

StatSoft®
Business White Paper

Categorical Predictors in *STATISTICA Data Miner* Regression Trees

Last Update: October 19, 2002

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Czech Republic: StatSoft Czech Rep. s.r.o.
France: StatSoft France

Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl

Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Poland: StatSoft Polska Sp. z o. o.

Portugal: StatSoft Iberica Ltda.
Russia: StatSoft Russia
Singapore: StatSoft Singapore
S. Africa: StatSoft S. Africa (Pty) Ltd.

Spain: StatSoft Espana
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Issue

Categorical predictors in tree models, when those predictors contain a large number of classes (e.g., over 3000).

Response

There are two important points that we would like to make.

First, in general, *STATISTICA* can usually handle in almost all analyses much larger sizes of designs than any of the competing data mining packages, so this difference in limitations you reported (regarding the number of categories in categorical predictors in regression trees) was initially a surprise to us. When we examined the issue more closely (see the section below on "*Categorical predictors with many categories*"), it appeared that we had set this limit (to 150) as a "more than a reasonable" limit with regard to various efficiency and estimation issues (these methods are generally not suitable to analyze data with that many categories, see further detailed below). However, if indeed this issue (the ability to include categorical predictors with thousands of categories in your regression and classification trees) is important for your particular work, then we can easily adjust these limits and provide an updated version of the program to you very quickly.

Also, I hope that you have had the opportunity to work with the various features of *STATISTICA Data Miner* (which - to the best of our knowledge - offers the most comprehensive selection of tree methods available on the market), and in particular the methods for building regression and/or classification trees; these are generally much more in-depth (and "careful") implementations of these techniques as compared to those offered in *Insightful Miner* (which has no CHAID or exhaustiveCHAID methods, no surrogate splits for dealing with missing data that we could find, no interactive tree building methods, etc.); we are also in the process of Beta testing another module for boosted trees (also referred to as treenets, forests, or stochastic gradient boosting), a method that has been developed over the past two years, and which appears to emerge as one of the most powerful general methods for predictive data mining.

Please let us know how your evaluation of the various other features of *STATISTICA Data Miner* is progressing as well; we are always anxious to learn from experience practitioners and practitioner-scientists how well we can address their needs (and, of course, how we may be able improve *STATISTICA Data Miner*).

PS: *Categorical predictors with many categories*.

In general, the C&RT algorithm will find for each particular split the category or combinations of k categories for each categorical predictor that yields the best split (purest sub-nodes). An exhaustive search of all possible combinations of categories can require a substantial amount of computing time, and past 20 or so categories becomes impractical (which is why other designated tree programs, such as, for example, SPSS AnswerTree, will limit the number of categories for categorical predictors to about 25 categories or so). To accommodate more categories in categorical predictors requires the program to resort to non-exhaustive searches for splits, using iterative "search" algorithms (for a good split). *STATISTICA's Interactive Trees* module, in fact, uses that approach which we have carefully tested for well over 100 predictor categories. However (and this is a very important "however"), the resulting split cannot be guaranteed to be the (global, i.e., over all possible splits) best split! In fact, in our tests we have found that, as the number of predictor categories increases, so will the chances of converging on a local optimum for each split (i.e., a "good" split, but not necessarily the best split).

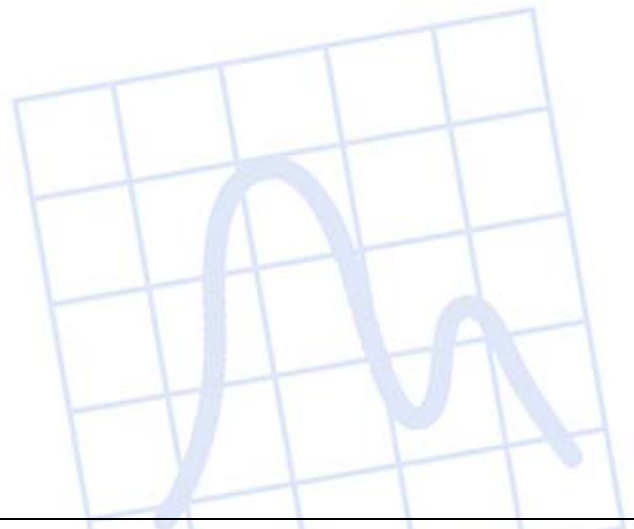
In practice, this means that the specific results may not be unique, and not invariant to, for example, the order in which data appear in the data file. So we made the decision to limit the number of categories for categorical predictors to 150, as a compromise to allow large numbers of categories, and yet to be reasonably sure that the results are stable (i.e., represent a true best tree).

There are also other issues to consider: To illustrate, suppose you performed an analysis with only 1 categorical predictor with 2000 or more categories. At that point, you can potentially at each split select from among billions of possible combinations of categories! Put another way, you are selecting one out of billions of possible predictors (considering that each combination of categories in the predictors is a unique "predictor" itself). Even with large datasets, it is not very likely that the resulting tree from such an analysis will have useful predictive validity. This is, in our experience, indeed a serious general problem when using categorical predictors with many categories for building predictive models.

As a practical matter, we would usually recommend that you recode your data into fewer meaningful categories for your analyses, which are then more likely to yield meaningful stable results. In fact, these types of data cleaning operations (applying domain-specific knowledge to, for example, apply meaningful recoding and transformations) usually requires the most time in any data mining project.

In general, the issue of "overlearning," in particular with machine learning algorithms, is a serious problem in predictive data mining, and the blackbox approach to data mining can sometimes yield promising results (with the "training sample") that prove not very useful for predicting new observations.

p.p.s. Just a short note to let you know that we have updated the limits for the maximum number of categories for categorical predictors in classification and regression trees (as well as CHAID) to around 10,000 (for now). To reiterate, exhaustive searches of all possible splits, or even simplified gridsearches for potential splits, are no longer efficient with so many predictor categories; instead the program will use an iterative search algorithm to find an optimal split (which cannot be guaranteed to be the optimal split). There is also the issue of predictive validity I described in my earlier message, and it is probably best to be very cautious when interpreting (or applying) the results of such analyses (and we would not, generally, recommend analyses involving categorical predictors with very many categories).



U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Czech Republic: StatSoft Czech Rep. s.r.o.
France: StatSoft France

Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl

Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Poland: StatSoft Polska Sp. z o. o.

Portugal: StatSoft Iberica Ltda.
Russia: StatSoft Russia
Singapore: StatSoft Singapore
S. Africa: StatSoft S. Africa (Pty) Ltd.

Spain: StatSoft Espana
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.