(6) = \pi_{\text{setosa}} \times P(6 \mid \text{setosa}) + \pi_{\text{versicolor}} \times P(6 \mid \text{versicolor}) + \pi_{\text{virginica}} \times P(6 \mid \text{virginica}) = 0.318$$

$$P(\text{setosa} \mid 6) = \pi_{\text{setosa}} \times P(6 \mid \text{setosa})/P(6) = 0.0$$

$$P(\text{versicolor} \mid 6) = \pi_{\text{versicolor}} \times P(6 \mid \text{versicolor})/P(6) = 1.0$$

$$P(\text{virginica} \mid 6) = \pi_{\text{virginica}} \times P(6 \mid \text{virginica})/P(6) = 0.0$$

$$I(6) = 1 - P(\text{setosa} \mid 6)^2 - P(\text{versicolor} \mid 6)^2 - P(\text{virginica} \mid 6)^2 = 0.0$$

$$\text{Loss}(6) = (0.0 + 1.0 + 0.0) - 1.0 = 0.0$$

The calculations for Node 7 are:

$$P(7 \mid \text{setosa}) = 0/28 = 0.0 \text{ (altered prior} = 0.0)$$

$$P(7 \mid \text{versicolor}) = 1/22 = 0.0455 \text{(altered prior} = (0.33 \times 1/22)/(26/75) = 0.0437)$$

$$P(7 \mid \text{virginica}) = 25/25 = 1.0 \text{(altered prior} = (0.33 \times 25/25)/(26/75) = 0.9615)$$

$$\text{Loss}(7) = (0.0437 + 0.9615) - 0.9615 = 0.0437$$

$$P(7) = \pi_{\text{setosa}} \times P(7 \mid \text{setosa}) + \pi_{\text{versicolor}} \times P(7 \mid \text{versicolor}) + \pi_{\text{virginica}} \times P(7 \mid \text{virginica}) = 0.3485$$

$$P(\text{setosa} \mid 7) = \pi_{\text{setosa}} \times P(7 \mid \text{setosa})/P(7) = 0.0$$

$$P(\text{versicolor} \mid 7) = \pi_{\text{versicolor}} \times P(7 \mid \text{versicolor})/P(7) = 0.04348$$

$$P(\text{virginica} \mid 7) = \pi_{\text{virginica}} \times P(7 \mid \text{virginica})/P(7) = 0.95652$$

$$I(7) = 1 - P(\text{setosa} \mid 7)^2 - P(\text{versicolor} \mid 7)^2 - P(\text{virginica} \mid 7)^2 = 0.0832$$

The "improvement" combines results from nodes 6 and 7 and is measured relative to node 3.

$$\text{Improvement} = I(3) - (P(6) \times I(6))^2 - (P(7) \times I(7)^2.$$

Substituting accordingly, $(0.5 - (0.318 \times 0.0)^2 - (0.3485 \times 0.0832)^2)$, the relative improvement is 0.304345, and the absolute improvement is I(3) × 75 = 22.826, as reported.

Theoretically, it is possible to continue splitting impure nodes until all cases are classified. However, this typically produces trees that are quite "bushy," hard to interpret, and harder to generalize to new data. Since the goal of classification trees is to generalize to unknowns, we want a tree that is "trained" to our learning dataset, but not so fine-tuned that the results generalize poorly to new data.

Breiman et al. (1984) demonstrated that trees with distant terminal branches snipped off ("pruned") typically generalize better to unknown data. As a result, they introduced the concept of tree pruning via "cost-complexity." This is now the primary way all recursive partitioning methods cut trees down to size. The principle is simple. As trees grow, they become more complex, while misclassification error rates decline. Breiman et al. (1984) proposed a strategy that attempts to balance the number of terminal nodes with the misclassification error rate. This is patterned after regression models that try to optimize the error sum of squares and the number of parameters. Models are penalized for additional parameter. By extension, Breiman et al. (1984) suggested penalizing models each time another split occurs. If the additional split does not improve the fit enough to overcome the penalty (the tree "cost"), the smaller tree is selected. Breiman et al. (1984) proposed this cost-complexity parameter (CP) as:

$$\text{CP} = \text{Training Misclassification Rate} + \alpha \times (\text{Number of Terminal Nodes}),$$

where α is the penalty for each additional terminal node. If α is 0, there is no penalty for additional nodes. If α is set arbitrarily large (up to ∞), the root node is preferred because it is the smallest tree. Thus, when α lies between 0 and ∞, different trees are selected.

In rpart, CP is computed as α / (root node relative error) (Venables and Ripley, 1999). This makes CP and α equivalent, since the root node relative error

TABLE 1. Complexity parameter table for Iris *classification tree*

	CP	Split	rel error	xerror	xstd
1	0.5000000	0	1.00000000	1.20467532	0.07127029
2	0.4772727	1	0.50000000	0.80922078	0.09055846
3	0.0100000	2	0.02272727	0.08545455	0.04166163

is normalized to 1.0. Rpart determines the complexity parameter when it calculates "improvement" and the resulting absolute and relative error. The absolute root node error is 0.67, as shown earlier. A single binary split of the root node decreases the absolute error to 0.33 and the relative error to 50%. From this, CP can be computed as:

$$CP_i = (RE_i - RE_{i+1})/((nsplit_{i+1}) - (nsplit_i)) \quad (4)$$

For the iris data, the CP table (Table 1) shows the root node CP as simply (1.0 − 0.5)/(1 − 0) = 0.5. Furthermore, after 1 split (2 branches), the CP is (0.5000 − 0.02272727)/(2 − 1) = 0.4772727.

The CP can be interpreted as the improvement in fit compared to a tree with one less split. The chief reason for using it is to "cost out" the improvement in fit resulting from the addition of another split in the data. It also measures the accuracy lost by removing one or more terminal nodes. An *Iris* tree with one terminal node decreases classification error by 50% over the root node; a tree with 2 terminal nodes improves the fit an additional 47.7%, while there is only a small gain from a third terminal node. By default, rpart stops growing trees when CP = 0.01.

The BFOS algorithm uses 10-fold cross-validation to determine whether and where to prune the tree. CP is useful for determining pruning sequences and for identifying potential "best trees," but it is not helpful for selecting the "best tree;" larger trees always yield better fits than smaller trees. The ideal solution is to use one sample to grow the tree, and a second sample to test it. Pruning is then based on the error rate of the test sample. Unfortunately, few of us have "spare" data. Cross-validation finesses this problem by splitting the original sample into k (10 in BFOS) parts, and combines k − n of these parts into a learning sample, and uses the remainder (n) as a test sample. The learning sample is used to develop trees, which are validated with the test sample. This process is repeated over all possible splits of the data into k − n and n partitions. By averaging all k test results, the final classification is a single, unbiased estimate of the error rate in the full, unsplit sample (Hastie et al., 2001).

Rpart reports xerror and xstd, the pooled 10-fold cross-validated error rate and the cross-validated standard deviation, for all trees of a particular size. The final tree size can be chosen to minimize xerror. This is typically the largest tree. Breiman et al. (1984), Therneau and Atkinson (1997), and Venables and Ripley (1999) recommended the 1 SE rule, which favors the largest tree with xerror within 1 standard deviation of the minimum. In this example, the 1 SE rule gives 0.0855 + 0.0417 = 0.1272. Since the tree with two leaves (line 2) has xerror = 0.8092, while the full tree has xerror = 0.0855, we could arguably select a tree with only two terminal nodes. However, since 1 SE finds xerror so close to the maximal tree (and so far from the next smaller tree), the larger tree is probably a better choice.

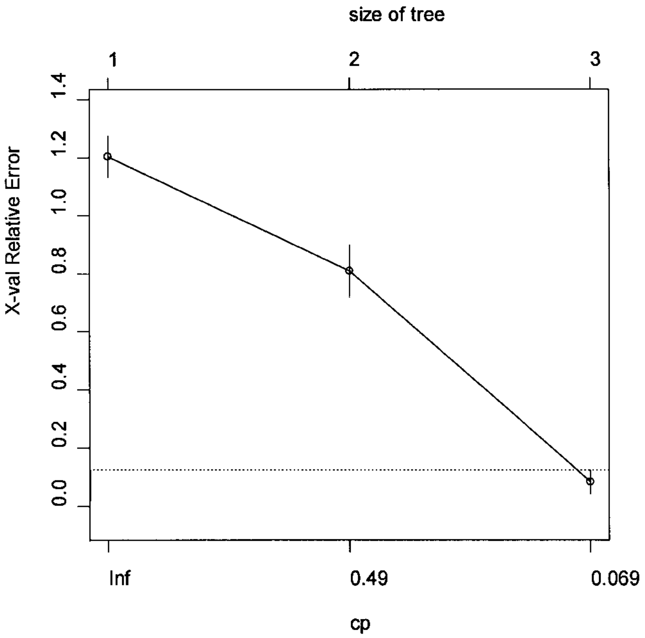

Fig. 4. Rpart plot of complexity parameter vs. cross-validated relative error and tree size for *Iris* data. Horizontal dotted line represents the 1 SE pruning point recommended by Breiman et al. (1984). Because the cross-validated error intersects 1 SE line at a complexity parameter only slightly smaller than the full (3 terminal node) tree, the larger tree is preferred.

TABLE 2. Resubstitution classification matrix for full Iris *classification tree*[1]

	1	2	3
Iris setosa	28	0	0
Iris versicolor	0	21	1
Iris virginica	0	0	25

[1] 98.7% correctly classified.

A visual tool in rpart allows for selecting the "best" tree. It plots CP against xerror and tree size, and places a horizontal line at the 1 SE boundary. The optimal tree size can be read directly from the graph by finding the largest tree with xerror not crossing the horizontal line. In Figure 4, xerror crosses just slightly to the left of a tree with 3 terminal nodes, our largest tree. This is further justification for choosing the maximal tree.

Table 2 shows the resubstituted classification matrix for the full *Iris* tree. Only one case (*I. versicolor*) is misclassified.

These results are biased because the cases classified are exactly the same cases used to grow the tree. By default, rpart also generates a classification matrix for the pooled 10-fold cross-validation results,

TABLE 3. Cross-validation classification matrix for full Iris *classification tree*[1]

	1	2	3
Iris setosa	28	0	0
Iris versicolor	0	19	3
Iris virginica	0	0	25

[1] 96% correctly classified.

which is an independent and unbiased estimate of the prediction error rate (Table 3).

These 10-fold cross-validation results barely differ from the full tree. Two additional specimens, both also from *I. versicolor*, are misclassified as *I. virginica*. Since Fisher's original hypothesis tested whether *I. versicolor* was a hybrid between *I. setosa* and *I. virginica* (but closer to *I. virginica*), these results are not unexpected.

Classification is quite straightforward when all cases have complete sets of predictor variables. Both known and unknown cases are literally "dropped" down the tree, following a binary path until they find a terminal node, which determines class assignment. Surrogate variables, which appear in Appendix 2, are relevant only if there are cases with missing primary split variables. Surrogate splitting variables are alternatives to the primary, chosen for their concordance with the primary splitting variable results on a case-by-case basis. Other variables may correlate more strongly with the primary splitter, or may classify more cases correctly than an alternative (nonprimary) splitter. However, surrogates get selected for how well they mimic the primary splitter behavior. When a primary splitter is missing, the algorithm merely selects the highest-ranking, nonmissing surrogate variable and uses its value to move a case down the tree. This approach is much simpler than imputing missing values, or deleting cases with missing predictors.

The calculations detailed for the *Iris* example have exact (if more complicated) parallels with the hominoid data. The hominoid results below are obtained in exactly the same way as with the *Iris* material. Only the computational details are omitted. As with the *Iris* sample, I assigned equal prior probabilities to the four hominoid taxa. Although the sample mixture is weighted heavily toward *Pan gorilla* and *Pan*, its composition is completely opportunistic, and there is no indication of the true frequency of any of these taxa in real populations.

After subjecting the full hominoid data set to LDA and rpart, I then applied rpart to a modified version of the hominoid data that had missing predictor variables. I selected 10% of the sample (25 cases) at random, and then applied 1 of 6 different missing value combinations randomly to these cases. These variables are frequently missing from fragmentary fossil humeri, or from incomplete modern forms. LATSUPRI requires a significant portion of the diaphysis and distal humerus to measure accurately, while HUMLENGT requires the entire bone. The latter measurements (size driven, of course) figure prominently in LDA results, and as primary splitters in the classification tree. Both are commonly missing in fragmentary fossils, especially those preserving only the distal humerus. On the distal humerus, the medial epicondyle is frequently broken off. When this occurs, BIEPI and MEDEPICO are unmeasurable, and measuring OLECRDEP may be problematic. As is shown, BIEPI, LATSUPRI, OLECRDEP, and HUMLENGT are main splitting variables, while MEDEPICO is a surrogate splitter. Paleoanthropologists typically face data like this, and this simulation will give some indication of how well the BFOS algorithm behaves under such conditions. There is no directly comparable simulation using LDA. In LDA, data imputation would be a necessary antecedent to predicting class assignments of cases with missing predictors. While this comparison would be interesting, it would be difficult to evaluate the results, since BFOS uses only available data, while the LDA would use cases made "whole" statistically.

TABLE 4. Coefficients of S-Plus canonical variates for 10-variable analysis

Variable	CV 1	CV 2
LATSUPRI	0.0479995	0.0479254
MEDEPICO	−0.0740558	0.0940120
PDHTCAPI	0.0041735	0.2970626
MLHTCAPI	0.2783228	0.0143315
APHTTROC	−0.1012576	−0.1714853
MLHTTROC	−0.0840499	−0.0955087
ANTARTBR	0.1109658	−0.0087553
OLECRDEP	−0.0745032	−0.2027351
HUMLENGT	−0.0383797	0.0267279
BIEPI	−0.0615746	−0.0999365
Eigenvalue	5.494	1.259
% trace	79.6%	18.3%

TABLE 5. Resubstitution classification statistics for LDA[1]

	1	2	3	4
Gorilla	82	0	0	4
Pan	0	90	0	14
Pongo	0	0	22	1
Homo	0	4	1	19

[1] 89.9% correctly classified.

RESULTS

Table 4 summarizes the results of the hominoid humerus LDA. It provides the canonical variates as well as the eigenvalues for each of the relevant canonical axes. The latter indicate the relative importance of each canonical axis. The LDA resubstituted classification statistics for these data are detailed in Table 5, which shows a very significant number (89.9%) of hominoid humeri correctly classified.

Figure 5 depicts a two-dimensional scatterplot of canonical axes 1 and 2, which visually demonstrates the group separations in discriminant space. Note that the confidence ellipses superimposed over the group centroids are not all oriented in the same

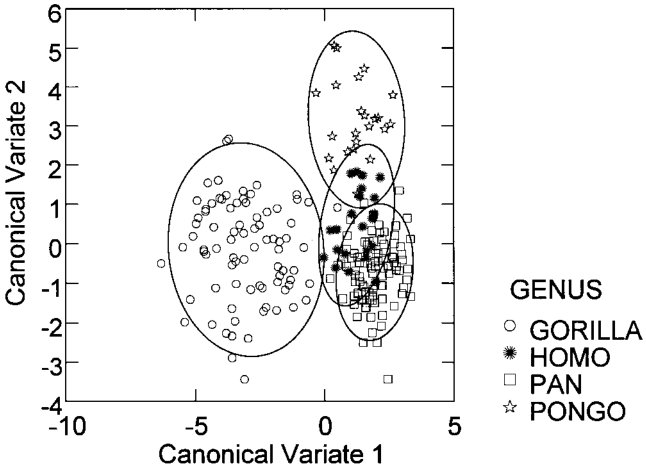

Fig. 5. Bivariate scatterplot of canonical variates 1 and 2, generated using the S-PLUS function discrim. Confidence ellipses surrounding each group centroid encompass 90% of the bivariate means for each group.

direction; *Homo* and *Pan* (the taxa with the smallest individuals) have major axes oriented about 20° away from those of *Pongo* and *Pan gorilla*. This visually confirms that the group covariance structures are unequal (Wilkinson et al., 1996; Ripley, 1996). Statistical confirmation of covariance inequality comes from a highly significant Box's M statistic (not reported).[5]

The eigenvalues and percent trace in Table 4, and the classification matrix in Table 5, show that LDA easily separates the four groups, and correctly classifies the vast majority of cases. Nearly 90% of the cases are resubstituted correctly: a convincing number in the face of nonnormality, covariance heterogeneity, and significant disparities in sample mixture. The resubstitution results also show that all groups have fairly high correct classification rates. Humans are misclassified most frequently, while *Pongo* and *Pan gorilla* have the fewest misclassifications.

The 10-fold cross-validation results (Table 6) provide additional strong support of the power of LDA. They show a small decline in classification accuracy, from 89.9% to 86.9%.[6] This decline is not evident under ordinary leave-one-out cross-validation, which barely changes from the resubstitution statistics (89.2%, but not tabulated here). The 10-fold cross-validation distributes misclassifications across all taxa, but *Pan gorilla* and *Homo* misclassifications are slightly more prevalent. The cross-validation reveals that "new" data would not perform as well as the more optimistically classified full training set, but the results are not dramatically different from the resubstitution statistics. This is reassuring information when the goal is to classify new and/or unknown data.

TABLE 6. Ten-fold cross-validation classification results for LDA[1]

	1	2	3	4
Gorilla	79	0	1	6
Pan	0	89	2	13
Pongo	0	1	21	1
Homo	0	5	2	17

[1] 86.9% correctly classified.

The rpart results compare quite favorably to the LDA. Figure 6 depicts the full rpart tree built from the complete hominoid data set, while Appendix 3 presents the detailed statistics from the tree. The intermediate and terminal nodes in Figure 6 are clearly labeled, and the classification rules are easy to decode.

The full binary tree in Figure 8 consists of 7 terminal nodes. The terminal node classifications (at nodes 2, 7, 13, 24, 50, 102, and 103), indicate 215 cases (91%) correctly assigned. This is slightly better than the resubstitution classification rate under LDA.

LATSUPRI, HUMLENGT, BIEPI, MLHTTROC, and OLECRDEP are the primary splitting variables for this tree; HUMLENGT is used twice. Under 10-fold cross-validation (Table 7), classification accuracy declines to 85%, which is a slightly steeper drop than occurs with cross-validated LDA.

The classification tree training sample error does not single out any one group for frequent misclassification. This can be seen easily in Figure 6; the misclassification rates for all groups except *Pan gorilla* (2%) range from 9–14%. The classification tree does a better job classifying humans correctly (88%) than does LDA (80%). Under cross-validation (Table 7), *Pan gorilla* and *Pan* classification accuracy declines slightly, while *Homo* and *Pongo* drop more significantly. The former drops from 12% misclassified to 33% misclassified under cross-validation, while the latter declines from 9% misclassified to 31% misclassified. This is in significant contrast to LDA, where there is a more uniform distribution of misclassification errors moving from resubstitution to cross-validation. It is tempting to link this high cross-validation misassignment rate to the original small sample sizes of *Homo* and *Pongo*. While this is certainly possible, more research is required to determine the relationship between sample size and classification error.

As noted earlier, one advantage of binary-recursive partitioning is the opportunity to consider smaller models, often making use of fewer variables. Figure 7 offers insight into how the tree could be pruned. Using the 1 SE criterion recommended by Breiman et al. (1984), the 7-terminal node tree could

[5] I am well aware that Box's M is highly sensitive to sample size disparities, which makes its interpretation tricky.

[6] Since 10-fold cross-validation results can vary with the way the sample is partitioned, I ran the analysis 10 times. The classification accuracy varied from 83–89%, with an 87% average.

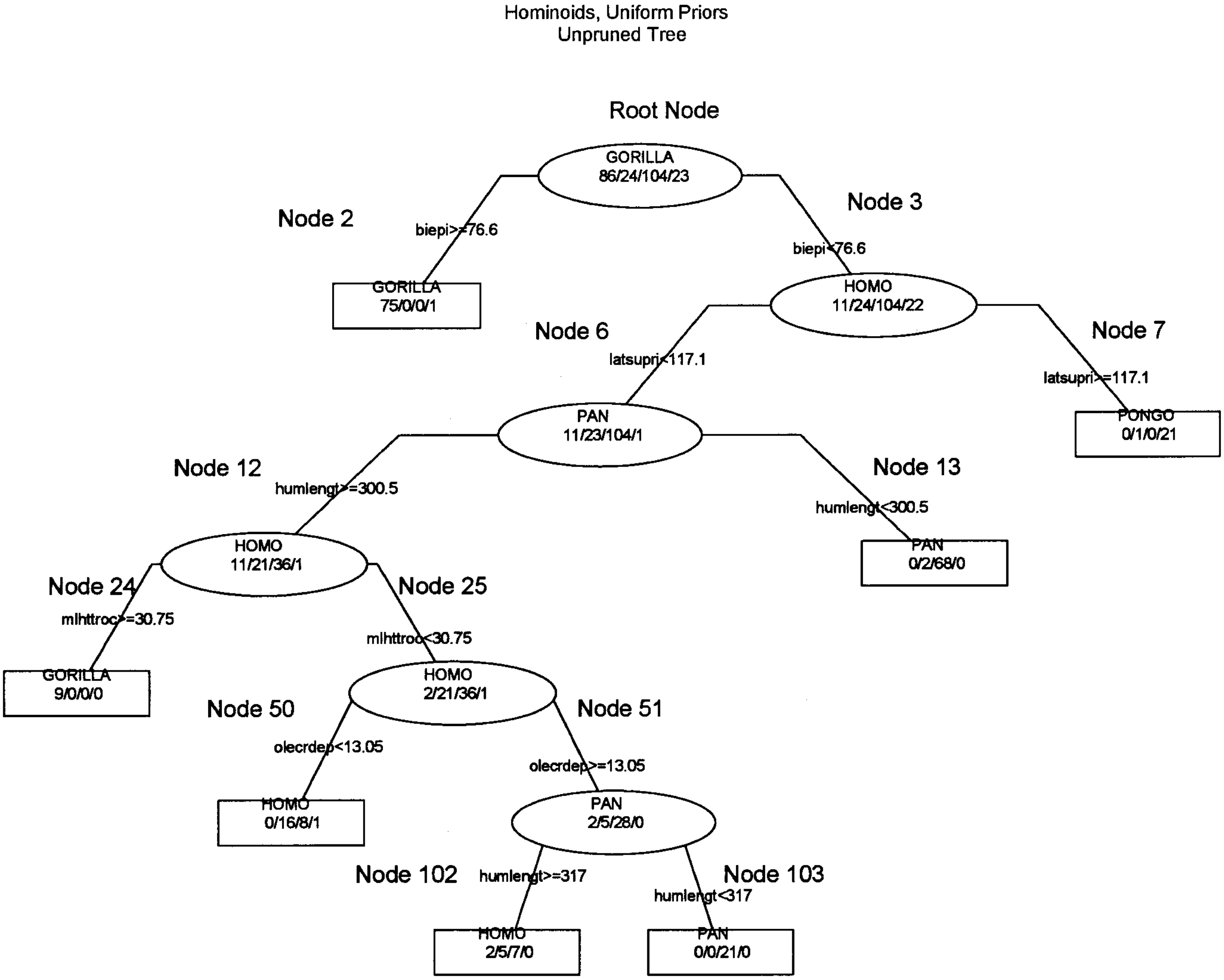

Fig. 6. Full rpart tree for hominoid data. Node labelling follows the convention used in rpart, and corresponds with information both in the text and in Figure 2.

TABLE 7. Ten-fold cross-validation results for Hominoid classification tree[1]

	1	2	3	4
Gorilla	82	4	0	0
Homo	2	16	6	0
Pan	0	16	88	0
Pongo	3	4	0	16

[1] 85.2% correctly classified.

be pruned down to a 4-leaf tree.[7] Accordingly, we could obtain a tree with 4 leaves by snipping off nodes 50, 51, 102, and 103. Although it is not figured here, it is quite easy to imagine. The resulting tree leaves 3 genuinely terminal nodes, with node 25 as a very heterogeneous leaf, classified as *Homo* on the basis of Bayesian probabilities. This has two effects. First, OLECRDEP is eliminated as a splitting variable. Second, the classification accuracy for *Pan* declines significantly, while the other 3 taxa retain their prepruning classification accuracy. This permits a researcher to trade uncertainty in classifying *Pan* for reliability in classifying the other taxa using fewer measurements.

Table 8 summarizes the missing values simulation. Of the 25 specimens missing one or more crucial variables, 21 were correctly classified: an initial accuracy rate of 84%. *Pan gorilla* is never misclassified, regardless of missing predictor. This suggests that size is so pervasive a factor for *Pan gorilla* that misassignment is nearly impossible. Only one error involves *Pongo*; the rest are either *Pan* or *Homo*, the two physically smallest and most size-similar taxa. Three of the four misclassifications occur when both LATSUPRI and HUMLENGT are missing; however, four other cases missing both variables are correctly

[7] Note the terminology here. All terminal nodes are leaves, but not all leaves are terminal nodes. A leaf is an endpoint of a tree that has either been split as far as it can be (a terminal node, by definition) or is at the end because tree-growing has been stopped prematurely via pruning. In this case, the leaf may be more heterogenous than homogenous.

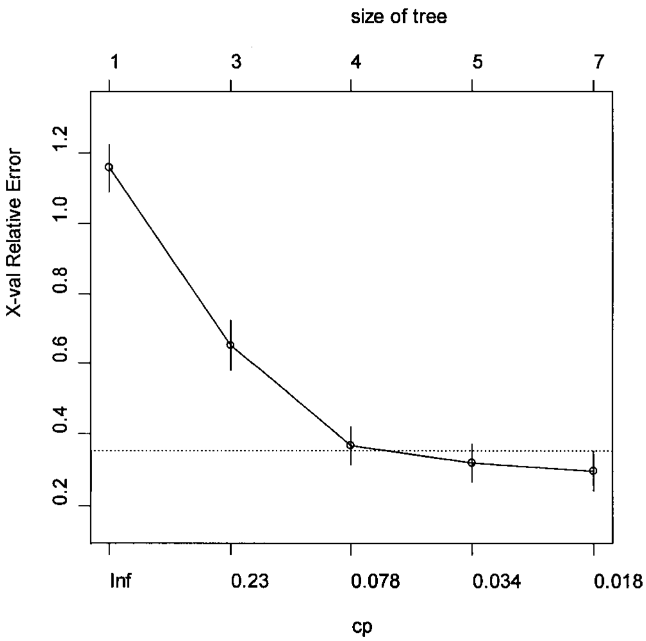

Fig. 7. Complexity parameter plot for hominoid classification tree. Labelling follows that of Figure 4. This suggests that a tree with 4 splits (5 nodes) is adequate. Such a tree would prune nodes 50, 51, 102, and 103 from the tree depicted in Figure 6.

classified. It is hard to argue from this that these variables are particularly significant for classification accuracy. Moreover, one of the misclassified cases is also misclassified with no missing data, and had no chance of being correctly assigned with fewer variables. The remaining misclassification involves LATSUPRI and BIEPI. Since these two variables are also missing in three other correctly classified cases, no generalizations are possible here either. The only relevant message is that even with key splitting variables missing, a high percentage of cases are correctly classified.

DISCUSSION

The differences in results between traditional LDA and nonparametric binary, recursive classification trees are relatively small here. In the hominoid sample, classification trees have a slight (but insignificant) advantage in classifying the full data set, while LDA has a slight (but insignificant) advantage under 10-fold cross-validation. The similarity of results begs an important question: when or why should we choose binary recursive trees over LDA? If our data fail to meet the requirements for LDA, then we cannot be confident in the classifications it produces, *unless* we compare them with results from another technique that requires either fewer or different assumptions. In principle, only then can we decide which technique best fits the data. The importance of the above cannot be overstated. To use LDA for classification, researchers are obliged to test their data for conformity to normal-theory assumptions. If the data meet the necessary conditions, there is nothing to be gained by using a less powerful nonparametric technique. If the data do not meet LDA's requirements, then the researcher should consider alternatives.

What happens if data do not meet the assumptions necessary for LDA, but when both LDA and classification trees yield similar results, as they do here? As it happens in the hominoid data set, neither approach trumps the other: BFOS is slightly better under resubstitution; LDA is better under cross-validation. Since both give excellent fits to the data, it probably doesn't matter which we choose, but power considerations favor the parametric model. With poorer fits, another procedure might be required to sort out subtle differences. Steinberg and Colla (1997) suggested combining the canonical variates (from LDA) with the raw predictor variables for each case into a single data set. They then subjected the combined data to a classification tree analysis, using the BFOS algorithm. If the canonical variates are unimportant, they will not be key splitting variables in the resulting tree. This favors the original binary tree results. On the other hand, if the canonical variates are key splitters while few, if any, original variables participate, then LDA is the best choice.

I performed Steinberg and Colla's (1997) tests by combining the 10 original humerus variables with the 3 canonical variates generated by LDA. Not surprisingly, the first two canonical variates are primary splitter variables at 3 of 4 major branches on the classification tree. Of the original variables, only HUMLENGT contributes to the new tree. This strengthens the case for LDA. The combined tree resolves the data in fewer terminal nodes (5 vs. 7), and is more accurate (93% resubstitution accuracy, and 87% cross-validated accuracy) than either analysis involving only the raw data. This suggests that LDA is the appropriate technique for these data. It also supports the assertion that LDA is robust to violations of required assumptions. This is precisely how Johnson and Wichern (1998, p. 665) defined robustness, as "the [resistance to] deterioration in error rates caused by using a classification procedure with data that do not conform to the assumptions on which the procedure was based." It is tempting to conclude that LDA is sufficiently robust for all classification problems, but we cannot use this to avoid testing the data. This claim is cannot be substantiated without the comparative analysis.

When predictor variables are missing, the answer is more complicated. Classification trees make no demands on the researcher confronted with an incomplete data set. If the missing values are not primary splitters, their status is irrelevant. The surrogates get used only when a primary splitting variable is unavailable. Classification results would not change, even if we had deleted every nonprimary splitting variable. The advantage of surrogacy is that the algorithm can recover if it encounters a missing primary variable. With strong concordance

TABLE 8. Classification tree analysis, using randomly selected cases with randomly assigned missing value combinations

Case number[1]	Correct classification	Classification in full analysis	Classification with missing values[2]	Missing variables
5	*Gorilla*	*Gorilla*	*Gorilla*	MEDEPICO, OLECRDEP, BIEPI
10	*Gorilla*	*Gorilla*	*Gorilla*	HUMLENGT, BIEPI, MEDEPICO
12	*Gorilla*	*Gorilla*	*Gorilla*	LATSUPRI, HUMLENGT
20	*Gorilla*	*Gorilla*	*Gorilla*	LATSUPRI, HUMLENGT
33	*Gorilla*	*Gorilla*	*Gorilla*	LATSUPRI, BIEPI
39	*Gorilla*	*Gorilla*	*Gorilla*	MEDEPICO, BIEPI
42	*Gorilla*	*Gorilla*	*Gorilla*	MEDEPICO, BIEPI
43	*Gorilla*	*Gorilla*	*Gorilla*	HUMLENGT, BIEPI, MEDEPICO
57	*Pan*	*Pan*	*Pan*	LATSUPRI, BIEPI
60	*Pan*	*Pan*	*Pan*	HUMLENGT, BIEPI, MEDEPICO
63	*Pan*	*Pan*	*Pan*	HUMLENGT, BIEPI, MEDEPICO
75	*Pan*	*Pan*	*Homo*	LATSUPRI, HUMLENGT
80[1]	*Pan*	*Homo*	*Homo*	LATSUPRI, BIEPI
108	*Pongo*	*Pongo*	*Pan*	LATSUPRI, HUMLENGT
121	*Gorilla*	*Gorilla*	*Gorilla*	LATSUPRI, HUMLENGT
140	*Pan*	*Pan*	*Pan*	MEDEPICO, OLECRDEP, BIEPI
151	*Pongo*	*Pongo*	*Pongo*	LATSUPRI, HUMLENGT
164	*Pan*	*Pan*	*Pan*	LATSUPRI, BIEPI
189	*Pan*	*Pan*	*Pan*	MEDEPICO, OLECRDEP, BIEPI
191	*Pan*	*Pan*	*Pan*	HUMLENGT
204	*Gorilla*	*Gorilla*	*Gorilla*	MEDEPICO, BIEPI
211	*Gorilla*	*Gorilla*	*Gorilla*	MEDEPICO, BIEPI
213	*Gorilla*	*Gorilla*	*Gorilla*	MEDEPICO, OLECRDEP, BIEPI
226	*Homo*	*Homo*	*Pan*	LATSUPRI, HUMLENGT
228	*Homo*	*Homo*	*Homo*	MEDEPICO, BIEPI

[1] Case was also misclassified when all variables were present.
[2] Classifications come from dropping cases with missing values down the original rpart tree generated from the full data set.

between a primary splitter and its surrogate(s), there is a good chance for correct classification. Moreover, since surrogates are selected on the basis of concordance and not correlation with the primary splitter, unrelated or weakly related variables may be surrogates for each other. This makes it more likely to find a reasonable surrogate. For this reason alone, classification trees are very attractive for data with incompletely observed predictors.

Missing predictor data complicate analysis via LDA. Since no conventional discriminant analysis program will use cases with incompletely observed predictor variables, the researcher who wants to maximize the "classification return" on limited data must use some form of data imputation. Regardless of the approach used for data imputation, the imputation step always precedes LDA.[8] Evidence is building that modern imputation techniques are capable of giving good estimates for missing values in certain circumstances (Schafer, 1997; Schimert et al., 2000). When researchers deal with fossil remains, classification is uncertain, the fraction of cases with missing values can be high, and the choice of a reference population can bias classification a priori. If the investigator chooses not to impute, the only LDA alternative is to reconfigure the analysis into a series of discriminants based on separate pieces of the data sets, as Stringer (1974) and Kidder et al. (1992) did. There is nothing inherently wrong with this strategy, but drawing inferences or conclusions from combining serial LDAs may introduce multiple comparison problems that inflate significance levels (e.g., Hsu, 1996). In either instance (imputation or serial discriminants), additional statistical testing is required to assess the results. By contrast, binary recursive trees not only allow the researcher to sidestep the missing values problem altogether, but they are actually designed for dealing with large data sets containing substantial amounts of missing

[8] Except for mean substitution, which can occur in tandem with the LDA.

data. A data set with missing values is an ideal candidate for classification trees, which a simple example illustrates.

The Stringer's (1974) set of hominoid cranial measurements from 131 *Homo erectus*, archaic *Homo sapiens*, Neanderthals, and anatomically modern *Homo sapiens* presents an interesting test of the power of classification trees. Using the first 9 measurements in this collection (GOL, NOL, BNL, BBH, XCB, XFB, AUB, ASB, and BPL) and response variables (taxonomic assignments) provided by Stringer (1974), only 54 cases are available for LDA; the remaining 77 are missing one or more measurement and are deleted casewise. Of the 54 cases, 46 are "modern," and LDA misclassifies only one of these; the remaining 8 cases fall in three other groups, and only two cases are correctly classified. By contrast, rpart assigns all 131 cases. Using surrogate variables, rpart correctly assigns 65% (85 individuals) of the cases, including the same 45 modern humans that LDA assigned correctly. Thus, while rpart handles modern specimens in exactly the same way as LDA, it improves the classification significantly by properly assigning an additional 40 (of 77) cases that standard LDA implementations delete. Not surprisingly, misclassifications occur with specimens missing the largest number of predictors, but classification accuracy could be improved by including additional measurements.

CONCLUSIONS

The analysis performed here demonstrates several important points: 1) data subjected to predictive LDA should always be tested for normality prior to analysis; 2) LDA is an extraordinarily robust technique that should be the first tool used for classification if there are complete data sets that are approximately normal; 3) nonparametric classification trees that use the BFOS algorithm yield excellent results that compare favorably to (or do better than) LDA with training data, and only slightly worse under rigorous cross-validation; and 4) the BFOS algorithm works well on data sets with incomplete observations without requiring data imputation. As demonstrated, even with crucial splitting variables missing from the analysis, the binary tree technique of using surrogate splitting variables allows a remarkably high percentage of cases to be correctly classified.

This leads me to recommend classification trees as an alternative or an adjunct to LDA in two instances. In the first, I recommend it whenever multivariate data depart from the essential normal-theory assumptions of classificatory LDA. When classification trees provide better cross-validated predictions than discriminant analysis, or when raw variables perform better than the canonical variates in a combined tree-structured analysis, I would report the classification tree results, not those from LDA. Either way, however, there is an additional analytical burden for the researcher, but modern software makes this straightforward. Unless these steps are taken, there is no way to assess the predictions resulting from an LDA of nonnormal data.

In the second case, I would strongly recommend classification trees instead of LDA whenever data sets are missing significant information, as they would be with fossils. This is not to discourage researchers from using modern data imputation techniques, but those who use these procedures (especially with fossils) have a special responsibility to understand how the imputations are obtained, and ask whether they make sense. The advantage of binary trees is that investigators are freed from these obligations and also from the need to defend the "sensibleness" of "filled in" values for fossils lacking an adequate reference population from which to develop imputations. Instead, researchers can concentrate on the meaning of class assignments in the context of observed (and defensible) data.

ACKNOWLEDGMENTS

I am grateful for all the assistance I received in various phases of this research. I thank Derek Howlett of the Powell-Cotton Museum (Birchington, UK), Paula Jenkins of the British Museum (Natural History, London, UK), Renate Angerman of the Alexander von Humboldt Museum (Berlin, Germany), W.N. van Neer of the Royal Central African Museum (Musée Royale de l'Afrique Centrale, Tervuren, Belgium), and Jan de Vos (Rijksmuseum van Natuurlijke Historie, Leiden, Netherlands), all of whom permitted me access to skeletal material in their charge. Geoff Kleckner assisted me with data collection during 1987 and 1989. Chris Stringer generously permitted me to use his hominid data set for the missing predictor example in the Discussion. I received invaluable advice and counsel from Drs. Leland Wilkinson (SPSS, Inc.), Charles Roosen, Tim Hesterberg, Bill Dunlap, and Kim Leader (Mathsoft, now Insightful), Warren Sarle (SAS, Inc.), and Dan Steinberg and Mikhail Goloyvna (Salford Systems, Inc.). Professor Brian Ripley (Oxford University) patiently advised me on subtleties of the LDA routines he wrote for S-Plus, as well on how best to exploit the capabilities of both S-Plus and R. I also thank Beth Atkinson and Terry Therneau of the Mayo Clinic for taking time to answer questions about technical details of rpart. Virginia Butler, Ken Ames, Bob Fountain, John Relethford, Geoff Kleckner, Karla Schilling, Susan Wolf, Lyle Konigsberg, and an anonymous reviewer read various drafts of this manuscript and made helpful comments and constructive criticisms.

LITERATURE CITED

Aiello LC, Wood B, Key C, Lewis M. 1999. Morphological and taxonomic affinities of the Olduvai ulna (OH 36). Am J Phys Anthropol 109:89–110.

Anderson E. 1935. The irises of the Gaspé peninsula. Bull Am Iris Soc 59:2–5.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. Classification and regression trees. New York: Chapman and Hall.

Campbell NA. 1984a. Canonical variates analysis—a general model formulation. Aust J Stat 26:86–96.

Campbell NA. 1984b. Canonical variate analysis with unequal covariance matrices: generalizations of the usual solution. Math Geol 16:109–124.

Campbell NA. 1984c. Some aspects of allocation and discrimination. In: van Vark G, Howells WW, editors. Multivariate statistical methods in physical anthropology. Dordrecht, Netherlands: Reidel. p 177–192.

Corruccini RS. 1978. Morphometric analysis: uses and abuses. Yrbk Phys Anthropol 21:134–149.

Feldesman MR. 1974. A multivariate morphometric study of the primate forelimb and elbow complex. Ph.D. dissertation, Department of Anthropology, University of Oregon. Ann Arbor, MI: University Microfilms.

Feldesman MR. 1976. The primate forelimb: a morphometric study of locomotor diversity. University of Oregon Anthropological Papers Number 10. Eugene, OR: University of Oregon Press.

Feldesman MR. 1982. Morphometric analysis of the distal humerus of some Cenozoic catarrhines: the late divergence hypothesis revisited. Am J Phys Anthropol 59:73–95.

Feldesman MR. 1986. The forelimb of the newly "rediscovered" *Proconsul africanus* from Rusinga Island, Kenya: morphometrics and implications for catarrhine evolution. In: Singer R, Lundy J, editors. Variation, culture, and evolution in African populations: papers in honour of Hertha de Villiers. Johannesburg: Witwatersrand University Press. p 179–193.

Feldesman MR. 1997. Bridging the chasm: demystifying some statistical methods used in biological anthropology. In: Boaz N, Wolfe L, editors. Biological anthropology: the state of the science, 2nd ed. Corvallis, OR: IIHER and Oregon State University Press. p 73–100.

Fisher RA. 1936. The use of multiple measurements in taxonomic problems. Ann Eugen 10:422–429.

Flury B. 1997. A first course in multivariate statistics. New York: Springer.

Hastie T, Tibshirani R, Friedman J. 2001. The elements of statistical learning. New York: Springer.

Holliday TW. 2000. Evolution at the crossroads: modern human emergence in Western Asia. Am Anthropol 102:54–68.

Hsu J. 1996. Multiple comparisons. New York: Chapman and Hall.

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314 (software version 1.3.0, June 26, 2001).

Johnson RA, Wichern DW. 1998. Applied multivariate statistical analysis, 4th ed. Saddle River, NJ: Prentice-Hall.

Kidder JH, Jantz RL, Smith FH. 1992. Defining modern humans: a multivariate approach. In: Brauer G, Smith F, editors. Continuity or replacement: controversies in *Homo sapiens* evolution. Rotterdam: Balkema. p 157–178.

Kowalski CJ. 1972. A commentary on the use of multivariate statistical methods in anthropometric research. Am J Phys Anthropol 36:119–132.

Lim TS, Loh WY, Shih YS. 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learn J 40:203–228.

Mardia K, Kent JT, Bibby JM. 1979. Multivariate analysis. New York: Academic Press.

Mathsoft. 1999. S-PLUS 2000 guide to statistics: volume 2, data analysis. Seattle: Data Analysis Products Division.

McLachlan GJ. 1992. Discriminant analysis and statistical pattern recognition. New York: John Wiley and Sons.

Morrison DF. 1990. Multivariate statistical analysis. Third edition. New York: McGraw Hill.

Ripley BD. 1996. Pattern recognition and neural networks. Cambridge: Cambridge University Press.

Schaafsma W, van Vark GN. 1977. Classification and discrimination problems with applications, part I. Stat Neerl 31:25–46.

Schaafsma W, van Vark GN. 1979. Classification and discrimination problems with applications, part IIa. Stat NeeRl 33:91–126.

Schafer J. 1997. Analysis of incomplete multivariate data. New York: Chapman and Hall.

Schilling KJ. 1997. Multivariate analysis of the role of size as a source of sexual dimorphism and genus-specific variance in the femur and humerus of *Pan* and *Pan gorilla*. Unpublished M.A. thesis, Department of Anthropology, Portland State University.

Schimert J, Schafer JL, Hesterberg TM, Fraley C, Clarkson DB. 2000. Analyzing data with missing values in S-Plus. Seattle: Insightful Corp.

Steinberg D, Colla P. 1997. CART: tree-structured non-parametric data analysis. San Diego: Salford Systems.

Stringer CM. 1974. Population relationships of later Pleistocene hominids: a multivariate study of available crania. J Archaeol Sci 1:317–342.

Therneau TM, Atkinson EA. 1997. An introduction to recursive partitioning using the rpart routines. Technical report #61. Rochester, MN: Mayo Foundation. Available at http://www-.mayo.edu/hsr/techrpt/61.pdf (software version 3.1.1, dated August 9, 2001).

van Vark GN. 1976. A critical evaluation of the application of multivariate statistical methods to the study of human populations and their skeletal remains. Homo 27:94–117.

van Vark GN. 1995. The study of hominid skeletal remains by means of statistical methods. In: Boaz N, Wolfe L, editors. Biological anthropology: the state of the science. Corvallis, OR: IIHER Press. p 71–90.

Venables WN, Ripley BD. 1997. Modern applied statistics with S-PLUS. Second edition. New York: Springer.

Venables WN, Ripley BD. 1999. Modern applied statistics with S-PLUS. Third edition. New York: Springer.

Wilkinson L, Blank G, Gruber C. 1996. Desktop data analysis with SYSTAT. Saddle River, NJ: Prentice-Hall.

Appendix 1. Summarized output directly from rpart for *Iris* example. Nodes correspond to numbered entries on Figure 2. "Split" defines criterion for branching, "n" represents number of cases assigned to node, "loss" represents number of misclassified cases adjusted for prior information, "yval" is classification given to node if it were terminal, and "yprob" represents Bayesian probabilities (explained in text) of each group. * Signifies a terminal node.

```
node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 75 50.000000 setosa (0.33333333 0.33333333 0.33333333)
  2) petlength<2.5 28  0.000000 setosa (1.00000000 0.00000000 0.00000000) *
  3) petlength>=2.5 47 25.000000 virginica (0.00000000 0.50000000 0.50000000)
    6) petlength<4.85 21  0.000000 versicolor (0.00000000 1.00000000 0.00000000)*
    7) petlength>=4.85 26  1.136364 virginica (0.00000000 0.04347826 0.95652174)*
```

Appendix 2. Node details for all *Iris* splits. This expands information provided in summary. This example is unusual because either Petal Length or Petal Width produce identical (smallest) Gini diversity indices and identical improvement scores for the primary split. In this rare circumstance, the program chooses predictors alphabetically.

```
         CP nsplit  rel error     xerror       xstd
1 0.5000000      0 1.00000000 1.20467532 0.07127029
2 0.4772727      1 0.50000000 0.80922078 0.09055846
3 0.0100000      2 0.02272727 0.08545455 0.04166163

Node number 1: 75 observations,    complexity param=0.5
  predicted class=setosa      expected loss=0.6666667
    class counts:    28    22    25
   probabilities: 0.333 0.333 0.333
  left son=2 (28 obs) right son=3 (47 obs)
  Primary splits:
      petlength < 2.5  to the left,  improve=25.00000, (0 missing)
      petwidth  < 0.8  to the left,  improve=25.00000, (0 missing)
      seplength < 5.45 to the left,  improve=19.46643, (0 missing)
      sepwidth  < 3.35 to the right, improve=10.26949, (0 missing)
  Surrogate splits:
      petwidth  < 0.8  to the left,  agree=1.000, adj=1.000, (0 split)
      seplength < 5.45 to the left,  agree=0.947, adj=0.857, (0 split)
      sepwidth  < 3.35 to the right, agree=0.827, adj=0.536, (0 split)

Node number 2: 28 observations
  predicted class=setosa      expected loss=0
    class counts:    28     0     0
   probabilities: 1.000 0.000 0.000

Node number 3: 47 observations,    complexity param=0.4772727
  predicted class=virginica   expected loss=0.5319149
    class counts:     0    22    25
   probabilities: 0.000 0.500 0.500
  left son=6 (21 obs) right son=7 (26 obs)
  Primary splits:
      petlength < 4.85 to the left,  improve=22.826090, (0 missing)
      petwidth  < 1.75 to the left,  improve=22.826090, (0 missing)
      seplength < 6    to the left,  improve= 7.318278, (0 missing)
      sepwidth  < 3.05 to the left,  improve= 3.224007, (0 missing)
  Surrogate splits:
      petwidth  < 1.6  to the left,  agree=0.979, adj=0.952, (0 split)
      seplength < 6    to the left,  agree=0.787, adj=0.524, (0 split)
      sepwidth  < 2.75 to the left,  agree=0.681, adj=0.286, (0 split)

Node number 6: 21 observations
  predicted class=versicolor  expected loss=0
    class counts:     0    21     0
   probabilities: 0.000 1.000 0.000

Node number 7: 26 observations
  predicted class=virginica   expected loss=0.04370629
    class counts:     0     1    25
   probabilities: 0.000 0.043 0.957
```

Appendix 3. Statistical details associated with classification tree in Figure 6. For explanation, see Materials and Methods.

```
          CP nsplit rel error    xerror       xstd
1 0.29752275      0 1.0000000 1.1572598 0.06812347
2 0.17628205      2 0.4049545 0.6513108 0.07259214
3 0.03488372      3 0.2286724 0.3676488 0.05315698
4 0.03365385      4 0.1937887 0.3212863 0.05285487
5 0.01000000      6 0.1264810 0.2971112 0.05543828

Node number 1: 237 observations,    complexity param=0.2975228
  predicted class=GORILLA  expected loss=0.75
    class counts:    86    24   104    23
   probabilities: 0.250 0.250 0.250 0.250
  left son=2 (76 obs) right son=3 (161 obs)
  Primary splits:
      biepi    < 76.6   to the right, improve=46.39846, (0 missing)
      humlengt < 336.5  to the right, improve=46.18487, (0 missing)
      mlhttroc < 30.65  to the right, improve=41.48050, (0 missing)
      antartbr < 51.1   to the right, improve=39.38355, (0 missing)
      latsupri < 117.35 to the left,  improve=38.64198, (0 missing)
  Surrogate splits:
      mlhttroc < 31.925 to the right, agree=0.958, adj=0.868, (0 split)
      antartbr < 52.95  to the right, agree=0.949, adj=0.842, (0 split)
      humlengt < 367.5  to the right, agree=0.928, adj=0.776, (0 split)
      aphttroc < 18.65  to the right, agree=0.924, adj=0.763, (0 split)
      olecrdep < 16.25  to the right, agree=0.920, adj=0.750, (0 split)

Node number 2: 76 observations
  predicted class=GORILLA  expected loss=0.03389588
    class counts:    75     0     0     1
   probabilities: 0.953 0.000 0.000 0.047

Node number 3: 161 observations,    complexity param=0.2975228
  predicted class=HOMO     expected loss=0.7670957
    class counts:    11    24   104    22
   probabilities: 0.041 0.324 0.324 0.310
  left son=6 (139 obs) right son=7 (22 obs)
  Primary splits:
      latsupri < 117.05 to the left,  improve=49.41500, (0 missing)
      humlengt < 329.5  to the left,  improve=35.94438, (0 missing)
      pdhtcapi < 25.15  to the left,  improve=31.64261, (0 missing)
      mlhttroc < 30.3   to the left,  improve=21.27751, (0 missing)
      antartbr < 47.8   to the left,  improve=20.13675, (0 missing)
  Surrogate splits:
      pdhtcapi < 25.15  to the left,  agree=0.932, adj=0.500, (0 split)
      humlengt < 350.5  to the left,  agree=0.913, adj=0.364, (0 split)
      mlhtcapi < 20.15  to the left,  agree=0.888, adj=0.182, (0 split)
      mlhttroc < 32     to the left,  agree=0.876, adj=0.091, (0 split)
      antartbr < 54.17  to the left,  agree=0.876, adj=0.091, (0 split)

Node number 6: 139 observations,    complexity param=0.1762821
  predicted class=PAN      expected loss=0.4815527
    class counts:    11    23   104     1
   probabilities: 0.060 0.450 0.470 0.020
  left son=12 (69 obs) right son=13 (70 obs)
  Primary splits:
      humlengt < 300.5  to the right, improve=19.491080, (0 missing)
```

(continued)

Appendix 3. (continued)

```
      olecrdep < 11.95  to the left,  improve= 7.797573, (0 missing)
      mlhttroc < 30.75  to the right, improve= 7.753742, (0 missing)
      biepi    < 70.85  to the right, improve= 5.718158, (0 missing)
      antartbr < 51.05  to the right, improve= 5.263866, (0 missing)
  Surrogate splits:
      pdhtcapi < 21.68  to the right, agree=0.683, adj=0.362, (0 split)
      antartbr < 44.65  to the right, agree=0.676, adj=0.348, (0 split)
      biepi    < 58.65  to the right, agree=0.669, adj=0.333, (0 split)
      aphttroc < 15.65  to the right, agree=0.662, adj=0.319, (0 split)
      mlhttroc < 27.35  to the right, agree=0.662, adj=0.319, (0 split)

Node number 7: 22 observations
  predicted class=PONGO    expected loss=0.1122159
    class counts:     0     1     0    21
   probabilities: 0.000 0.044 0.000 0.956

Node number 12: 69 observations,    complexity param=0.03488372
  predicted class=HOMO     expected loss=0.4444086
    class counts:    11    21    36     1
   probabilities: 0.092 0.628 0.249 0.031
  left son=24 (9 obs) right son=25 (60 obs)
  Primary splits:
      mlhttroc < 30.75  to the right, improve=8.597229, (0 missing)
      olecrdep < 13.05  to the left,  improve=7.565060, (0 missing)
      humlengt < 337.5  to the right, improve=7.413283, (0 missing)
      antartbr < 42.9   to the left,  improve=6.344064, (0 missing)
      biepi    < 69.745 to the right, improve=5.939034, (0 missing)
  Surrogate splits:
      humlengt < 341    to the right, agree=0.971, adj=0.778, (0 split)
      biepi    < 71.05  to the right, agree=0.971, adj=0.778, (0 split)
      antartbr < 51.05  to the right, agree=0.957, adj=0.667, (0 split)
      latsupri < 72.11  to the left,  agree=0.899, adj=0.222, (0 split)
      medepico < 15.5   to the right, agree=0.899, adj=0.222, (0 split)

Node number 13: 70 observations
  predicted class=PAN      expected loss=0.07053571
    class counts:     0     2    68     0
   probabilities: 0.000 0.113 0.887 0.000

Node number 24: 9 observations
  predicted class=GORILLA  expected loss=0
    class counts:     9     0     0     0
   probabilities: 1.000 0.000 0.000 0.000

Node number 25: 60 observations,    complexity param=0.03365385
  predicted class=HOMO     expected loss=0.4077268
    class counts:     2    21    36     1
   probabilities: 0.018 0.679 0.269 0.034
  left son=50 (25 obs) right son=51 (35 obs)
  Primary splits:
      olecrdep < 13.05  to the left,  improve=6.972128, (0 missing)
      antartbr < 42.9   to the left,  improve=5.279082, (0 missing)
      humlengt < 321    to the right, improve=5.000037, (0 missing)
      mlhtcapi < 17.05  to the left,  improve=3.864870, (0 missing)
      biepi    < 60.45  to the left,  improve=3.665587, (0 missing)
  Surrogate splits:
      antartbr < 42.25  to the left,  agree=0.750, adj=0.40, (0 split)
```

(continued)

Appendix 3. (continued)

```
      mlhttroc < 25.5   to the left,  agree=0.733, adj=0.36, (0 split)
      biepi    < 60.85  to the left,  agree=0.733, adj=0.36, (0 split)
      mlhtcapi < 16.65  to the left,  agree=0.700, adj=0.28, (0 split)
      latsupri < 79.225 to the left,  agree=0.650, adj=0.16, (0 split)

Node number 50: 25 observations
  predicted class=HOMO     expected loss=0.2853512
    class counts:     0    16     8     1
   probabilities: 0.000 0.847 0.098 0.055

Node number 51: 35 observations,    complexity param=0.03365385
  predicted class=PAN      expected loss=0.3920473
    class counts:     2     5    28     0
   probabilities: 0.046 0.416 0.538 0.000
  left son=102 (14 obs) right son=103 (21 obs)
  Primary splits:
      humlengt < 317    to the right, improve=7.798651, (0 missing)
      biepi    < 64.04  to the right, improve=5.593440, (0 missing)
      medepico < 11.95  to the right, improve=4.793755, (0 missing)
      latsupri < 93.85  to the right, improve=4.705908, (0 missing)
      mlhtcapi < 18.53  to the right, improve=3.965978, (0 missing)
  Surrogate splits:
      medepico < 11.23  to the right, agree=0.771, adj=0.429, (0 split)
      biepi    < 64.04  to the right, agree=0.771, adj=0.429, (0 split)
      latsupri < 103.5  to the right, agree=0.743, adj=0.357, (0 split)
      pdhtcapi < 23.5   to the right, agree=0.743, adj=0.357, (0 split)
      mlhtcapi < 18.53  to the right, agree=0.743, adj=0.357, (0 split)

Node number 102: 14 observations
  predicted class=HOMO     expected loss=0.3832777
    class counts:     2     5     7     0
   probabilities: 0.078 0.697 0.225 0.000

Node number 103: 21 observations
  predicted class=PAN      expected loss=0
    class counts:     0     0    21     0
   probabilities: 0.000 0.000 1.000 0.000
```