Summary 63

VITICULTURE & ENOLOGY
UNIVERSITY OF CALIFORNIA DAVIS

Use of decision tree analysis for determination of critical enological and viticultural processing parameters in historical databases

By: V. Subramanian, K. Buck, and D. Block

In: Am. J. Enol. Vitic. 52(3)175-184. 2001

Funded by: American Vineyard Foundation, California Competitive Grants Program for Research in Viticulture and Enology, and the University of California.

The authors show how a data management technique can be used to determine the most relevant viticulture and winemaking parameters affecting quality. This is a rather technical paper, but the outcome is quite useful.

• As our ability to collect large amounts of data increases, it is becoming critical to take advantage of current methods of information management. By developing computer methods able to analyze large databases, it should be possible to use *past* information to optimize *future* grape-growing and winemaking practices. That means we should be able to identify and repeat the factors that lead to quality, while avoiding the factors that hurt quality.

• There are many variables that are measured during the production of wine that may not affect the final product quality. Thus, the first step is to do some “database mining” to come up with what the authors call the “working database”. There are certain computer programs that help us do that. Then, the following decisive step is to “train the program”. This involves finding out the role –or weight- that each individual input plays in the final outcome. This is, once again, something the compute does for us. After the “database mining” and the “training”, the model is ready to predict future outcomes based on historical inputs.

• Winemaking databases can include both “*continuous*” (for example, temperature, Brix) and “*categorical*” variables, which are the ones that cannot be expressed by a number (for example, vineyard block, clone, type of barrel). Some commercial statistical techniques have the shortcoming of not being able to handle “categorical” variables. But the technique called Decision Tree Analysis (DTA) is able to quantify the flow of information when both “categorical” and “continuous” variables are present.

• But what the heck is a Decision Tree Analysis? A DTA is an *algorithm* that determines which variable(s) are most important to a given outcome or result. An algorithm is any mathematical process that involves a “series of steps for doing something”. We can think of an algorithm as a “cooking recipe”: add this, add that, and this dish will come out.

• The way a DTA works is by **finding which input variable best classifies the data** at each step -or branch- of the tree. Each step of the way, the program finds which additional variables improve the classification of the remaining data, and this “branching” goes on until no further classification is possible. The way DTA monitors that information gets progresively “classified” is by tracking the parameter called *information content*, which gets smaller and smaller with each step. The system stops seeking further classification when it arrives at classifying at least 80% of the data points in a given category (when *information content* comes close to zero).

• Luckily, the authors offer an **excellent example that is crucial to understanding the content of this paper**. I would like to ask for your special attention since, if you understand this example, you will have understood the author's message and what an algorithm is all about. Let's imagine that all fermentations with a low initial Brix were "good" and all fermentations with a high initial Brix were "bad". Then the input "initial Brix" will completely classify the data. This is the type of input that DTA helps us identify. If, on the other hand, half of the low initial Brix fermentations were "good" and half "bad", then the input "initial Brix" does not help classify the database at all, and DTA needs to continue searching for another input that does, until it comes up with a hierarchical set of inputs, or parameters, that would give future fermentations the most likelihood of being "good".

• The goals of this paper were: 1) to develop a decision tree analysis (DTA) to identify the critical inputs in an existing database, 2) to verify that the DTA works properly by using an existing enological database, and finally, 3) to use the DTA to identify critical viticultural parameters in a commercial vineyard database.

• **Verification of the DTA**. To verify that the computer DTA could identify critical parameters in an enological database, the authors applied it to a Sauvignon blanc database in which tartaric acid was measured as a function of initial Brix, skin contact time, and fermentation temperature, data for which there was a parallel statistical analysis available. The resulting decision tree classified the variables studied in decreasing order of importance in causing low or high acidity in the wines. The most important input was found to be maturity level (Brix), followed by fermentation temperature, and lastly, skin contact time. Most importantly, **the critical variables pointed out by the DTA analysis correlated well with those from the statistical analysis**.

• **Behavior of the DTA in the presence of random inputs**. To make sure that the DTA program only classified data when the inputs made sense, the authors studied how the DTA would behave if they were to shuffle all the inputs and outputs. As expected, the **DTA was not able to classify the above Sauvignon blanc database when the inputs had been changed randomly**. That is, the DTA was unable to find correlations, meaning none of the inputs were significant in causing high or low tartaric acid.

• **Use of the DTA with continuous variables**. Next the authors studied how the DTA behaved with "continuous" variables. (In this case, the program tends to choose a value for which the data is best classified (say, 22 ^{o}Brix) and splits that branch into 2 groups at that point ("Brix< 22" and "Brix>22"). The continuous database the authors used consisted of the quality scores for the last 22 vintages of Napa Valley Cabernet Sauvignon (average of *Wine Spectator* and *Wine Advocate* scores) as a function of the temperatures registered during 9 defined growing season periods. The goal was to determine which periods had the most effect on vine growth and fruit development leading to the highest-quality Napa Valley Cabernet. The DTA pointed that **Period 9 (just prior to harvest) had the most relevance in determining the highest scores**. The next branch of the tree showed the infrequent situation of a "tie" among three entries (periods 2, 4, and 7). The authors also analyzed the same data statistically, and found that Periods 9 (just prior to harvest) and 2 (just after budbreak) were the most significant. So there was a good overall agreement between the DTA and the statistical analysis.

• **Use of DTA to identify viticultural variables important for wine quality**. The next logical step was to test the DTA in a real, production situation. This database contained 4 years of information about multiple parameters used in the farming of 3 Pinot noir vineyards (ranch, clone, rootstock, vine spacing, trellis system, phonological information, and pesticide applications). The output was the quality scores of the resulting wines (low, medium, and high price point), as judged by the collaborating winemaking team. The goal was to find the viticultural practices most responsible for determining wine quality. The DTA results showed that the parameter "clone" was the one that best determined quality in this instance. Other important parameters included: year planted, pruning weight, ratio fruit/pruning weights, mite spraying, irrigation, and length of time between veraison and harvest.

In summary, these studies proved the efficacy of a computer-generated program, called a decision tree algorithm, to manage databases to understand the relationships between inputs and outputs. This method worked well with a categorical database (Sauvignon blanc tartaric acid levels), a continuous database (Napa Valley Cabernet Sauvignon scores), and a commercial winery database (Pinot noir wine price points based on viticultural parameters). Based on the successful results, Dr. Block and his team believe it should now be possible to apply this type of analysis to optimize a program capable of extracting the most desirable set of grapegrowing and winemaking conditions from existing historical databases. Even though this paper is rather technical, and the techniques used have names unfamiliar to most of us, the potential for extracting extremely useful information from what used to be a meaningless “bunch of numbers” seems invaluable.

Author: Bibiana Guerra, Editor: Kay Bogart. This summary series funded by J. Lohr Vineyards & Wines.