Background on Classification Tree Analysis:

Classification tree analysis comprises a set of model-free methods for analyzing multivariate data (Fisher and Lenz, 1996; Biggs et al., 1991; Cox, 1989) and "mining" large databases for useful knowledge (Elder and Pregibon, 1996). Classification tree algorithms search for combinations of values of independent variables that best predict the value of the dependent variable. Ability to predict is measured by criteria such as the entropy, variance, or statistical significance of the conditional frequency distribution of values for the dependent variable, conditioned on the answers to questions asked. Questions are represented by partitioning the possible values of variables into subsets and asking which subset the value for a particular individual belongs to. Based on the answer, a new question is asked. The result is a classification tree, with nodes representing questions and branches at each node representing possible answers. Internal nodes are also called splits, while leaf nodes ("tips" of the tree) represent probabilistic classifications or predictions of the value of the dependent variable. The tree stops growing when no additional questions will improve the ability to predict the value of the dependent variable, as measured by the selected criterion. Alternatively, trees may be grown larger than needed and then pruned back until the estimated error rate is minimized (Breiman et al., 1984). Different splitting criteria and pruning or stopping criteria lead to different specific classification tree algorithms. All the classification trees discussed in this paper were created using the specific algorithm of Biggs et al. (1991), which uses estimated statistical significance (based on F- and chi square statistics for continuous and categorical variables, respectively), with a Bonferroni adjustment to correct for multiple comparisons in choosing class boundaries, for its splitting and stopping criteria. (See Biggs et al., 1991 for details and Monte-Carlo validation of the performance of this specific algorithm.) Each tip of the tree, corresponding to a unique branch or path through it, represents a sequence of questions and answers. The conditional frequency distribution of the value of the dependent variable, based on the questions and answers leading to the tip, constitutes the (probabilistic) prediction from the classification tree at that tip.

Several features make classification trees particularly well suited to avoid many of the threats to valid causal inference identified in Table 1. Specifically:

- Multiple hypothesis testing and multiple comparisons bias can be avoided by incorporating Bonferroni estimates of true p-values directly into node- splitting criteria (Biggs et al., 1991) and by pruning back large trees to reduce false positives to the desired global significance levels. Several commercial classification tree algorithms allow the use of hold-out samples and/or cross-validation techniques to provide data-driven estimates of the true p-values achieved (Breiman et al., 1984).
- Model specification errors may be reduced or eliminated because classification trees do not require or assume any specific parametric form for the relation between independent and dependent variables. Thus, they provide"model-free" approximations to the multivariate response surface.
- Aggregation errors are reduced by treating individuals rather than aggregate populations or groups as the units of analysis.

Some imperfections due to quantization of variables remain, however. For example, classification trees approximate the true multi-factor response surface for cancer by, in effect, piecewise-constant multivariate histograms. Other classification tree algorithms yield piecewise-linear approximations of the response surface (Breiman et al., 1984), while the newer Multiple Adaptive Regression Splines (MARS) technique provides smooth approximations (Elder and Pregibon, 1996). Such approximations introduce a potential source of error compared to the true but unknown response surface, but the errors are often small (since ones that are large enough to be statistically significant lead to additional branches) and are inherently local, in contrast to the global errors introduced by wrong model forms or other specification errors in parametric statistical models.

From Classification Trees To Causal Graphs

Formally, a causal graph is defined as a directed acyclic graph (DAG) in which nodes represent variables and each node is conditionally independent (CI) of its ancestors, given the values of its parents, i.e., its immediate predecessors in the directed graph (Jensen, 1996). (An ancestor of a node is here defined recursively as a parent of a parent or a parent of an ancestor.) Such DAGs are among the most widely used structures for representing causal knowledge (e.g., Fisher and Lenz, 1996) and for "mining" useful patterns from large data bases (Fayyad et al., 1996). Their precise meaning and interpretations have been placed on a useful philosophical foundation by Shafer (1996). They may be used not only to represent CI relations among subsets of variables, but also to indicate that changing the levels of parent variables will change the conditional probability distributions for the levels of their children (Pearl, 1996). These and other interpretations and the relations among them are more fully developed by Shafer (1996).

Classification tree algorithms can be used to test the CI relations in DAG models by testing whether the hypothesized parents of each node (if any) form a minimal sufficient subset with respect to its ancestors for predicting its value. The following two-phase tree-growing algorithm accomplishes this test:

Algorithm A: Two-Phase Tree Growing Procedure for Testing Node Parents

For each variable (node) having incoming arrows in the hypothesized DAG: 1. Grow a classification tree with the selected node as the dependent variable and with only the hypothesized parents of the node allowed as potential independent variables (possible splits). Call the resulting tree T1. 2. Expand the set of allowed independent variables to include all ancestors of the selected node. Then, resume the tree-growing algorithm, starting at the leaf nodes in the tree generated in step 1. Call the resulting tree T2. 3. If none of the ancestors introduced in Step 2 enters the tree (i.e., T2 = T1), then the hypothesized DAG structure specifying the parents of the node is confirmed; otherwise, it is rejected. Repeat until all nodes with incoming arrows have been tested.

Since the binary relation "is a parent of" between nodes suffices to determine the entire DAG structure, Algorithm A provides a useful tool for helping to determine DAG models from multivariate data. Conversely, the insights derived from multiple classification trees can often be combined and succinctly summarized in a single causal graph. Although any single classification tree is limited to a single dependent variable, causal graphs provide a natural way to show multiple dependent variables and the causal relations (in terms of shared parents or ancestors) among them.

To learn more, read our page on A DATA-MINING APPROACH TO FORECASTING TELECOMMUNICATIONS DEMANDS AND NETWORK LOADS or download Dr. Cox's paper on LEARNING IMPROVED DEMAND FORECASTS FOR TELECOMMUNICATIONS PRODUCTS FROM CROSS-SECTIONAL DATA