Abstract: Application of Numerical Taxonomy to Systematically Research and Verify Performance Based Handicaps

A Presentation at the Fall 2003 Meeting of the ISAF Empirical Handicap Subcommittee

Paul J. Ansfield

Taxonomy is most often thought of as the effort of biologists to construct an acceptable and ordered classification of plants and animals. More broadly, taxonomy is the body of knowledge and methods to classify naturally occurring systems.

Early attempts to apply numerical methods date back to the rise of biometrics in the last century to find a systematic way through measurement to support classification within biological systems. With advancing statistical methods and the evolution of computing technology a contemporary methodology developed that succeeds. Pivotal work on these numerical classification methods was that of Robert R. Sokal in the United States and Peter Sneath in England and was described in their 1963 book: *Principles of Numerical Taxonomy (1973; Freeman).* The techniques have further evolved, and are robust. Often called cluster analysis, these techniques do not require rigid statistical assumptions to be met with exactitude as do factor analysis and discriminate function analysis. Examples are block clustering, tree clustering, and hierarchical clustering. These methods provide ways of grouping similar profiles of measures collected on multiple cases that are examined for the purpose of classification.

When a collection measured **profiles** is **comprised of the critical hull, rig and sail plan dimensions** for wind driven boats that race, numerical clustering may lead to statistical confirmation of various classes or types, validation of existing empirically derived handicaps, and the discovery of new classes or types derived statistically. The overall purpose of clustering critical dimensions of sailboats is to increase the precision of systematically assigned empirical handicaps as a predictor of potential boat speed for scoring competitive events.

While statistical clustering methods are sound, there are **problems that arise** in their application to classify sailboats that race. These problems are within the nature of the data set used for performance handicapping. Often measures for handicap assignment are self reported and carry an inherent reduction in precision. A greater problem lies with the selection of the critical dimensions from the data set for creating clusters of boat types or classes. A sample of dimensions tied to manufactures model that might be used include hull parameters, rig type, sail plan variables, and calculated ratios such as sail area to displacement and length to displacement, etc. The number of dimensions that may contribute to each boat model's profile used in the clustering procedure can be quite extensive. The immediate heuristic task involves selecting those few that may efficiently form clusters for examination.

In **one exercise** I, ISP, J, JC, P, Y, SPL, SL, SMG, displacement and maximum draft dimensions for 250 different masthead and fractional sloops were utilized. Other dimensions could have been selected. Clusters were formed by the statistical process. The most striking finding was the inclusion within the clusters of boats that did not experientially seem to fit. Whether this was an artifact of the method, selection of critical dimensions or measurement error needs determination. Perhaps these outliers are representative of those boat types that trouble the empirical assignment of initial handicaps. They are thus marked for evaluation of their assigned handicap.

Summary: An evaluation of the initial effort indicates that hierarchical clustering of sailboat critical dimensions to verify similar groups of boats and assigned empirical handicaps shows promise. Criteria for selection of critical hull, rig, and sail plan dimensions for clustering need further development. The precision of self-reported measures needs to be addressed and dealt with to refine the method for statistical classification of sailboats that race.