DATA MINING IN RECREATION RESEARCH: DETERMINANTS OF PLACE ATTACHMENT IN RECREATIONAL CAMPS AND COTTAGES

Jeremy S. Flaherty
Department of Recreation Resource & Ski Resort Management
Lyndon State College
P.O. Box 919
Lyndonville, VT 05851

Rodney R. Zwick
Lyndon State College

Ralph B. Brown
Brigham Young University

Abstract

This paper uses data collected by Lyndon State College and the Vermont Fish and Wildlife Department from the Northeast Kingdom Camp Owner Survey to show how Classification and Regression Tree (CART) analysis, a data-mining program, can be used in social research. Unlike traditional regression techniques, CART is not constrained by typical parametric assumptions or missing data. Not only can CART be used as a stand-alone application, it can be used to improve the performance of traditional quantitative methods like OLS regression.

1.0 Introduction

One of the most important and most commonly used statistical tools in the social sciences is regression analysis. Much of the data being used, however, are complex and contain missing values, and the relationships between variables may be non-linear and involve interactions. The commonly used regression techniques have difficulty dealing with such data. Classification and Regression Tree (CART) analysis, a data mining program, is a more modern statistical technique ideally suited for exploring and modeling this complex data.

Traditional regression methods, such as linear or logistic regression, have several assumptions that the data must meet in order to work; and several problems may arise in the data which are detrimental to the success of modeling with traditional methods. The data must be normally distributed and must be linearly related to the dependent variable, or it has to be transformed to make it so, and missing data will cause a case to be dropped from the analysis. Moreover, traditional methods are difficult to use to model interaction effects, much less discover an unsuspected interaction; and variable selection can be cumbersome when faced with numerous potential independent variables, especially when there is no theory to dictate which variables should be important. In the end, assuming we have normalized all the data and made it linear, included any important interaction terms, and selected the best model, interpretation is still less than intuitive.

CART does not have these problems. CART automatically determines the most important variables and uncovers interactions of any complexity, handles both linear and non-linear relationships, includes methods for dealing with missing data, and makes no assumptions about the distributions of the independent variables. The resulting regression tree is very easy to interpret even for someone with no quantitative expertise.

CART has been used with a great deal of success in other fields. It has been used to develop triage rules for AIDS and cardiac patients (Lewis 2000, Long, Griffith, Selker, & D'Agostino 1993), and for identifying cancerous tissue (Zhang, Yu, Singer, & Xiong 2001); to determine high and low credit risks (Arminger, Enache, & Bonne 1997); to determine optimal nesting areas of smallmouth bass so that they can be protected (Rejwan, Collins, Brunner, Shuter, & Ridgway 1999); it has even been used to determine which milk cows should be culled from a herd (Tronstad & Gum 1994). In this paper we introduce CART by using it to discover the determinants of place attachment among camp owners in northeastern Vermont. Using data collected in the Northeast Kingdom Camp Owner Survey, we will compare CART, particularly the regression tree tool, to multiple linear regression to show the pros and cons of this method, and show that CART is a valuable tool to be used in conjunction with other techniques.

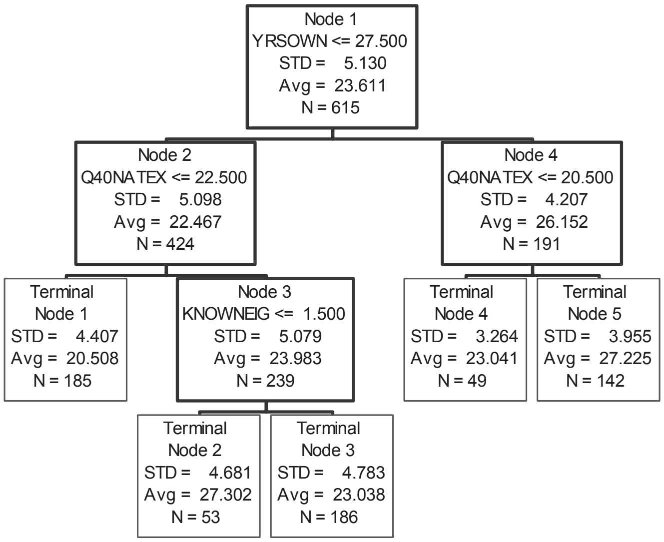

Figure 1.—Regression Tree Predicting Score on Location-Historic Index.

2.0 What is CART

There are three ways that CART can be used that will be discussed in this paper: 1) CART can be used as a stand-alone tool for analysis; 2) CART can be used to select from a long list the most important variables to be included in a regression model; and 3) CART can be used to split the data into multiple groups, each of which can then be used to build separate regression models. First, we will describe what CART is.

Classification and Regression Trees are made up of nodes. Nodes simply represent groups of data. The root node, the original node at the top of the tree, contains every case in the data set (see Fig. 1 for an example of a tree). Every node that is split is a parent node, and the two nodes that come out of it are child nodes. The two resulting child nodes together contain all of the cases that were in their parent node. The final nodes that do not spit are terminal nodes. All of the terminal nodes combined contain all of the cases that were in the root node. Each and every case ends up in only one terminal node.

CART is a form of recursive binary partitioning. Classification and regression trees are grown by repeatedly splitting the data into two mutually exclusive groups. A node, or the data, is split at a point in any independent variable that creates two new nodes that are more homogeneous in the dependent variable than the was the original node. The new nodes are then spit again using the exact same process. Tree growth ends when no spit can increase the level of homogeneity.

CART chooses the best splitting point from all possible points in the data. Within a categorical variable, the number of possible spits is equal to 2k-1-1, where k is the number of categories in the variable. So a variable with only two categories, such as a yes or no answer to a survey question, can be spit only one way, while a variable with, say, five categories could be split 15 different ways, allowing for every possible combination of categories. For quantitative variables, there are u-1 possible splits, where u is the number of unique values the variable may take. CART looks at every possible split in every variable to find the one that produces the greatest homogeneity in the two resulting child nodes. In a regression tree, which is what will be described in this paper, homogeneity is measured by the standard deviation of the dependent variable. A lower standard deviation represents greater homogeneity. When CART splits a node, the two child nodes will have a combined weighted standard deviation that is lower than the parent node's standard deviation. The weighted standard deviation takes into account the number of cases in the child node.

CART continues splitting the data until no more splits are possible. This happens when all of the cases in a node have the same value in the dependent variable, or when no possible split will increase the homogeneity. The resulting tree overfits the data, just as continuing to add independent variables to a regression model even though there is no important increase resulting in the r-square will do. The overfit tree fits the data very well, but would not be very useful if applied to a new set of data because it takes into account all of the idiosyncrasies of a particular set of data. To find an optimal tree, CART prunes the tree, eliminating nodes further down the tree that are capturing the idiosyncrasies of the data. To do this, CART can use a set of test data, if available, that was not used to build the tree, or it uses cross-validation, which basically mimics test data using subsets of the data available. This produces the tree with the smallest estimated error. The pruned tree generally works much better than the unpruned tree if you wish to make inferences.

3.0 Northeast Kingdom Camp Owner Study

The data used below were obtained from the Northeast Kingdom Camp Owner Survey. The purpose of the study was to examine the underlying structure of place attachment among camp owners in rural northeastern Vermont, and to relate sub-dimensions of place attachment to a series of demographic, activity, place attribute and social interaction variables. The data were collected through interviews and mail surveys of a representative sample of camp owners.

A place attachment scale was used to measure the strength of attachment to the camp. This scale was based on scale items developed by Williams et al. (1992), Cuba and Hummond's (1993) work on Cape Cod, and Kaltenborn's (1997) examination of place attachment to recreational homes in Norway. The 23-item scale was formatted as a 5-point Likert-type scale ranging from strongly agree (1) to strongly disagree (5). In addition, a place attribute scale was developed that was thought of as encompassing both socio-cultural and bio-physical "magnets" that draw people to a place (Beckley 2003). Items from this scale were derived form both Kaltenborn's study and Beckley's paper, and were scored on a Likert-type scale of not at all important (1) to very important (5). Principal component analysis was used to reveal two sub-dimensions of both the place attachment and the place attribute scale. One of the sub-dimensions derived from the place attachment scale will be used as the dependent variable in the analysis below. The place attribute sub-dimensions are used as independent variables.

CART and multiple linear regression were used to examine the 23 demographic, activity and social interaction variables', and the place attribute sub-dimensions' relationships to one of the place attachment sub-dimensions.

A two-factor solution was chosen for the place attachment scale based on an examination of the scree plot and interpretability of the rotated factor solution. The two factors accounted for 49 percent of the variance. The first factor reflected attachment on an emotive-symbolic level similar to that of Schreyer, Jacob and White's (1981) emotional/symbolic sub-dimension and Williams et al.'s (1992) place identity. The second factor loaded on historical and geographical area statements and was labeled as a location-historic dimension, a dimension that approaches the functional dimensions of Schreyer, Jacob and White, and overlaps with the area and history dimensions of Kaltenborn's study (see Table 1). This latter location-historic dimension of place attachment was used as the dependent variable in the analysis presented here.

3.1 CART Analysis

Figure 1 shows the pruned regression tree built by CART using the 23 demographic, activity and social interaction variables', and the place attribute sub-dimensions to predict the value of the subject's score on the location-historic index of place attachment. Each node is numbered, with terminal nodes numbered separately from the splitting nodes. Listed below each node number is the variable that the node was split on. If the description in the node is true of a given case (for example, in the root node, if YRSOWN—the number of years the respondent has owned or leased the camp—is less than or equal to 27.5), the case goes to the left child node. Listed next in each node is the standard deviation and the mean of the dependent variable for the cases in the node. The growing homogeneity is apparent from the decreasing standard deviations as the tree descends. The mean in the terminal nodes represents the predicted value on the location-historic index for the cases in each node.

Figure 2 is a scatter plot of YRSOWN and Q40NATEX (the index of the place attribute sub-dimension "Nature," which includes the importance of the natural environment, wildlife viewing opportunities, serenity and peacefulness of one's camp's location). The root node splits on YRSOWN at 27.5 years. This split is represented in Figure 2 by the vertical line drawn at 27.5 on the horizontal axis. The subjects to the left of the line are in the left child node, node 2, in Figure 1, and the subjects to the right of the line are in the right child node, node 4. In node 2 in Figure 1, the cases are again split, this time on Q40NATEX at 22.5 (the possible range is 5 to 25). This split is represented in the scatter plot in Figure 2 by the line drawn at 22.5 on the vertical axis. Note that this line terminates at the line drawn previously representing the root node split. Only the cases in the left section of the scatter plot are split. The cases below the new split go to node 2's left child node, terminal node 1, and the cases above the split go to node 3. Node 4 in Figure 1, the right child node of the root node, also splits on Q40NATEX, but at 20.5. This split is also represented in Figure 2 by the line drawn at 20.5 on the vertical axis, this time to the right of the root node split. The cases above this split go to terminal node 5 in Figure 1, and the cases below go to terminal node 4. The final split in the regression tree in Figure 1 cannot be shown on the scatter plot because it is in a new variable, but it would split the cases in the upper left section of the scatter plot on the variable concerning how well the respondent knows his or her neighbor.

The r-square for the regression tree can be examined two ways. One way is the resubstitution r-square, and this is most similar to the r-square we look at for a multiple regression model. It is calculated by looking at how well the original data fits the pruned tree. The other is the cross-validated r-square. This is calculated from the cross-validation procedure that is used to select the optimal tree. It gives an indication of how well the regression tree would do if it were used with an entirely separate set of data. The resubstitution and cross-validation r-squares for the tree in Figure 1 are 0.274 and 0.215, respectively.

Using the regression tree as a stand-alone tool for analysis can be done at this point. Table 2 gives the statistics for the terminal nodes for the regression tree in Figure 1. They have been ordered in ascending value based on the mean value of the dependent variable within each terminal node. The result of the tree which led to these terminal

Table 1.—Factor Loadings.

Experiences	Factor I Emotive-symbolic	Factor II Location-historic
Going to my camp is one of the most pleasant things I can think of	.71	.08
Camp is a refuge when the rest of my life becomes too stressed	.71	.05
I get more pleasure out of doing the things I do at camp, than doing the same things anywhere else	.76	.11
Much of my life centers on the immediate location of my camp	.51	.38
Doing what I do at camp is more important than doing the same things anywhere else	.66	.33
Camp is my favorite place in my time off	.78	.20
I would prefer spending more time at camp, if I could	.64	.10
When I am at camp I can really be myself	.62	.24
Being at camp provides more satisfaction than visiting other places	.60	.36
Camp makes me feel like no other place can	.64	.35
Camp is very special to me	.74	.22
My family has a long lasting attachment to the general area where my camp is located	.08	.67
I identify more strongly with my camp than my home	.49	.51
My camp is significant to me because of its history and tradition	.16	.71
I identify strongly with the general area of my camp	.36	.62
Memories of camp are more vivid than those of home	.41	.57
No other places can be compared with the general area where my camp is located	.34	.62
The general area where my camp is located means more to me than my actual camp	.01	.54
Cronbach Alpha	.90	.81

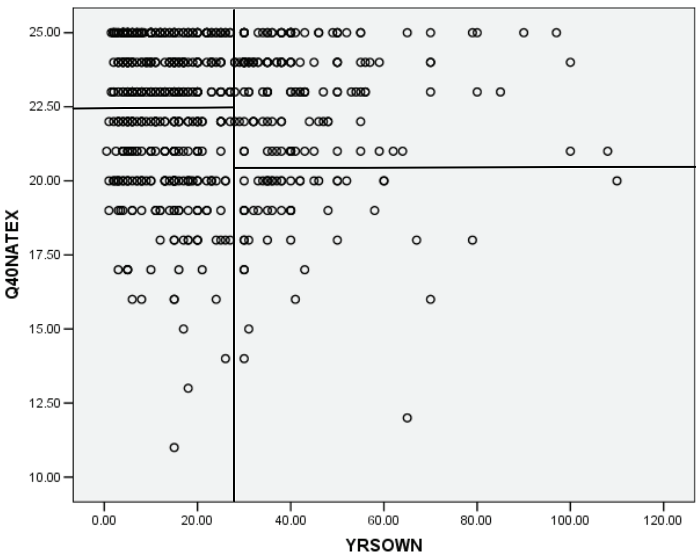

Figure 2.—Scatter plot showing splits from regression tree.

nodes is the creation of splitting rules. The splitting rules leading to each terminal node are shown in the right column of Table 2. The subjects with the lowest estimated scores on the location-historic sub-dimension of place attachment, those in terminal node 1, had owned or leased their camps for less than 28 years and did not have extremely high scores on the nature dimension of place attribute (at least a 23 out of a possible 25). Two of the nodes show groups that had high scores on the location-historical sub-dimension of place attachment. Node 5 is made up of people who had owned or leased their camps for at least 28 years and who scored a 21 or higher on the nature dimension of place attribute. Node 2 is made up of people who owned or leased their camps for less than 28 years, but who had extremely high scores on the nature sub-dimension. They also responded that they knew their neighbor "very well," which is the highest level in a 5-point Likert-type scale. These results show the importance of the nature sub-dimension index of place attribute and the number of years and individual owned or leased their camp. The importance of how well someone knows his or her neighbors to the location-historic index of place attachment comes out particularly among those respondents who were relatively short-term owners of their camps and who scored high on the nature dimension of place attribute. This analysis using CART as a stand-alone tool uses only three variables to split the data and developed an easily interpreted model with a resubstitution r-square of 0.274. Subsequently, we approach the same problem with linear regression.

3.2 Multiple Linear Regression

One of the uses of CART is to select from a long list of potential independent variables to be entered into a regression model. In this section, a model derived from CART will be compared to two models developed using common variable selection techniques—backward regression and stepwise regression (see Table 3). Variable selection for the CART model is based on the CART output, which tells what variables were the most important in creating the splitting rules for the regression tree (Table 4 is an abbreviated list showing variable importance in the CART model). Based on the r-squares in Table 3, the stepwise and backward models did better than the CART-derived regression model, but not by much, especially considering the CART model used fewer variables. The more linearly related the data is, the less well CART will do as compared with traditional linear techniques. CART will do better with non-linear data, though.

Table 2.—Terminal node statistics.

Terminal Node	Count	Mean	SD	Splitting rules
1	185	20.508	4.407	YRSOWN < 28, Q40NATEX < 23
3	186	23.038	4.783	YRSOWN < 28, Q40NATEX > 22, KNOWNEIGH ≠VERY WELL
4	49	23.041	3.264	YRSOWN > 27, Q40NATEX < 21
5	142	27.225	3.955	YRSOWN > 27, Q40NATEX > 20
2	53	27.302	4.681	YRSOWN < 28, Q40NATEX > 22, KNOWNEIGH = VERY WELL

Note: The values in the right column have been adjusted to represent values in the actual data set. For example, YRSOWN splits on 27.5 in the regression tree, but no case has this value, so those who were ≤ 27.5 are actually < 28. KNOWNEIGH = 1 means the respondent knows his neighbor very well.

Table 3.—Multiple regression models estimating score on location-historic index.

	STEPWISE	BACKWARD	CART
Q40NATEX	0.666**	0.654**	0.642**
Q40HARVSX	0.266**	0.262**	0.253**
KNOW NEIGH	-0.485**	-0.427**	-1.116**
YEARS OWN	0.080**	0.076**	0.087**
COMPARE	-1.077**	-1.004**	—
PARNT OWN	-1.272**	-0.793*	—
NATUR STDY	-1.250**	-1.081**	—
XC SKI	—	-1.013**	—
STAY CAMP	—	-0.352*	—
OCCUPY	—	0.006*	—
AGE	—	—	-0.021
TOGETHER	—	—	-0.057
CONSTANT	6.174**	6.562**	6.286**
	ADJ. R2=.389	ADJ. R2=.382	ADJ. R2=.359

*p<.10,
**p<.05

Table 4.—Variable importance.

Q40NATEX	100.000
YRSOWN	96.990
Q40HARVSX	
KNOWNEIG	39.169
TOGETHER	31.210
AGE	23.071

The numbers represent an index of importance, with the most important variable set at 100.

The last use of CART to be illustrated here is to divide the original data set into separate categories based on the splits from the regression tree and use each group to estimate a regression model. Three separate models are shown in Table 5, and selection rules for each model are shown in the column headers. If the regression models based on the entire data set are an accurate description of the data, then we should see similar results for each of these three groups, but we do not. In fact, as can be seen in Table 5, there is quite a bit of difference between the models built for each of the three groups in terms of the variables selected for each model. Furthermore, the amount of variance explained for each group was quite different. Although the middle model did very well, with an r-square of 0.38, the other two groups were relatively weak, with r-squares under 0.19. This shows us that the models built for the entire data set are a bit deceptive, because they tell us that the change in the dependent variables due to the changes in the independent variables is the same for all the data. We see here that this is not the case. In fact, we see that for certain sets of the data, there are unique relationships between the dependent and independent variables. The one exception is in the case of the variable Q40HARVSX, one of the sub-dimensions of the place attribute scale, which is important for all three subgroups in Table 5 and has a similar coefficient in each case.

4.0 Conclusion

CART is not a replacement for traditional methods, but a tool that can be used to compliment and extend the usefulness of traditional methods. CART excels when the relationships between variables are non-linear, while traditional methods typically perform better with linear relationships. As the last example shows, however, even when the relationships are linear, they may not be the same across all of the data, and traditional regression techniques are unable to detect this. CART helps determine what the important subgroups of the data are. Classification and Regression Trees are a valuable addition to the statistical toolbox of every social scientist.

Table 5.—Multiple regression models for subsets based on CART splits.

	YRSOWN < 28, Q40NATEX < 23	YRSOWN > 27	YRSOWN < 28, Q40NATEX > 22
Q40NATEX	—	0.686**	—
Q40HARVSX	0.234**	0.243**	0.273**
KNOW NEIGH	—	—	-0.532
YEARS OWN	—	0.051**	—
COMPARE	—	-1.108**	-1.200**
PARNT OWN	—	—	-2.335**
NATUR STDY	—	-1.560**	—
XC SKI	—	-0.978	—
STAY CAMP	-0.817**	—	—
OCCUPY	0.010**	0.014**	—
WILDLIFE OBS.	—	—	-2.259**
BERRY/MUSHROOM	1.368**	—	—
TOGETHER	-0.460**	—	—
CONSTANT	19.022**	5.443*	25.701**
	ADJ. R2=.170	ADJ. R2=.380	ADJ. R2=.189

*p<.10,
**p<.05

5.0 Citations

Arminger, G., Enache, D., and Bonne, T. 1997. **Analyzing credit risk data: a comparison of logistic discrimination, classification tree analysis, and feedforward networks.** Computational Statistics. 12(2): 293-310.

Beckley, T.M. 2003. **The relative importance of sociocultural and ecological factors in attachment to place.** In Understanding Community-Forest Relations, USDA Forest Service Gen. Tech. Rep. PNW-GTR-566: 105-126.

Cuba, L., and Hummon, D.M. 1993. **Constructing a sense of home: Place affiliation and migration across the life cycle.** Sociological Forum. 8(4): 547-572.

Kaltenborn, B.P. 1997. **Nature of place attachment: A study among recreation homeowners in southern Norway.** Leisure Sciences. 19: 175-189.

Lewis, R.J. 2000. **An introduction to classification and regression tree analysis.** Paper presented at the meeting of the Society for Academic Emergency Medicine in San Francisco. Retrieved February 2005, from http://www.saem.org/download/lewis1.pdf

Long, W.J., Griffith, J.L., Selker, H.P., and D'Agostino, R.B. 1993. **A comparison of logistic regression to decision-tree induction in a medical domain.** Computers in Biomedical Research. 26; 74-97.

Rejwan, C., Collins, N.C., Brunner, J., Shuter, B.J., and Ridgway, M.S. 1999. **Tree regression analysis on the nesting habitat of smallmouth bass.** Ecology. 80(1): 341-348.

Schreyer, R., Jacob, G., and R. White 1981. **Environmental meaning as a determinant of spatial behavior in recreation.** Proceedings of the Applied Geography Conferences. 22: 294-300.

Tronstac, R., and Gum, R. 1994. **Cow culling decisions adapted for management with CART.** American Journal of Agricultural Economics. 76: 237-249.

Williams, Daniel R., Patterson, Michael E., Roggenbuck, Joseph W. and Alan E. Watson. 1992. **Beyond the commodity metaphor: Examining emotional and symbolic attachment to place.** Leisure Sciences. 14: 29-46.

Zhang, H., Yu, C., Singer, B., and Xiong, M. 2001. **Recursive partitioning for tumor classification with gene expression microarray data.** Proceedings of the National Academy of Sciences. 98(12): 6730-6735.

Citation:
In: Peden, John G.; Schuster, Rudy M., comps., eds. Proceedings of the 2005 northeastern recreation research symposium; 2005 April 10-12; Bolton Landing, NY. Gen. Tech. Rep. NE-341. Newtown Square, PA: U.S. Forest Service, Northeastern Research Station