Regression Tree Cartography

Denis WHITE and Jean C. SIFNEOS

We illustrate several types of cartographic displays that can enhance understanding from hierarchical analysis techniques such as regression trees. When the observations have spatial locations, maps of the predicted values, maps of the residuals, and maps of the predicting relationships of the tree may help to reveal associations between predictors and response. We propose an objective method for constructing maps that may help to show the geographical similarities and differences between observations based on their positions in the prediction tree. This mapping method divides the color spectrum to assign colors to the leaves, using the same hierarchical pattern as the prediction tree does to divide the data. We illustrate regression tree cartography with two examples and suggest how the prediction tree mapping method could be used for classification trees and for dendrograms produced by hierarchical clustering methods.

Key Words: Color symbolism; Classification and regression trees; Hierarchical clustering.

1. INTRODUCTION

Classification and regression trees (Breiman, Friedman, Olshen, and Stone 1984) are useful for looking at hierarchical models of relationships between predictors and response. In these models, different subsets of the observations can have different predicting conditions. The form of the relationship at each node in the hierarchical tree, in the Breiman et al. (1984) formulation, is a univariate function between one of the predictors and the response. The relationship is either a threshold function of a continuous variable or a level in a categorical variable. Classification trees have a categorical variable as a response and regression trees have a continuous variable.

In applications where the observations are spatial locations, visualizing the geography of the prediction tree may reveal insights into mechanistic relationships or other kinds of associations between the predictors and the response. Mapping residuals from the prediction tree may also help to identify missing variables or gaps in knowledge. Previous work in mapping classification and regression trees includes Davis, Michaelsen, Dubayah, and

Denis White is a Geographer, U.S. Environmental Protection Agency, Corvallis, OR 97333 (E-mail: white.denis@epa.gov). Jean C. Sifneos is a Statistician, Department of Geosciences, Oregon State University, Corvallis, OR 97331 (E-mail: sifneos.jeannie@epa.gov).

In the Public Domain

Journal of Computational and Graphical Statistics, Volume 11, Number 3, Pages 600–614
DOI: 10.1198/106186002484

Dozier (1990), Walker (1990), Moore, Lees, and Davey (1991), Skidmore, Gauld, and Walker (1996), O'Connor et al. (1996), Iverson and Prasad (1998), and Rathert, White, Sifneos, and Hughes (1999). We explore the cartography of regression trees by proposing an objective method for assigning map symbols to the leaves of regression trees in order to display the different prediction sequences. We contrast this type of cartography with the more usual mapping of predicted responses and their residuals, and we illustrate these different cartographic methods with two examples. We also discuss how maps of prediction relationships can aid in determining the size of the prediction tree. The presentation will focus on regression trees although similar techniques could be used with classification trees.

2. METHODS

2.1 CARTOGRAPHY OF REGRESSION TREES

In building a regression tree, the midpoints between all values of all of the predicting variables that are present in the data form the possible splits for the tree. In the first step, sums of squares of differences between the observations and their means are computed for all binary divisions of the observations formed by all of the splits. The minimum sum determines the split. The observations are then divided into two subsets based on the split and the process recursively repeats on the two descendent subsets. One kind of possible split is between observations with valid data values and those with missing values. Splitting continues until one or more stopping criteria are reached. The criteria are defined to balance predictive power and parsimony. We used the cross-validation pruning techniques of Breiman et al. (1984), as implemented by Clark and Pregibon (1992), and with recommendations by Sifneos, White, and Urquhart (in press), to determine the optimal size of trees.

One way to map the predicting relationships when the observations represent spatial locations is to assign a cartographic symbol, such as a color or gray tone, to each leaf or terminal node of a tree so that the map shows the pattern of the observations in the different leaves. We would like to satisfy two goals with the cartographic symbolism: (1) to distinguish qualitatively between different paths of prediction from the root of the tree to a leaf, and (2) to indicate, also, which prediction paths are similar because they descend from the same upper level nodes.

Mapping the predicting relationships in this way is different from mapping the predicted values of the response variable, as we will illustrate in the examples. In the Breiman et al. (1984) formulation of regression trees, the predicted value of the response at each leaf is the mean of the observations contained in the leaf. A map of the predicted response is then simply a univariate map of the levels of the response predicted at each leaf. Cartographic design principles (Ware 1988; Brewer 1994) suggest using a sequence of gray tones from light to dark, or a set of colors from gray to full saturation in a single hue, to represent the predicted response levels. Mapping the predicting relationships, on the other hand, is an attempt to reveal the geographical similarities and differences in observations according to the sequences of explanatory variables that define the positions of the observations in the tree.

Our first approach to developing a cartographic symbolism for mapping regression tree relationships was to select equally spaced points on the gradient from black to white in order to create a gray tone sequence that could be applied to the leafs in positional order from left to right, for example (White and Sifneos 1997). This symbolism technique is limited by the number of discrete gray tones that can be perceived. To obtain more resolution we developed a recursive partition of the hue spectrum to mimic the recursive partitioning of the observation space by a regression tree. We started with the total range of hues that were available and chose the hue at the midpoint of the range to represent the root or top node (thinking of the tree as inverted so that the root is at top). The two descendants of the root were assigned hues that were the midpoints of the two halves of the original range of hues. Division of hues continued to the maximum depth of the hue tree. Because the hue tree is a binary tree with two descendants at each node, like a regression tree, we can use as colors for each leaf of a regression tree the hue in the corresponding position of the hue tree.

A spectral sequence in hues assigned to the leaves of regression trees meets the two goals of a cartographic symbolism for regression trees: (1) different prediction conditions (sequences of splits leading to the leaves) have qualitatively different hues, and (2) the hues have an ordering so that the relationship between the color symbols and the structure of the tree may assist in map interpretation. If the splits have a consistent ordering such that, for example, observations with values less than the splitting value are always placed in the left node of the split, then the assignments of colors will be unambiguous. Of course, if the ordering is the opposite such that observations with values greater than the splitting value are placed on the left, then the color assignments will be different, and the apparent nearness of leaves descending from nodes higher in the tree will be different. Therefore, observations that are close in color space will not always be close in prediction space. The relationship between color space and prediction space is that two leaves that are close in the prediction tree, that is, with a short path through the tree connecting them, will have similar colors. But similar colors need not be associated with a short path. In reading maps of prediction space, then, it is important to use, as a companion graphic, a diagram of the prediction tree showing the colors assigned to the leaves.

2.2 COLOR CALIBRATION

To implement the recursive partitioning of the color spectrum, we needed a method to obtain approximately equal perceptual steps along the spectrum. Color science suggests using one of the uniform color spaces developed from empirical color-matching data (Wyszecki and Stiles 1982; Tajima 1983; Hill, Roger, and Vorhagen 1997). In the CIELAB color space (a device independent international standard for color measurement), for example, dividing the hue dimension into equal steps and translating these coordinates into one of the color models used in computer graphics such as RGB (red, green, and blue as primary variables) or HLS (hue, lightness, and saturation) could provide a satisfactory

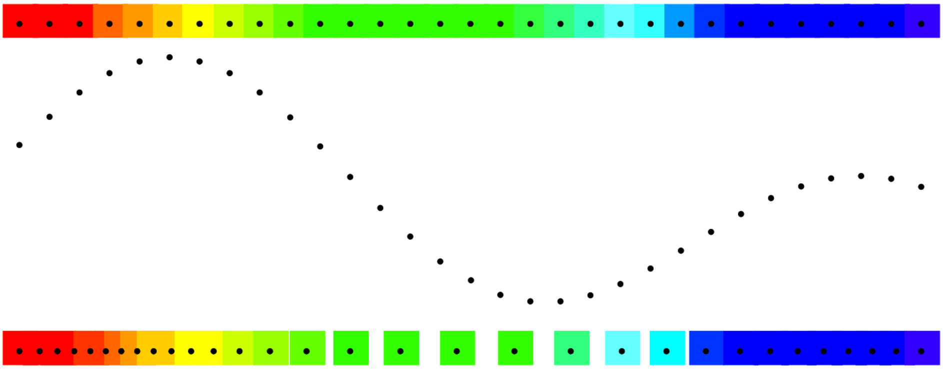

Figure 1. The unmodified hue spectrum above the curve was resampled by the weighting function represented by the curve to achieve the more equally spaced spectrum below. The points along the curve of the weighting function define samples at equal intervals along the unmodified spectrum. The weighting resulted in an unequal spacing of those sample points as indicated in the transformed spectrum below.

spectrum (Robertson and O'Callaghan 1986; also see http://www.inforamp.net/~poynton/ColorFAQ.html). Without a color management system, however, the properties of specific color graphics output devices, their inks or toners, and their media cause differences in perception of identically specified colors from device to device. Because we did not have a color management system to overcome these problems, we used an adaptive weighting method to obtain an equally spaced spectrum. Our approach was purely empirical; we modified the color spectrum until we judged by our eyes that the spacing of hues was approximately equal.

The weighting function we used was the product of an exponential and a sine function with several parameters: $\exp(px) \bullet \sin(4psd(x + f))$, where x was a value in the interval $[0, 1]$ representing the hue in the HSB (hue, saturation, and brightness) color model, and p, s, d, and f were parameters for tuning the function to obtain a desired spectrum. The procedure was to plot an unmodified hue spectrum and then adjust the parameters of the weighting function to transform the hues into a more perceptually equal sequence (Figure 1). For our output, the parameters were set to $p = -0.75$, $s = 0.85$, $d = 1.0$, and $f = -0.01$. Because hue is represented in many color models as a circular function starting at red, passing through other hues, and returning to red, and because we did not want duplicate colors at each end, we truncated the hue sequence to 80% of its length, ending in purple rather than red. After respacing the hue spectrum, we produced a color tree from a recursive partition of the spectrum (Figure 2). This color tree formed the basis for assigning colors to leaves in the corresponding positions of regression trees.

3. EXAMPLES

3.1 NATIVE BIRD SPECIES RICHNESS IN OREGON

In the first example the data were from a biodiversity research program made up of investigators in a number of organizations (White, Preston, Freemark, and Kiester 1999).

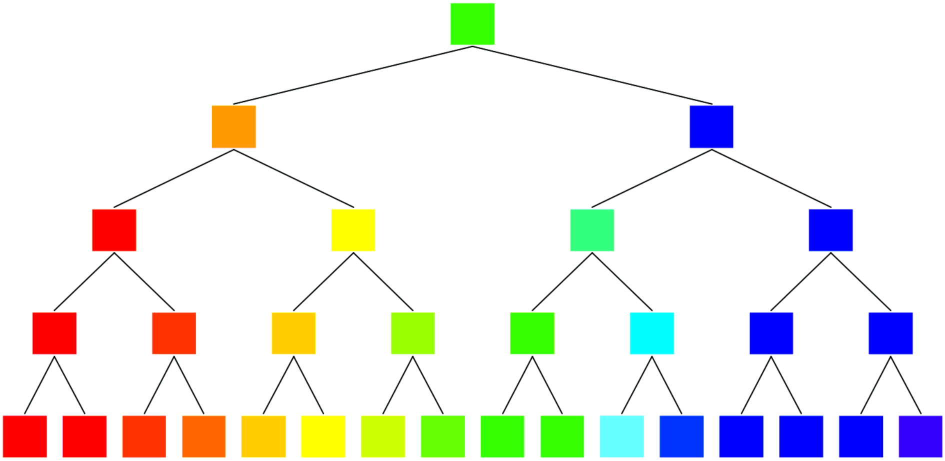

Figure 2. The color tree obtained by recursively partitioning the transformed hue spectrum. The range of hues is bisected at each level in the tree and the hue at each node is the midpoint in the range assigned to the node.

The data were collected on a sampling grid of 391 cells, each of approximately 650 km^2 in area, across the state of Oregon, USA. For a response variable we used the number of native breeding bird species per cell. As predictors we used 12 functions of climate, remotely sensed radiation, and vegetation and land cover classes.

The geography of the bird prediction tree confirmed major environmental regions in Oregon (Figure 3). Warmer winter areas (minimum January temperatures greater than −6.5 Celsius) were identified by the first tree split to be primarily in western Oregon and at lower elevations in eastern Oregon. The first split accounted for 25.1% of the deviance (analogous to variance explained) in the data and predicted areas with colder winters to have 21 more species, on average. This result was somewhat counterintuitive since, in general, species richness for most vertebrate species increases from cooler regions to warmer regions. In this case the result may have reflected the more heterogeneous habitats of forested areas in eastern Oregon than in western Oregon and the western and northern extensions of ranges of Rocky Mountain and California species, respectively.

In the warmer winter areas, the next split, on the right side of the tree, divided most of the coastal mountains, western valleys, and the Columbia plateau (in purple) from scattered areas in southwestern, central, and southeastern Oregon (in blue). This split was on the annual range in normalized difference vegetation index (NDVI), a variable that was derived from two bands of remotely sensed radiation response to be a surrogate for the biomass or productivity of vegetation. In the cooler winter areas, the next split was on the annual total of NDVI. This split divided areas of more forest cover in the Cascade and Blue Mountains (yellow and green) from more arid areas primarily in southeastern Oregon (red and orange). Higher species richness was associated with higher NDVI. The two subsequent splits below this level both used the annual temperature range. In the more forested areas a greater temperature range was associated with more species, but the opposite was the case for less forested areas. The nonlinear relationship between annual temperature range and species richness is similar to that between other environmental variables and species richness.

The map of the predicted values of species richness (Figure 4a) had only six levels, corresponding to the means at the six leaves. The two leaves with similar means (115 and 116 species) were in different parts of the tree (Figure 3) and at opposite ends of the state (Figure 4a). In the map of the residuals from the prediction (Figure 4b), there was some clustering of areas of over- and under-prediction.

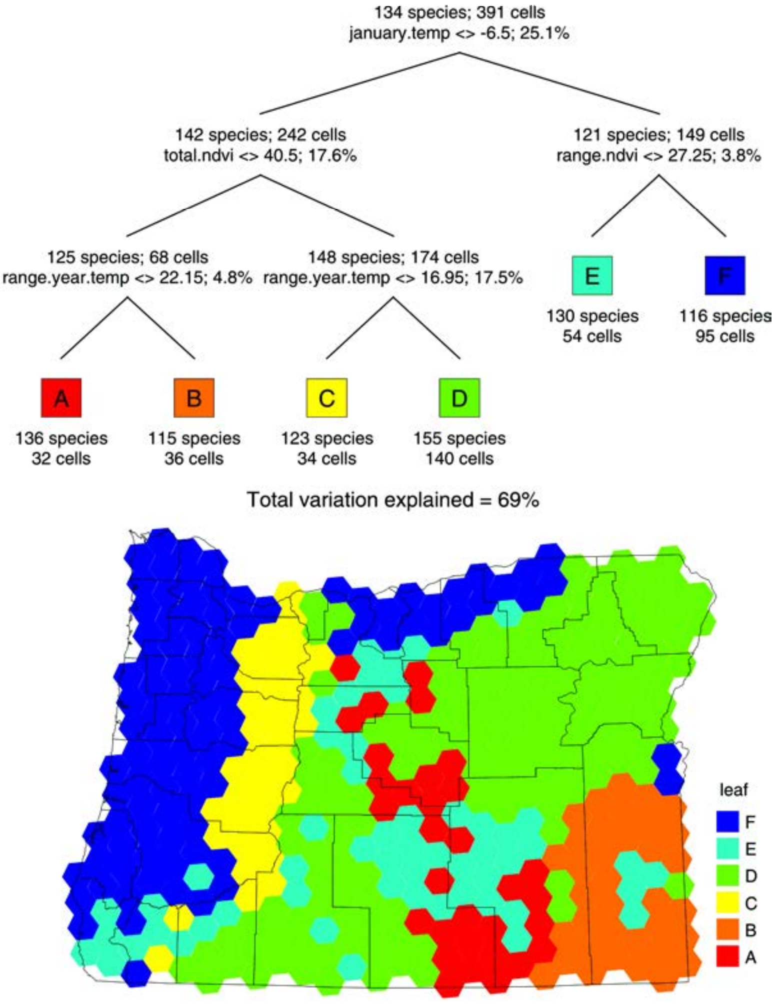

Figure 3. Regression tree and map of prediction relationships for native bird species richness in Oregon. At each node in the regression tree the first line of text indicates (a) the mean of the response variable for all observations at that node; and (b) the number of observations. The second line of text indicates (a) the variable selected for splitting, (b) the splitting value (introduced by the symbol "<>"); and (c) the percent of deviance attributed to the node. At the leaves a color box indicates the color and letter assigned to the node. The first line of text below the box indicates the mean of the response. The second line indicates the number of observations.

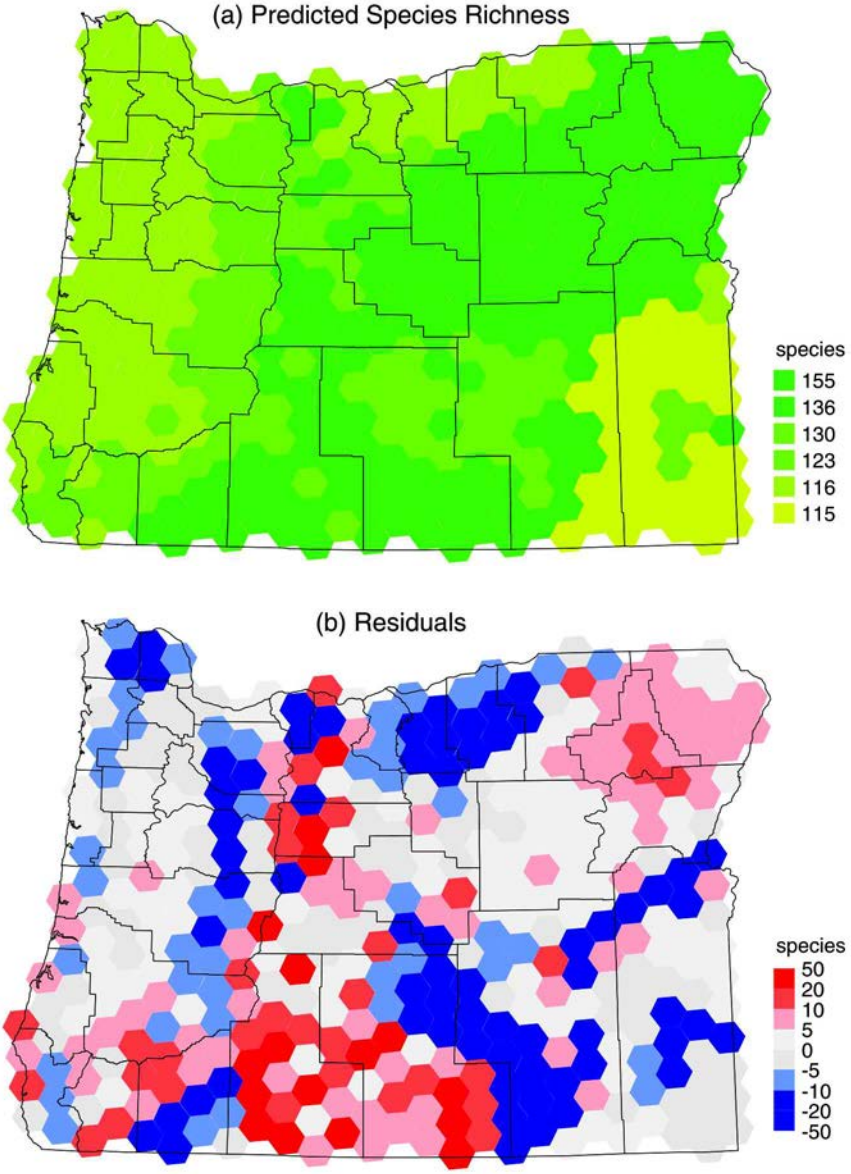

Figure 4. (a) Map of the predicted value of the response based on the regression tree for birds in Oregon. Unless identical values are predicted, there will be as many levels of the predicted response as leaves in the tree. (b) Map of residuals from the predicted response. Under-predictions are shown in increasing intensities of red and over-predictions in increasing intensities of blue.

3.2 Threatened Bird Species in Countries of the World

The second dataset was developed from the World Resources Institute (1994) and National Oceanic and Atmospheric Administration (Climatological Baseline Station Data over Land, www.ncdc.noaa.gov). For 118 countries of the world for which there were complete data, we used the number of species of threatened birds as a response variable and 15 functions of climate, human demography, military forces, energy use, and land use and disturbance as predictors.

The land area of each country (in km^2) created the first split, accounting for the largest amount of deviance (25.9%) in the data (Figure 5). Larger land area was associated with more threatened species on average, a result that could be expected since the total number of species, threatened and not, would also be expected to positively correlate with land area. The larger countries were further divided between those with higher annual precipitation (purple) and those with lower (blue). More precipitation was associated with more species. The smaller countries were first split on the size of their armed forces (in thousands of soldiers); larger armies were associated with more threatened species. Countries with larger armed forces were further divided by January temperatures. Countries with cooler January temperatures (orange) had fewer species on average. Countries with warmer January temperatures were finally split by percentage of urban population. Countries with greater urban population (dark green) were associated with more threatened species than those with lesser (light green).

Predicted response and residuals from the fit (Figure 6) showed underprediction in middle and low latitudes in the western hemisphere and overprediction in high latitudes. A number of countries with large land areas were not well predicted. A separate model that did not use land area as a predicting variable did not significantly change the residual pattern for large countries.

4. SIZING THE TREE

An important issue in choosing the size of regression trees is the intended use. Larger trees provide greater accuracy for predicting to sites where the response has not been measured. On the other hand, smaller trees may be more appropriate for understanding and interpreting relationships among the data (Clark and Pregibon 1992) because nodes accounting for less deviance provide less information. Breiman et al. (1984) developed the method of determining tree size by initially overfitting a tree and then pruning it back using one of several techniques. Sifneos et al. (in press) evaluated these techniques and others using simulated and published data. In determining the size to which to prune trees, graphical methods for visualizing choices are helpful. One graphical aid is a diagram of the overfitted tree indicating the order of pruning for each node. In the Oregon bird example, the tree diagram shows the sequence of pruning from 13 leaves to one (Figure 7).

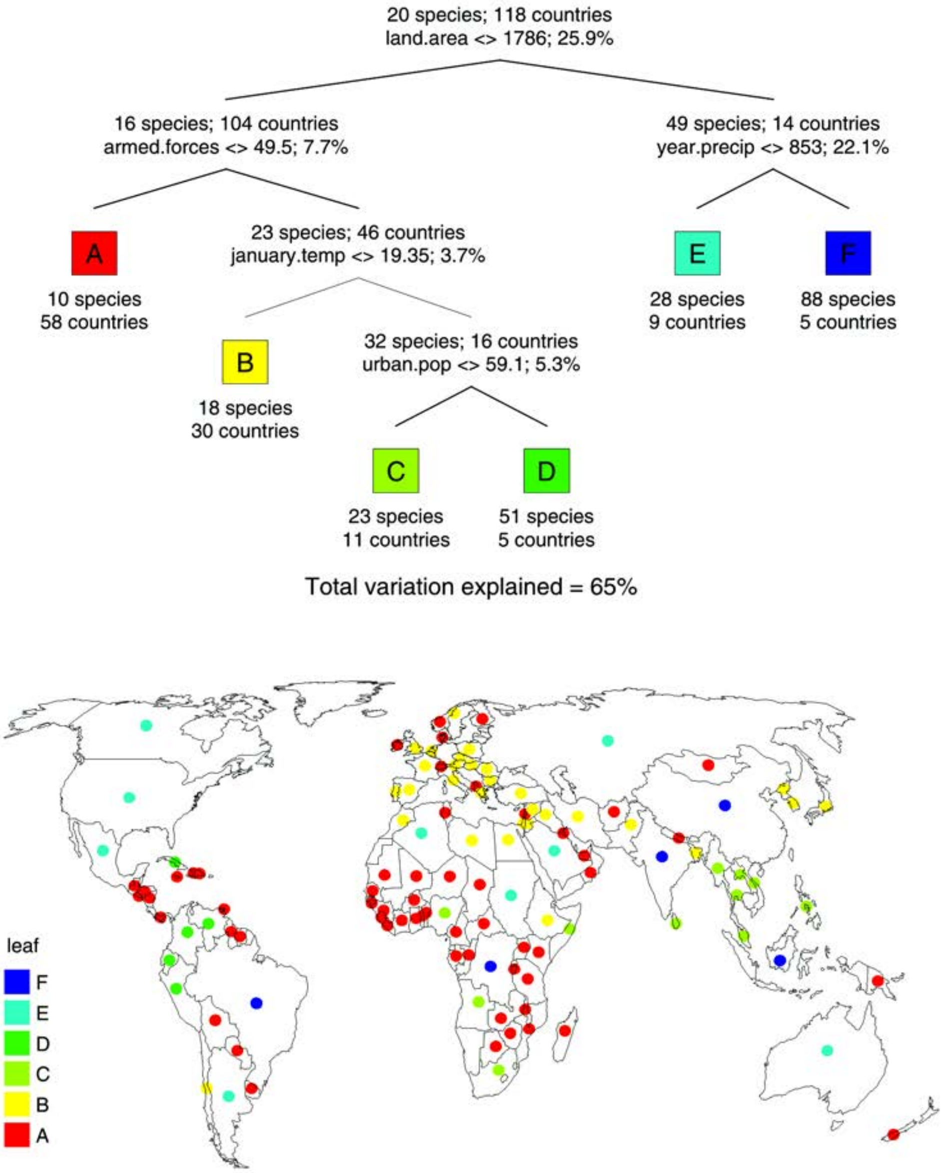

Figure 5. Regression tree and map of prediction relationships for threatened bird species richness in 118 countries in the world. Countries are symbolized by a colored point rather than a color applied to their complete land area (which method would be a choropleth map in cartographic terminology). This choice of the geometry of symbolism emphasizes the equality of countries as legal entities rather than their physical inequalities of size.

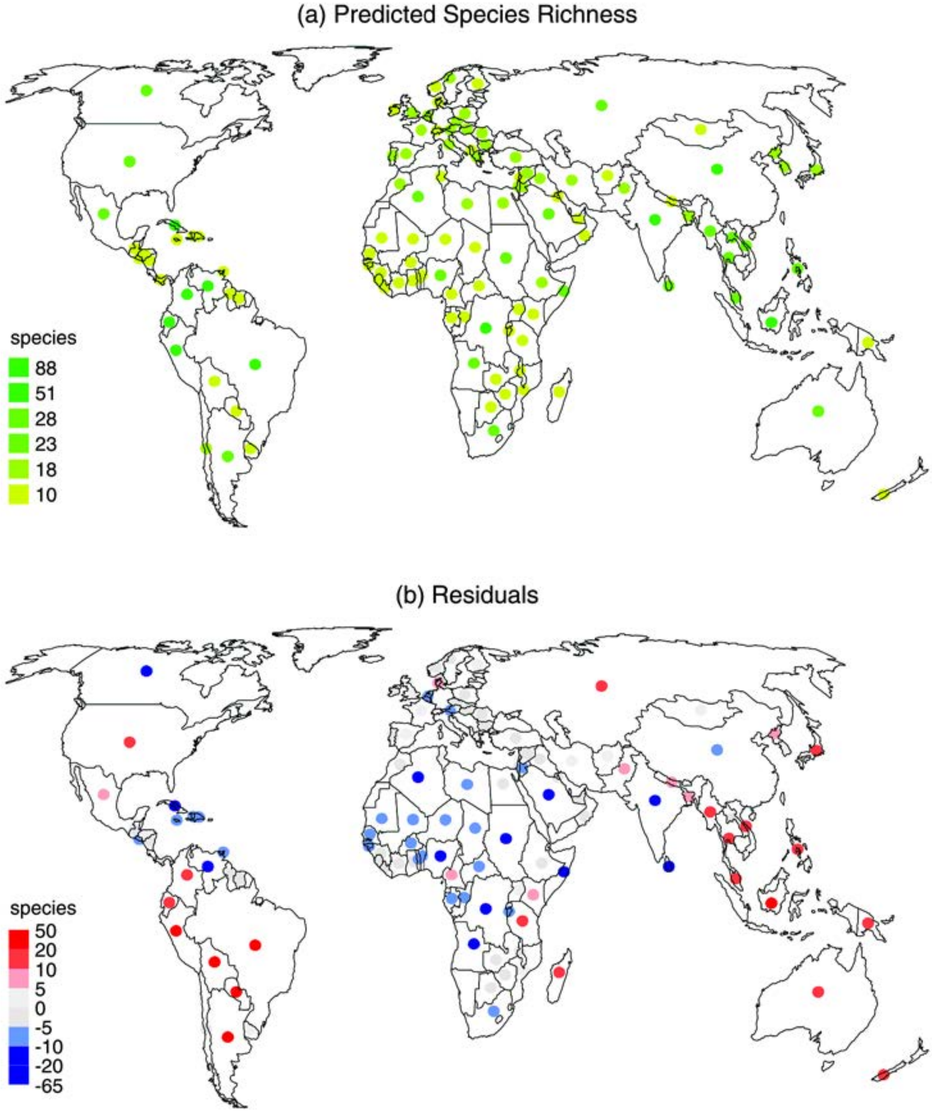

Figure 6. Maps of (a) the predicted response and (b) residuals from the threatened bird species regression tree.

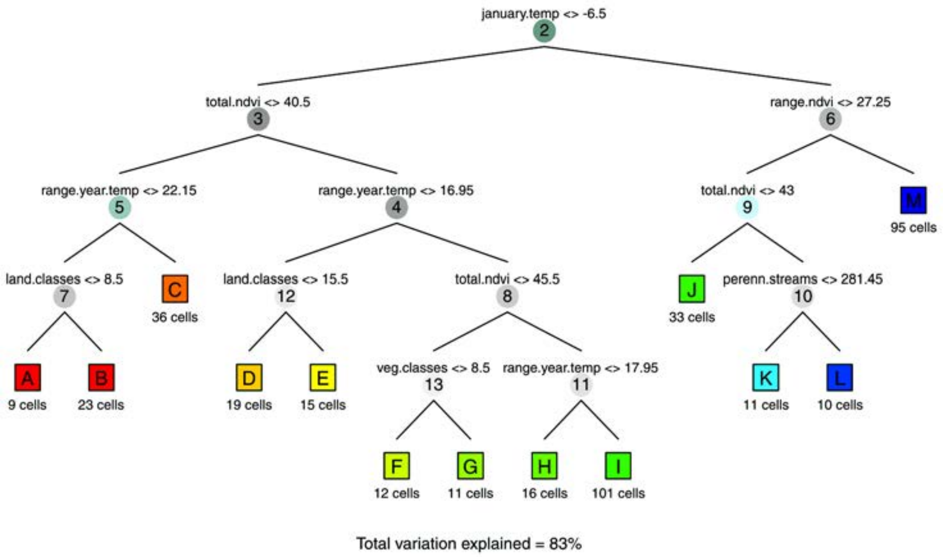

Figure 7. Regression tree diagram for birds in Oregon showing the order in which nodes will be pruned from the overfitted tree. The line of text at each node indicates the splitting variable. Below the split is a number indicating the size of the tree when the node is the next to be pruned. These numbers are shaded from darkest (last to prune) to lightest (first to prune) to aid in seeing the pruning order. At the leaves, the number of observations and the color and letter assigned to the node are indicated. In pruning the tree, the node that is directly above leaves F and G would be the first node to be pruned because it is the largest number indicating that the full tree has 13 leaves; the node above leaves D and E would be the second to be pruned and the tree would have 12 leaves when it is next to be pruned, and so on.

A companion to the pruning tree diagram is the set of maps corresponding to each tree in the pruning sequence. For the Oregon bird example, the map sequence shows which parts of the state remain intact as the state is successively partitioned by additional leaves (Figure 8). Examining the increasing spatial fragmentation as tree size is increased can be helpful in deciding where to prune the tree. Based on these graphical aids and measures recommended in Sifneos et al. (in press) we selected the tree sizes for both of the example datasets.

5. DISCUSSION

How do maps of the regression trees help reveal spatial patterns of the contingent relationships among the data? In the Oregon bird data, for example, there is spatial clustering of observations defined by the leaves of the regression tree (Figure 3). The yellow grid cells are almost all contiguous as are the green cells. Furthermore the yellow and green areas are close in prediction space as well as in geographical space compared to areas from other pairs of less similar colors. Of course, neighbors in geographical space can be from quite different parts of the prediction tree, for example, North American versus Central American countries in the example of threatened bird species of the world (Figure 5).

The cartography of regression trees also can provide a regionalization of the prediction relationships. In the Oregon dataset, the regions revealed by the regression tree corresponded, to a moderate degree, to ecological regions that were developed from different criteria (Omernik 1987). This suggested that the geography of environmental associations of bird species richness was represented reasonably well by these regions. In a study using similar data, regression tree regions for Oregon fish species and their associations also showed correspondence with more general ecological regions (Rathert et al. 1999).

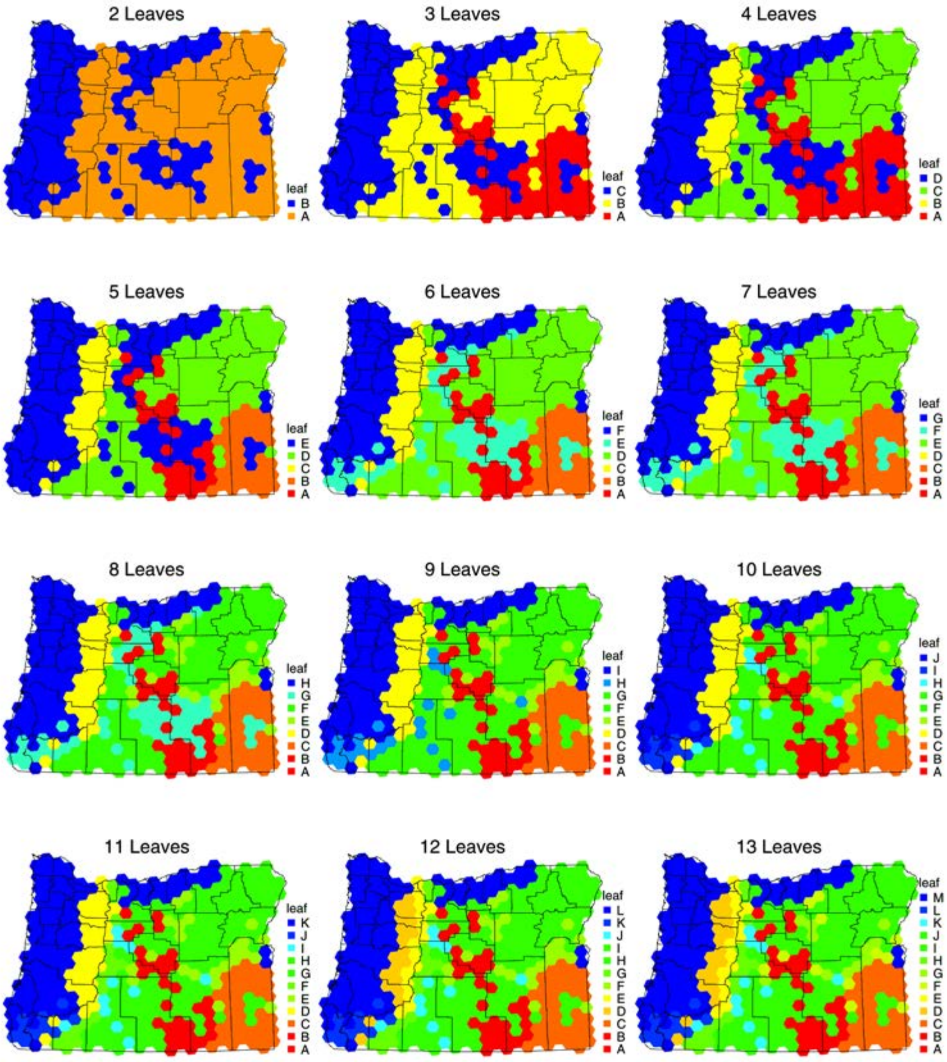

Figure 8. The sequence of maps of prediction relationships corresponding to the pruning sequence in Figure 7. The map corresponding to the tree in Figure 7 with all 13 leaves is in the bottom row at the right. The map corresponding to the tree in Figure 7 with the first node pruned off, leaving 12 leaves, is in the bottom row in the center, and so on.

Mapping of regression tree results can suggest interesting geographical associations. In the tree for threatened bird species there was a group of countries from central and western Europe (excluding Scandinavia, Ireland, and Switzerland), the middle east (excluding Lebanon), Iran, Pakistan, Japan, and the Koreas, that had in common larger armed forces and smaller numbers of threatened species.

The color mapping method presented here using recursive partitioning of the color space limits the depth of trees that can be effectively displayed. At depths greater than about six levels it is difficult to distinguish adjacent leaves using our method and our printing devices. Additional or alternative symbolism types would be needed such as repeating line or dot patterns or leaf identifiers.

The mapping procedure we have proposed could be used on treed regression as well. Alexander and Grimshaw (1996) proposed that, rather than the mean of the subset of observations at each leaf, the best linear regression of any one of the predicting variables be used for estimating the response. They suggested that trees produced by their method would be expected to be of lesser depth than trees using the Breiman et al. (1984) methods. Potential geographical relationships could still be detected by mapping the subsets of observations in the leaves, and the prediction accuracy would be likely to be enhanced by the univariate regression estimates of the response.

Another potential application of spectral tree mapping is for dendrograms from hierarchical clustering methods. Assuming that the clustering method creates an unambiguous ordering of the composition of observations into clusters, then the observations belonging to each cluster can be assigned colors in the same way as in the regression tree. If the dendrogram has more than two descendants at a node, then the hue spectrum for the node would be divided into a corresponding number of equal intervals.

Recent developments in cartographic display, as well as in other types of statistical graphics, have emphasized dynamic user interaction in graphic displays (e.g., Dykes 1998; Unwin and Hofmann 1998). The methods we propose for displaying predicted values, residuals, and prediction relationships in regression trees (and other hierarchical models) could be enhanced with dynamic color selection and dynamic linking between observations on the map and positions in the hierarchical tree model. Our contribution is in the display of the geography of the prediction or clustering relationships. As an aid to the interpretation of spatial data, the cartography of regression trees supplements the exploratory power of the hierarchical prediction structure of tree models by communicating spatial relationships in these models.

Software to prepare graphs of tree models from regression trees, classification trees, and hierarchical clustering methods, along with associated maps of the prediction or clustering space, is available as a contributed extension, called “maptree,” to the R language at http: //www.R-project.org/.

ACKNOWLEDGMENTS

We appreciate helpful comments and reviews from Raymond O'Connor, Anthony Olsen, Scott Urquhart, John Van Sickle, three anonymous reviewers, and editor Andreas Buja. We acknowledge support from the National Science Foundation under a grant to Raymond O'Connor at the University of Maine, and additional support from the USDA Forest Service, the U.S. Environmental Protection Agency, and the U.S. Department of Defense. This article has been subjected to review by the U.S. EPA and approved for publication. The conclusions and opinions are solely those of the authors and are not necessarily the views of the Agency.

[Received January 2000. Revised April 2001.]

REFERENCES

Alexander, W. P., and Grimshaw, S. D. (1996), "Treed Regression," *Journal of Computational and Graphical Statistics*, 5, 156–175.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, New York: Chapman & Hall.

Brewer, C. A. (1994), "Color Use Guidelines for Mapping and Visualization," in *Visualization in Modern Cartography*, eds. A. M. MacEachren and D. R. Fraser Taylor, London: Pergamon, pp. 123–147.

Clark, L. A., and Pregibon, D. (1992), "Tree-Based Models," in *Statistical Models in S*, eds. J. M. Chambers and T. J. Hastie, Pacific Grove, CA: Wadsworth & Brooks, pp. 377–419.

Davis, F. W., Michaelsen, J., Dubayah, R., and Dozier, J. (1990), "Optimal Terrain Stratification for Integrating Ground Data from FIFE," in *Symposium on FIFE, First ISLSCP Field Experiment*, Boston: American Meteorological Society, pp. 11–15.

Dykes, J. F. (1998), "Cartographic Visualization: Exploratory Spatial Data Analysis With Local Indicators of Spatial Association Using Tcl/Tk and CDV," *The Statistician*, 47, 3, 485–497.

Hill, B., Roger, T., and Vorhagen, F. W. (1997), "Comparative Analysis of the Quantization of Color Spaces on the Basis of the CIELAB Color-Difference Formula," *ACM Transactions on Graphics*, 16, 109–154.

Iverson, L. R., and Prasad, A. M. (1998), "Predicting Abundance of 80 Tree Species Following Climate Change in Eastern United States," *Ecological Monographs*, 68, 465–485.

Moore, D. M., Lees B. G., and Davey, S. M. (1991), "A New Method for Predicting Vegetation Distributions Using Decision Tree Analysis in a Geographic Information System," *Environmental Management*, 15, 59–71.

O'Connor, R. J., Jones, M. T., White, D., Hunsaker, C., Loveland, T., Jones, B., and Preston, E. (1996), "Spatial Partitioning of Environmental Correlates of Avian Biodiversity in the Conterminous United States," *Biodiversity Letters,* 3, 97–110.

Omernik, J. M. (1987), "Ecoregions of the Conterminous United States," *Annals, Association of American Geographers*, 77, 118–125.

Rathert, D., White, D., Sifneos, J. C., and Hughes, R. M. (1999), "Environmental Correlates of Species Richness for Native Freshwater Fish in Oregon, USA," *Journal of Biogeography*, 26, 257–273.

Robertson, P. K., and O'Callaghan, J. F. (1986), "The Generation of Color Sequences for Univariate and Bivariate Mapping," *IEEE Computer Graphics and Applications*, 6, 24–32.

Sifneos, J. C., White, D., and Urquhart, N. S. (in press), "A Comparison of Pruning Methods for Regression Trees: Evidence from Simulation and Published Studies, " *Biometrical Journal*.

Skidmore, A. K., Gauld, A., and Walker, P. (1996), "Classification of Kangaroo Habitat Distribution Using Three GIS Models," *International Journal of Geographical Information Systems*, 4, 441–454.

Tajima, J. (1983), "Uniform Color Scale Applications to Computer Graphics," *Computer Vision, Graphics, and Image Processing*, 21, 305–325.

Unwin, A., and Hofmann, H. (1998), “New Interactive Graphics Tools for Exploratory Analysis of Spatial Data,” in *Innovations in GIS 5*, ed. S. Carver, London: Taylor & Francis Ltd., pp. 46–55.

Walker, P. A. (1990), “Modelling Wildlife Distributions Using a Geographic Information System: Kangaroos in Relation to Climate,” *Journal of Biogeography*, 17, 279–289.

Ware, C. (1988), “Color Sequences for Univariate Maps: Theory, Experiments, and Principles,” *IEEE Computer Graphics and Applications*, 8, 41–49.

White, D., Preston, E. M., Freemark, K. E., and Kiester, A. R. (1999), “A Hierarchical Framework for Conserving Biodiversity,” in *Landscape Ecological Analysis: Issues and Applications*, eds. J. M. Klopatek and R. H. Gardner, New York: Springer-Verlag, pp. 127–153.

White, D., and Sifneos, J. C. (1997), “Mapping Multivariate Spatial Relationships from Regression Trees by Partitions of Color Visual Variables,” *Proceedings, Auto-Carto 13*, Seattle, WA: American Congress on Surveying and Mapping, pp. 86–95.

World Resources Institute, (1994), *World Resources 1994-95*, New York: Oxford University Press.

Wyszecki, G., and Stiles, W. S. (1982), *Color Science: Concepts and Methods, Quantitative Data and Formulae* (2nd ed.), New York: John Wiley.