Old and New Applications of Multivariate Statistics to Problems in QSAR

Gerald J. Niemi[1] **and Jarad B. Niemi**[2]

[1] *Natural Resources Research Institute and Department of Biology, University of Minnesota, Duluth, MN, USA.*

[2] *Department of Statistical Science, Duke University, Durham, NC, USA.*

Abstract

The analyses of quantitative structure -activity relationships or quantitative structure-property relationships are inherently multivariate statistical problems. Chemicals can be quantified in many ways using mathematical techniques which results in hundreds of variables that can be calculated for every chemical. Similarly, each chemical, whether it is naturally or synthetically produced, has physical, chemical or biological activity potential. These include biological activity such as toxicity, biodegradability, or carcinogenicity or chemical properties such as melting point, boiling point, vapor pressure, or partition coefficients. Among the goals of structure-activity relationships is to relate variables that describe structure (independent variables) with activity variables (dependent variables). Obvious statistical procedures for examining relationships between structure and activity include correlation or, when predictions of behavior are desirable, regression can be used. Unfortunately, the sheer numbers of variables often render standard statistical approaches useless or result in severe violations of statistical assumptions. Here we briefly review some of the multivariate statistical procedures used to explore problems in structure activity relationships and provide several examples of their use. In addition, we explore how Bayesian statistics could be used to further develop relationships in QSAR.