FURTHER READING

Balon, E. K. 1995. Origin and domestication of the wild carp, *Cyprinus carpio*: From Roman gourmets to the swimming flowers. *Aquaculture* 129: 3–48.

Bajer, P. B., and P. W. Sorensen. 2009. The superabundance of common carp in interconnected lakes in Midwestern North America can be attributed to the propensity of adults to reproduce in outlying habitats that experience winter hypoxia. *Biological Invasions* 12: 1101–1112.

Billard, R., ed. 1995. *Carp Biology and Culture.* Paris: Springer-Verlag.

Buffler, J. R., and T. J. Dickson. 1990. *Fishing for Buffalo.* Minneapolis: Culpepper Press.

Cooper, E. L. 1987. *Carp in North America*. Bethesda, MD: American Fisheries Society.

Haas, K., U. Köhler, S. Diehl, P. Köhler, S. Dietrich, S. Holler, A. Jaensch, M. Niedermaier, and J. Vilsmeier. 2007. Influence of fish on habitat choice of water birds: A whole system experiment. *Ecology* 88: 2915–2925.

Koehn, J. D. 2004. Carp (*Cyprinus carpio*) as a powerful invader in Australian waters. *Freshwater Biology* 49: 882–894.

Koehn, J., A. Brumley, and P. Gehrke. 2000. *Managing the Impacts of Carp*. Canberra, Australia: Bureau of Rural Sciences.

McCrimmon, J. R. 1968. *Carp in Canada.* Bulletin 165. Ottawa: Fisheries Research Board of Canada.

Sorensen, P. W., and N. E. Stacey. 2004. Brief review of fish pheromones and discussion of their possible uses in the control of non-indigenous teleost fishes. *New Zealand Journal of Marine and Freshwater Research* 38: 399–417.

CART AND RELATED METHODS

VOJTĚCH JAROŠÍK

Charles University, Prague, Czech Republic

Classification and regression trees (CART) are a computer-intensive data-mining tool originally designed for analyzing vast databases of often incomplete data, with an aim to find financial frauds, suitable candidates for loans, potential customers, and other uncertain outputs. Searching for potential invasive species and their traits responsible for invasiveness, predicting their potential distributions in regions where they are not native, or identifying factors that distinguish invasible communities from those that resist invasion are similar risk assessments. This is perhaps one reason why CART and related methods are becoming increasingly popular in the field of invasion biology. Identifying homogeneous groups with high or low risk and constructing rules for making predictions about individual cases is, in essence, the same for financial credit scoring as for pest risk assessment. In both cases, one searches for rules that can be used to predict uncertain future events.

PRINCIPLES OF CART

CART models use decision trees to display how data may be classified. Their method is technically known as *binary recursive partitioning*: the data are successively split along coordinate axes of the explanatory variables so that, at any node, the split is selected that maximally distinguishes the response variable in the left and the right branches. If the response variable is categorical, the tree is called a *classification tree*; if continuous, it is called a *regression tree*. Explanatory variables can be both categorical and continuous. The process is binary because parent nodes are always split into exactly two child nodes by asking questions that have a "yes" or "no" answer, and it is recursive because the process can be repeated by treating each child node as a parent.

Making a decision when a tree is complete is best achieved by growing the tree until it is impossible to grow it further and then examining smaller trees obtained by gradually decreasing the size of the maximal tree in a process called pruning. A single *optimal tree* is then determined by testing for misclassification error rates of candidate trees. When the data are sufficiently numerous (i.e., greater than 3,000 records), they are divided into a learning (also called training) sample and a test sample, created by a completely independent set of data or a random subset (e.g., 20%) of the input data. To calculate the misclassification error rate, the model is fitted to the learning sample to predict values in the test sample. The error rate is calculated for the largest tree as well as for every smaller tree. When the data are not sufficiently numerous to allow for separate test samples, cross-validation is employed. Cross-validation involves splitting the data into a number (e.g., 10) of smaller samples with similar distributions of the response variable. Trees are then generated, excluding the data from each subsample in turn. For each tree, the error rate is estimated from the subsample excluded in generating it, and the cross-validated error for the overall tree is then calculated. The cross-validated or test sample errors are then plotted against tree sizes (insets in Fig. 1). The optimal tree is the one with the lowest error rate (SE rule 0 in the inset) or that is within one standard error of the minimum (SE rule 1 in the inset). A series of 50 or 100 validations of error rates for both rules is recommended, of which the modal (most likely) optimal tree is chosen for description. This tree is then represented graphically (Fig. 1), with the root

From Daniel Simberloff and Marcel Rejmánek, editors, *Encyclopedia of Biological Invasions*, Berkeley and Los Angeles: University of California Press, 2011.

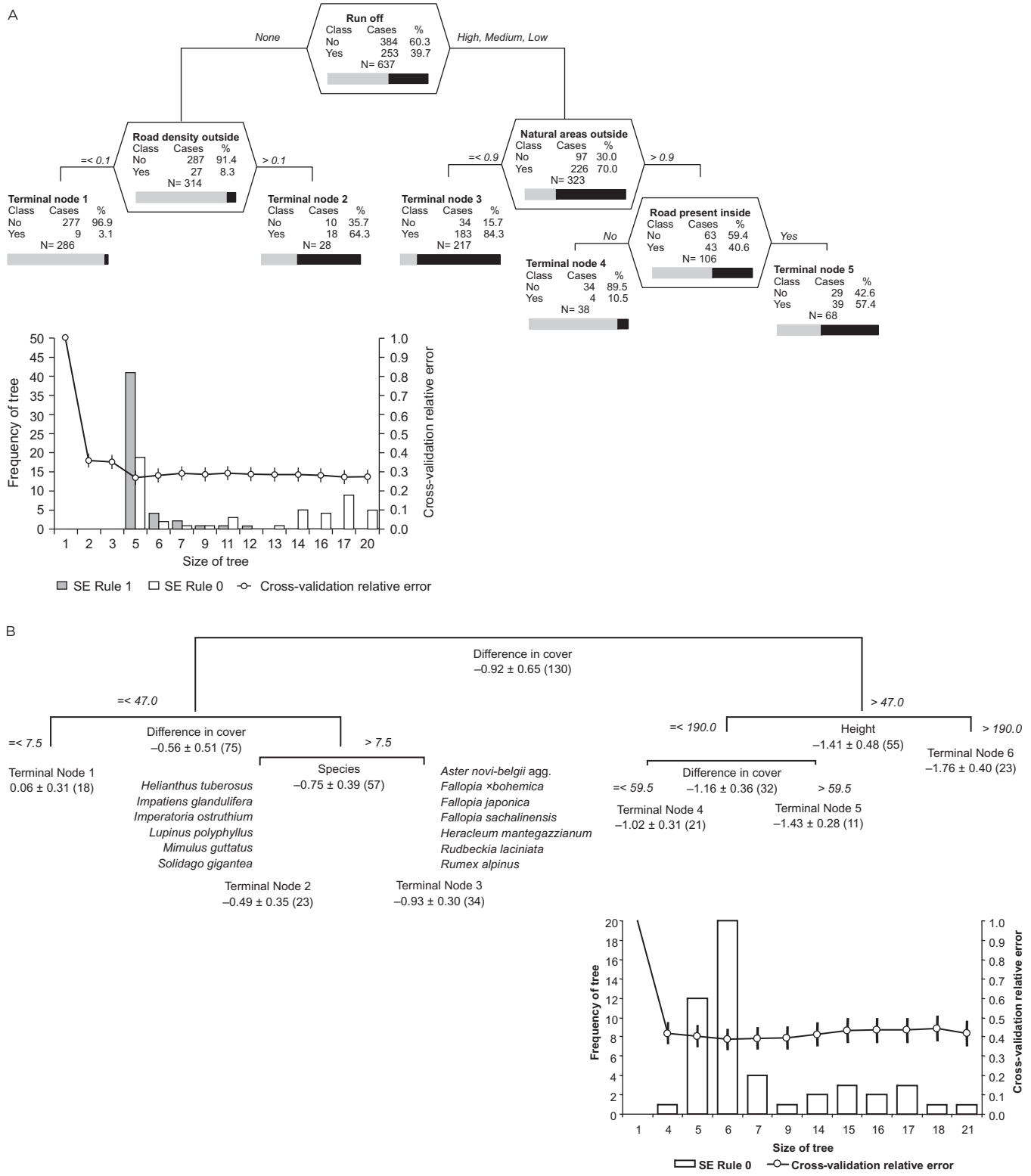

FIGURE 1 (A) Classification tree describing the probability of an alien plant presence (No/Yes) in boundary segments of Kruger National Park (KNP) based on explanatory variables from within the KNP and areas adjacent to the park (after Foxcroft et al., Protected Area Boundaries as a Natural Filter of Plant Invasions from Surrounding Landscapes, under review). (B) Regression tree describing the impact of individual invading plant species on diversity of native plant communities (Shannon's index of diversity *H'*) based on absolute and relative population performances of the invaders. (After Hejda et al. 2009. *Journal of Ecology* 97: 393–403.) For both trees, the vertical depth of each node is proportional to its improvement value that corresponds to explained variance at the node. *(Continued on next page)*

standing for undivided data at the top, and the terminal nodes, describing the most homogeneous groups of data, at the bottom of the hierarchy.

The quality of each split can be expressed based on deviance explained by the split and visualized as a vertical depth of the split. The overall quality of the best classification tree (Fig. 1A) can be expressed as its misclassification rate by comparing the misclassification rate of the optimal tree with the misclassification rate of the null model (e.g., 50% null misclassification rate for a presence–absence response variable) and with the misclassification rate for each category of response variable. The overall quality can be also described as specificity and sensitivity of cross-validated or test samples. Specificity is defined as the true positive rate, or the proportion of observations correctly identified (for instance, the ability of the model to predict that a species is not invasive when it is not). Sensitivity is defined as the true negative rate (e.g., the ability to predict that a species is invasive when it is). The overall quality and appropriateness of the optimal regression tree (Fig. 1B) can be expressed similarly to analogous features of a linear model. We can express explained variance r^2 of the tree, and because we know observed and predicted values for each terminal node, we can also calculate residuals as a difference between observed and expected values and use them as a diagnostic check of the model.

RELATED METHODS AND THEIR COMBINATIONS WITH CART

When explanatory variables have no missing values and either have or can be transformed to an approximately normal distribution, CART models can be replaced by linear or multivariate statistics. Linear models can then substitute for regression trees. If we have a categorical response variable with two classes and at least

FIGURE 1 *(Continued)* Insets: Cross-validation processes for the selection of the optimal trees. The lines show a single representative tenfold cross-validation of the most frequent (modal) optimal tree with standard error (SE) estimate for each tree size. Bar charts are the numbers of optimal trees of each size (frequency of tree) selected from a series of 50 cross-validations based on the minimum-cost tree, which minimizes the cross-validated error (white, SE rule 0), and 50 cross-validations based on SE rule 1 (gray, SE rule 1), which minimizes the cross-validated error within one standard error of the minimum. The most frequent (modal) classification tree, established based on both SE rules, has five terminal nodes (A), and the modal regression tree, established based on SE rule 0, has six terminal nodes (B).

(A) Each node (polygonal table with splitting variable name) and terminal node with node number shows table with columns for class (No/Yes), number of cases and percent of cases for each class, total number of cases (*N*), and graphical representation of the percentage of No (gray) and Yes (black) cases in each class (horizontal bar). Except for the root node standing for undivided data at the top, there are the splitting variable name and split criterion above each node. "Run-off" is the categorical measure of mean annual runoff from the surrounding watershed (million m^3), "natural areas outside" refers to the percentage of natural areas within a 5 km radius outside the KNP boundary, "road density outside" refers to the density of major roads within a 10 km radius outside the KNP boundary, and "road present inside" refers to the presence or absence of all roads in the segment inside the KNP. In segments with no river, the probability of being invaded depended on the density of major roads within a 10 km radius outside the KNP boundary. If there was a river, invasion was unlikely only in segments with more than 90 percent of protected natural areas with natural vegetation in the adjacent 5 km radius outside the KNP, and with roads absent within the park. The overall misclassification rate of the model is 13.5 percent, compared to 50 percent for the null model; the sensitivity (true positive rate, defined as proportion of observations correctly identified as suitable) is 0.90; the specificity (true negative rate) is 0.80; the misclassification rate for the presence of an alien species is 0.10; the misclassification for the absence is 0.20. This model is an alternative tree after dropping a primary splitter, the continuous explanatory variable "mean annual runoff," from the optimal tree. The categorical surrogate "run-off" appeared at Node 1 of the optimal tree with an association value of 0.86, and it explained 86.8 percent of the variability of the primary split. The optimal tree had a bit higher misclassification rate (14%), but higher sensitivity (0.92) and specificity (0.81), and a lower misclassification rate for the presence (0.075) and absence (0.19) of alien species, with only three compared to five terminal nodes for the alternative tree. The chosen continuous explanatory variables of the optimal tree, "mean annual runoff" and "road density outside," lacked collinearity and could be used as explanatory variables in a logistic regression.

(B) Each node shows the splitting variable, splitting criteria, mean ± standard deviation of the difference in species diversity between invaded and uninvaded pairs of plots (negative value indicates a decrease due to invasion), and number of plots in brackets. "Difference in cover" is a cover difference between an invading species and the dominant native species in uninvaded plots (in percent), "height" is height of the invading species (in centimeters), and "species" are the scientific names of the invading plants. To reduce the splitting power of the high categorical variable "species" (13 factor levels), the species were adjusted to have no inherent advantage over continuous explanatory variables. The impact was first divided based on the cover difference of approximately 50 percent (≤47%). The group with the small differences in cover exhibited no impact on diversity if the cover of the invading and native dominant species differed by less than or equal to 7.5 percent; for cover differences between 8 and 47 percent, the impacts were species-specific. The group with differences in cover above 47 percent indicated the absolutely highest impact on diversity if the invading species was taller than 190 cm. If the invading species was shorter than 190 cm, then the impact on diversity was further divided based on differences in cover. The tree explains 74 percent of the variance. The alternative linear model explained 76 percent of the variance, but all the variance was included in interactions in a way that rendered the model noninterpretable.

one explanatory variable is continuous, then a suitable method to replace a classification tree is a binary logistic regression; if all explanatory variables are continuous, then a suitable alternative is also multivariate discriminant function analysis. For missing values, CART and a related technique called random forests (RF) can be applied. RF gives, at least for small samples, more robust results than CART and allows for a ranking of explanatory variables, but it does not have the CART virtue of easily followed graphic presentation. The ability to treat missing values, however, makes both RF and CART invaluable tools when one tries to identify and rank traits associated with invasiveness, impact of invasive species, and similar tasks in which we usually deal with very incomplete data.

When classification trees are compared with logistic regressions or discriminant analyses, the optimal tree often performs better on the learning sample, and because of the CART model verifications on validated samples, the tree is usually more accurate on new data. The same is usually true for RF, owing to its self-testing procedure based on an extension of cross-validation. When we compare linear models and regression trees on data determining susceptibility to invasions of different habitats by plant invaders, we also find slightly higher explanatory power for regression trees, and the trees are much easier to interpret than linear models. The reason is that, unlike linear models, which uncover a single dominant structure in the data, CART models are designed to work with data that might have multiple structures. The models can use the same explanatory variable in different parts of the tree, dealing effectively with nonlinear relationships and higher-order interactions. In fact, provided there are enough observations, the more complex the data and the more variables that are available, the better tree models will do compared to alternative methods. With a complex data set, understandable and generally interpretable results often can be found only by constructing trees (Fig. 1B).

However, trees are also excellent for initial data inspection. CART models are often used to select a manageable number of core measures from databases with hundreds of variables. A useful subset of predictors from a large set of variables can then be used in building a formal linear model (Fig. 1A). CART can also suggest further model simplification by converting a continuous variable to a categorical one at cut-points used to generate splits and by lumping together subcategories of categorical variables that are never distinguished during splitting.

IMPORTANT PROPERTIES OF CART

CART models are nonparametric. Consequently, unlike with parametric linear models, nonnormal distribution and collinearity do not prevent reliable parameter estimates. Because the trees are invariant to monotonic transformations of continuous explanatory variables, no transformation is needed prior to analyses. Outliers among the response variables generally do not affect the models because splits usually occur at non-outlier values. However, in some circumstances, transformation of the response variable may be important to alleviate variance heterogeneity.

Surrogates of each split, describing splitting rules that closely mimic the action of the primary split, can be assessed and ranked according to their association values, with the highest possible value 1.0 corresponding to the surrogate producing exactly the same split as the primary split. Surrogates can then be used to replace an expensive primary explanatory variable by a less expensive, although probably less accurate, one, and then to build alternative trees on surrogates (Fig. 1A). Unlike in a linear model, a variable in CART thus can be considered highly important even if it never appears as a primary splitter. Surrogates also serve to treat missing values, because the alternative splits are used to classify a case when its primary splitting variable is missing.

However, as it is easier to be a good splitter on a small number of records (e.g., splitting a node with just two records), to prevent missing explanatory variables from having an advantage as splitters, the power of explanatory variables can be penalized in proportion to the degree to which they are missing. High-level categorical explanatory variables have inherently higher splitting power than continuous explanatory variables and therefore can also be penalized to level the playing field (Fig. 1B). Finally, proportions calculated from larger samples give more precise estimates, and therefore, proportional response variables can be weighted by their sample sizes; a similar approach can be applied to stratified sampling on strata having different sampling intensities. CART models can also accommodate situations in which some misclassifications are more serious than others. For instance, if invasion risks are classified as low, moderate, and high, it would be more costly to classify a high-risk species as low-risk

than as moderate-risk. This can be achieved by specifying a differential penalty for misclassifying high, moderate, and low risk.

LIMITATIONS

CART models are good, but not as good to solve completely all problems with data that violate a basic assumption of the independence of errors of observations due to temporal or spatial autocorrelation, or due to a related problem of phylogenetic relatedness. Fortunately, as we do not need formal parametric tests of statistical significance, spatial autocorrelations cannot prevent correct use of CART and related nonparametric methods (e.g., for prediction of species distributions). However, CART cannot distinguish fixed and random effects and thus cannot be used with mixed effect and nested statistical designs. Trees thus do not allow for phylogenetic corrections using mixed-effect general linear models in which taxonomic hierarchy is included as nested random effects. They are also inefficient when used on principal coordinate axes derived from phylogenetic trees to account for relatedness. The reason is that the principal coordinate axes are orthogonal, and trees exhibit their greatest strengths with a highly nonlinear structure and complex interactions. Their usefulness decreases with increasing linearity of the relationships, and consequently, on mutually independent principle coordinates, no trees are built. Because, in invasion biology, we usually need values for each individual species, phylogenetic contrasts derived from related species are also usually not helpful. The only way to include phylogeny in CART models seems to be to use the hierarchical taxonomic affiliations of the individual species. Considering phylogenetic effects is important not only to treat a lack of statistical independence, but also to solve practical implications (e.g., when one predicts whether a species belonging to a particular family, order, or class would be more predisposed to invasion than other species belonging to other taxa at the same hierarchical level).

The tree-growing method is data intensive, requiring many more cases than classical regression. While for multiple regression it is usually recommended to keep the number of explanatory variables six to ten times smaller than the number of observations, for classification trees, at least 200 cases are recommended for binary response variables and about 100 more cases for each additional level of a categorical variable. Efficiency of trees decreases rapidly with decreasing sample size, and for small data sets, no test samples may be available.

SEE ALSO THE FOLLOWING ARTICLES

Invasiveness / Life History Strategies / Remote Sensing / Risk Assessment and Prioritization

FURTHER READING

Bourg, N. A., W. J. McShea, and D. E. Gill. 2005. Putting a CART before the search: Successful habitat prediction for a rare forest herb. *Ecology* 86: 2793–2804.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. G. Stone. 1984. *Classification and Regression Trees*. Pacific Grove: Wadsworth.

Cutler, D. R., T. C. Edwards Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. *Ecology* 88: 2783–2792.

De'ath, G., and E. Fabricius. 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology* 81: 3178–3192.

Křivánek M., P. Pyšek, and V. Jarošík. 2006. Planting history and propagule pressure as predictors of invasions by woody species in a temperature region. *Conservation Biology* 20: 1487–1498.

Kumar, S., S. A. Spaulding, T. J. Stohlgren, K. A. Hermann, T. S. Schmidt, and L. L. Bahls. 2009. Potential habitat distribution for the freshwater diatom *Didymosphenia geminata* in the continental US. *Frontiers in Ecology* 7: 415–420.

Reichard, S. H., and C. W. Hamilton. 1997. Predicting invasions of woody plants introduced into North America. *Conservation Biology* 11: 193–203.

Rejmánek, M., and D. M. Richardson. 1996. What attributes make some plant species more invasive? *Ecology* 77: 1655–1661.

Steinberg, G., and P. Colla. 1995. CART: *Tree-Structured Non-Parametric Data Analysis*. San Diego, CA: Salford Systems.

Vall-llosera, M., and D. Sol. 2009. A global risk assessment for the success of bird introductions. *Journal of Applied Ecology* 46: 787–795.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*, 4th ed. New York: Springer.

CHEATGRASS

RICHARD N. MACK

Washington State University, Pullman

Cheatgrass (*Bromus tectorum*), or downy brome, is a cleistogamous (i.e., almost totally self-pollinating) annual grass that occupies an enormous native range in Eurasia and the northern rim of Africa. In the past 200 years it has been transported worldwide, almost always as an accidental introduction, and it now has a naturalized range that includes North and South America, Australia, and temperate environments in Oceania. It has become a widespread invader in arid North America, especially in the largely treeless region between the Rocky Mountains and the Cascade and