MAPPING MULTIVARIATE SPATIAL RELATIONSHIPS FROM REGRESSION TREES BY PARTITIONS OF COLOR VISUAL VARIABLES

Denis White and Jean C Sifneos
Department of Geosciences
Oregon State University
Corvallis, OR, 97331 USA

ABSTRACT. In classification and regression tree (CART) analysis, the observations are successively partitioned into a prediction tree. At each node in the tree, the CART algorithm searches for the value of one of the predictor variables that explains the greatest amount of variation in the response variable. The observations are split into two groups at each node according to this splitting criterion until the tree reaches a size that balances predictive power and parsimony. We illustrate a method for mapping the spatial relationships in a prediction tree when the cases are spatial. Each leaf in the tree has a unique set of predictor variables and corresponding value ranges that predict the value of the response variable at the observations belonging to the leaf. If the tree is arranged such that observations with lower values of the splitting variable are always on the left at each node, then there is an unambiguous ordering to the tree. One method for assigning mapping symbols to the observations of the leaves is by locating each leaf in a corresponding position along the continuum of one of the color visual variables. Observations that are closer in perceptual value to others indicate a closer relationship in the structure of prediction.

INTRODUCTION

Classification and regression trees (CART) are a multivariate analysis technique made popular by Breiman *et al.* (1984). Applications are varied: examples include machine learning (Crawford, 1989), medicine (Efron and Tibshirani, 1991), optical character recognition (Chou, 1991), soil classification (Dymond and Luckman, 1994), forest classification (Moore *et al.*, 1991), vegetation ecology (Davis *et al.*, 1990; Michaelson *et al.*, 1994), animal distribution (Walker, 1990), biodiversity (O'Connor *et al.*, 1996), and others.

In an application where the cases are spatial locations, the geography of the prediction tree results may reveal insights into mechanistic relationships 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 CART results includes Davis *et al.* (1990), Walker (1990), Moore *et al.* (1991) and O'Connor *et al.* (1996). We explore this idea by proposing an objective method for assigning map symbols to the leaves of regression (or classification) trees. We illustrate this mapping method with both simulated and real data.

METHODS

In regression tree development, the midpoints between all values of all of the predictor 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 sets based on the split and the process recursively repeats on the two descendent sets. Splitting continues until a stopping criterion is reached. We used the cross-validation pruning techniques of Breiman *et al.* (1984), as implemented by Clark and Pregibon (1992), and as investigated by Sifneos *et al.* (in preparation), to determine the optimal size of trees.

We prepared two simulated data sets as examples. The first set consisted of three predictor variables defined as two level (x1), or three level (x2 and x3), step functions. The response variable (y) was defined as a four level step function. All variables were defined on a 10 by 10 grid, simulating a spatial surface. The steps were defined on one quarter or one half of the grid (Figure 1). The second simulated set also consisted of three predictors, but these were samples from a lognormal distribution (x1) and two different normal distributions (x2 and x3), respectively. The response (y) was defined differently in each quadrant of the grid to simulate the contingent effects of hierarchical interactions that CART is well suited to analyze. The first quadrant was defined as $y = x1 + 2x2 + 3x3$, for example, and the other quadrants as indicated in Figure 2.

In addition, we used portions of a data set from a fish biodiversity study (Rathert *et al.*, in preparation). For illustrating the regression tree mapping method we used total fish species richness, including native and introduced species, as a response variable. We used 20 predictor variables representing climatic, elevational, hydrographic extent, and human impact effects (Figure 4). All variables were provided for 375 equal area sample units covering the state of Oregon. (The variable representing the length of 4th and higher order streams in each sampling unit is not shown in Figure 4.)

We can think of the mapping of regression trees in the framework of measurement scales. The terminal nodes, or leaves, of a tree contain observations that have a unique chain of prediction rules with respect to other leaves. The uniqueness property confers at least a nominal scaling on the leaves. Because the predictor splits can be arranged in an unambiguous order with lower values of continuous variables, for example, always appearing in the left branches, an ordinal scaling can be imposed as well. (Nominal predictor variables can meet this criterion with an arbitrary ordering of categories.) More ambitiously, we may attempt to convey distance in prediction space by mapping leaf positions to an interval scale. Color visual variables are good candidates to symbolize these scaling distinctions. In this paper we present tree mapping using

ordinal scaling of the value or brightness dimension. We select a number, equal to the number of leaves, of equally-spaced division points along the value scale. In another paper we present a more refined implementation of this idea using a recursive partition of the hue spectrum to mimic the recursive partitioning of the observation space by the regression tree (White and Sifneos, in preparation).

RESULTS

The stepped prediction and response surfaces (Figure 1) produced a simple tree that has perfect prediction. That is, the variation explained, computed as the ratio of sums of squares in the leaves to that of the root node, subtracted from one, is exactly equal to 1. The tree diagram expressing the prediction relationship (Figure 1) followed the pattern of predictors precisely: the first split recognized the division of the observation grid into two halves by x1; the left branch of the tree representing the top half of the grid was split by x2; and the right branch representing the bottom by x3. A multiple linear regression on this data also achieved perfect prediction with an R-Squared of 1.

The contingent response from normally and lognormally distributed predictors (Figure 2) produced, in one realization, a tree with six leaves (Figure 3). We applied the value scaling to the leaves and mapped the prediction groups of observations on the simulated study area grid (Figure 3). The variation explained by the tree was 0.71. A multiple regression with no interactions between predictors produced a R-Squared of 0.28. (A multiple regression including interactions between predictors would have a higher R-Squared.)

A regression tree analysis of the fish data set produced a tree with seven leaves (Figure 5). Each of the six splits used a different predictor variable. The variation explained by the tree was 0.72. A multiple regression fit with no interactions had a R-Squared of 0.50, using seven predictor variables determined through stepwise procedures. The map of prediction groups from the regression tree revealed a strong east-west structure in Oregon (Figure 6). On the west side of the Cascades, climate and elevation variables formed the prediction, while on the east side, stream length variables. The value scale mapping of leaf prediction groups with gray tones helped to identify this structure. Comprehensive analysis of this data and an interpretation of the biogeography will be found in Rathert *et al.* (in preparation).

ACKNOWLEDGEMENTS

We acknowledge support from agreements CR 821672 between US EPA and Oregon State University, PNW 92-0283 between US Forest Service and OSU, DW12935631 between US EPA and USFS, and DOD SERDP Project #241-EPA. This research has not been officially reviewed by US EPA and no endorsement should be inferred.

REFERENCES

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

Chou, P A. (1991). Optimal partitioning for classification and regression trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:340-354.

Clark, L A, D Pregibon. (1992). Tree-based models. In: *Statistical models in S.* JM Chambers & TJ Hastie, editors. Wadsworth & Brooks, Pacific Grove, CA. pp. 377-419.

Davis, F W, J Michaelsen, R Dubayah, J Dozier. (1990). Optimal terrain stratification for integrating ground data from FIFE. In: *Symposium on FIFE, First ISLSCP Field Experiment.* Amer. Meteor. Soc., Boston, MA. pp. 11-15.

Dymond, J R, P G Luckman. (1994). Direct induction of compact rule-based classifiers for resource mapping. *Int. J. Geog. Info. Sys.*, 8:357-367.

Efron, B, R Tibshirani. (1991). Statistical data analysis in the computer age. *Science*, 253:390-395.

Michaelsen, J, D S Schimel, M A Friedl, F W Davis, R C Dubayah. (1994). Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. *J. Veg. Sci.*, 5:673-686.

Moore, D M, B G Lees, S M Davey. (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, M T Jones, D White, C Hunsaker, T Loveland, B Jones, E Preston. (1996). Spatial partitioning of environmental correlates of avian biodiversity in the conterminous United States. *Biodiversity Letters,* in press.

Rathert, D, D White, J C Sifneos, R M Hughes. (In preparation). Environmental correlates of species richness in Oregon freshwater fishes.

Sifneos, J C, D White, N S Urquhart, D Schafer. (In preparation). Selecting the size of regression tree models.

Walker, P A. (1990). Modelling wildlife distributions using a geographic information system: kangaroos in relation to climate. *J. Biogeog.*, 17:279-289.

White, D, J C Sifneos. (In preparation). Mapping multivariate spatial relationships from regression trees by recursive binary partitions of the spectrum.

Figure 1

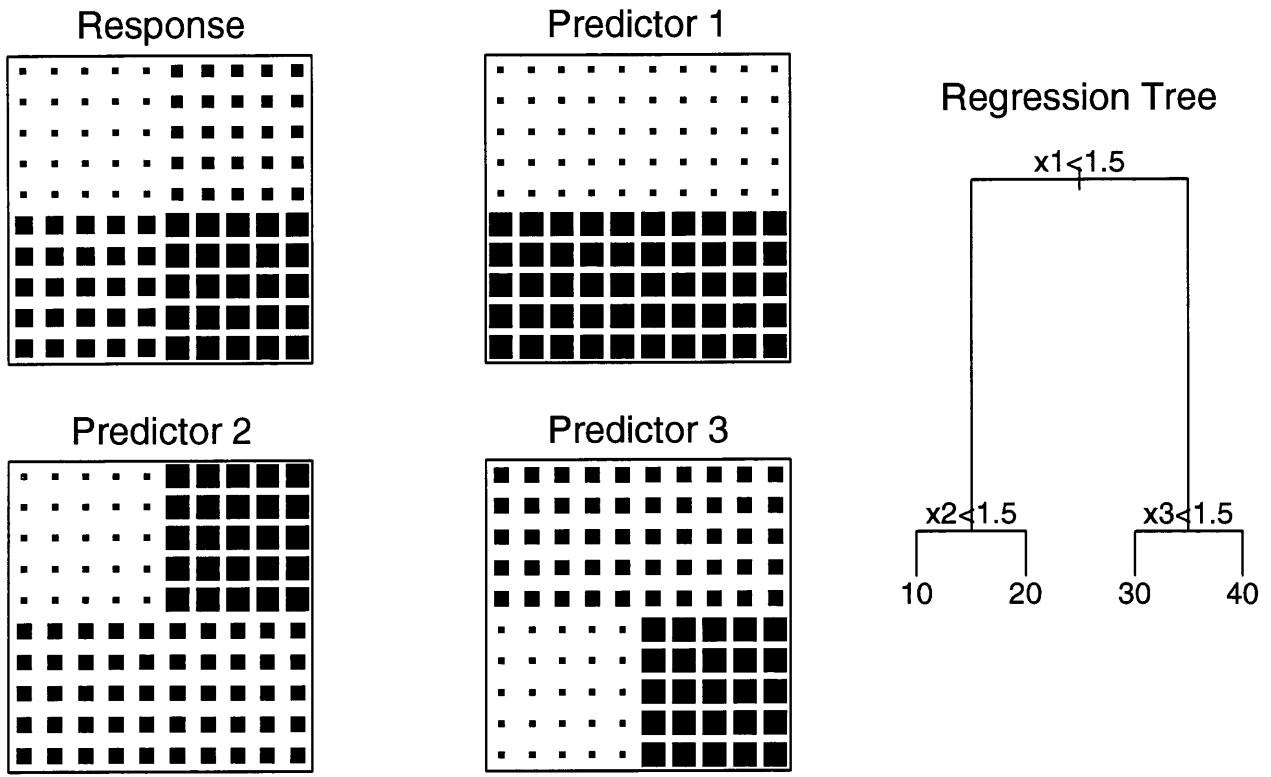

Stepped Response and Predictors
Response
Predictor 1
Predictor 2
Predictor 3
Regression Tree
x1<1.5
x2<1.5
x3<1.5
10
20
30
40

Figure 2

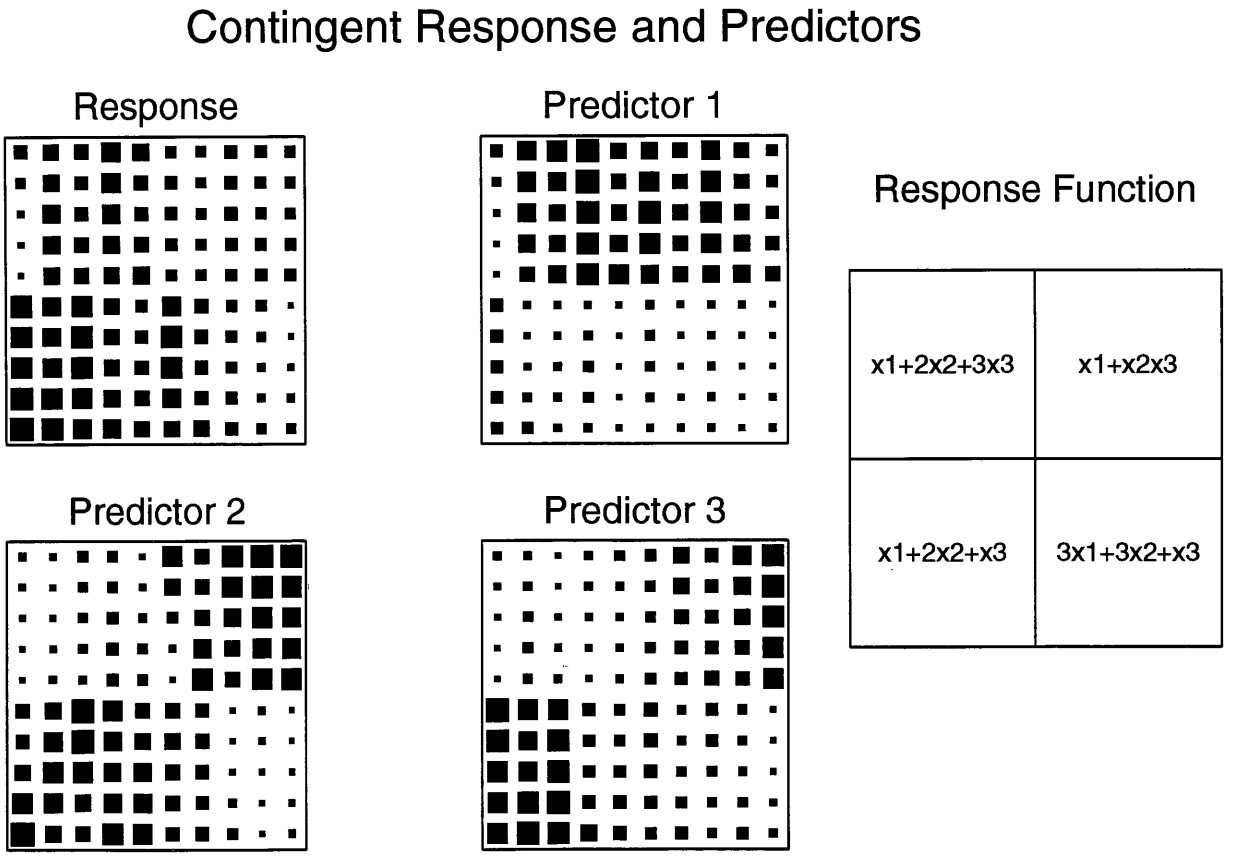
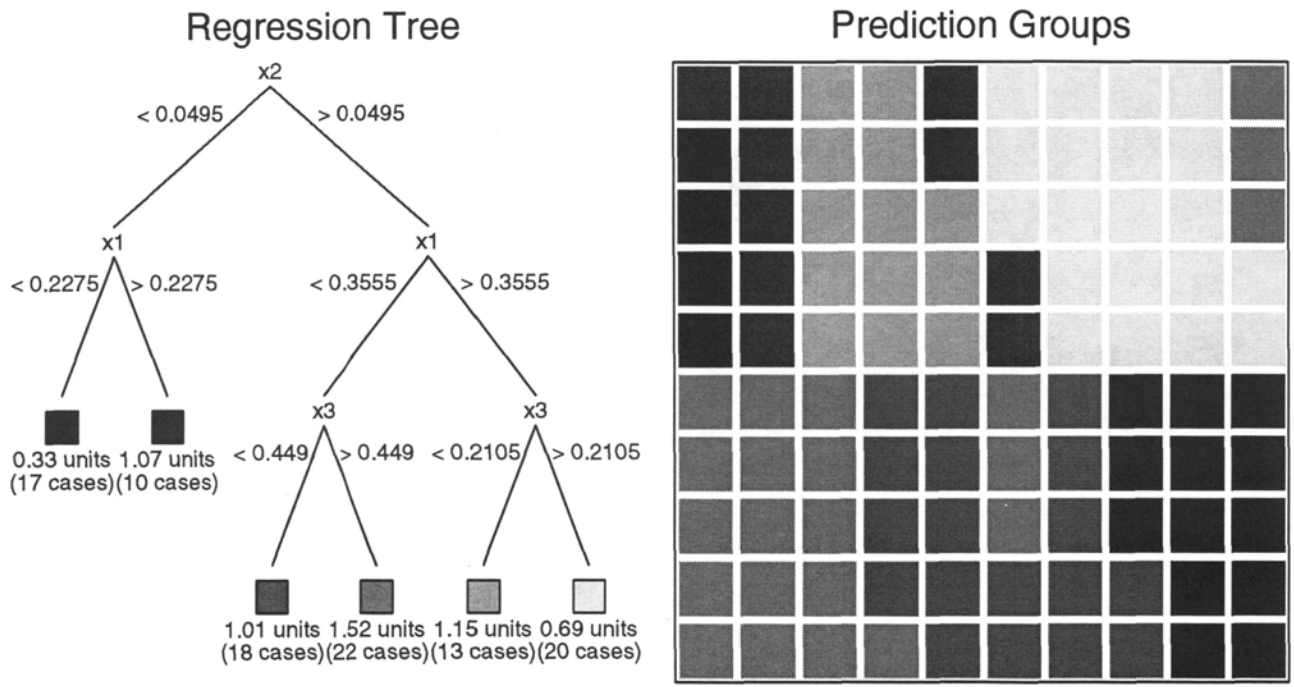

Contingent Response and Predictors
Response
Predictor 1
Predictor 2
Predictor 3
Response Function
x1+2x2+3x3
x1+x2x3
x1+2x2+x3
3x1+3x2+x3

Figure 3


Contingent Response
Regression Tree
Prediction Groups
x2
< 0.0495
> 0.0495
x1
x1
< 0.2275
> 0.2275
< 0.3555
> 0.3555
x3
x3
0.33 units
(17 cases)
1.07 units
(10 cases)
< 0.449
> 0.449
< 0.2105
> 0.2105
1.01 units
(18 cases)
1.52 units
(22 cases)
1.15 units
(13 cases)
0.69 units
(20 cases)

Fish Species Richness and Predictors
Fish.Species
Jan.Temp.Min
Jul.Temp.Max
Annual.Temp.Range
Ann.Temp.Rng.Std.Dev
Monthly.Temp.Std.Dev
Annual.Precip.Min
Annual.Precip.Max
Annual.Precip.Range
Annual.Precip.Mean
Elevation.Min
Elevation.Max
Elevation.Range
First.Order.Streams
Second.Order.Streams
Third.Order.Streams
Intermittent.Streams
Perennial.Streams
Road.Length
Human.Population

Figure 4

Fish Species Richness
Regression Tree
Annual.Temp.Range
< 16.625
> 16.625
Elevation.Range
< 484.022
> 484.022
Perennial.Streams
< 45.33
> 45.33
Jan.Temp.Min
< -7.5
> -7.5
Higher.Order.Streams
< 35.785
> 35.785
Annual.Precip.Range
< 953
> 953
29 species
(16 cases)
10 species
(20 cases)
2 species
(69 cases)
19 species
(24 cases)
6 species
(105 cases)
14 species
(43 cases)
9 species
(98 cases)

Figure 5

Fish Species Richness

Prediction Groups

Figure 6