Sorting out bad data; data mining with recursive partitioning

Jeffrey M. Yarus[1] (1) Quantitative Geosciences, LLP, Houston, TX

With the vast amounts of information available on historically drilled wells, finding the “right” data can be like looking for a needle in haystack. In particular the task is daunting when recorded data are erroneous, yet not obviously problematic. Recursive partitioning (RP) is a multivariate statistical technique that can help sort out good data from bad particularly when bad data takes the form of misinformation.

RP is a method that was developed in the 1980’s; it is often referred to by the acronym CART, coming from the title of the first book on the subject, Classification and Regression Trees (Friedman, et al., 1983, 1993). RP begins by creating a “tree” of questions that classify (or partition) a testing data set into several subsets. Once this tree is created, it can be used for prediction by taking a new observation, classifying it to the same set of questions, and using the behavior of the samples in the same node to predict the behavior of the new observations. When used in this way, RP can be seen as a method that extracts from the training data set the “closest cousins” - those most similar to the new observations.

In the case presented, six oil producing wells are believed to be underachievers with respect to other nearby wells with similar reservoir characteristics. RP is used to uncover the possible underlying reasons which potentially can be due to unforeseen poor geological or reservoir conditions, or misreported data.