# Fit Model to Data, Not Data to Model (Newsletter)*

November 12, 2014

To make informed regulatory decisions it is necessary to understand differences between ecological and environmental data. Analyses of environmental data historically use models developed by numerical ecologists for ecological data collected by academic and research agency scientists. These numeric and statistic models require well-structured data collected to fit assumptions and requirements of the models. This works for researchers who identify a question to be answered and work forward from that to determine when, where, and how much data need collecting to answer that question. The research approach of fitting data to models has leaked into the analyses of environmental data gathered in response to statutory and regulatory requirements. Most often, the results are mis-leading or incorrect. Regulatory decisions based on these results are ineffective at best or economically and socially harmful at worst.

Environmental data are messy and unstructured, collected to support environmental permit applications and monitor compliance with permit conditions. Locations change over time, data collection frequency is irregular, and chemical or biological data elements can cease being collected and re-instated at a future time. Such data cannot be fit to research models such as species diversity, indices of biotic integrity (IBI) or community indices (CI), predictive models based on expected taxa (RIVPACS), hydroelectric fish passage models (CRiSP), or pit lake water quality (PITLAKQ). For real-world environmental regulatory decision-making it is necessary to fit the model to the data.

It is difficult (or impossible) to get reliable, consistent, generally applicable analytical results of environmental data from numeric models. Therefore, an appropriate statistical model is used. There is such a large choice of statistical models (the R project alone has over 6,000 application-specific model packages for analyzing data of every type) that one appropriate for regulatory decisions based on environmental data can be identified and used to produce technically sound and legally defensible results.

Not all environmental data analysts are aware of the broad selection of available statistical models, nor of the differences in what they measure. It is

not uncommon to read a report to regulators that compares sets of water chemistry data using analysis of variance (determining if they are similar because they come from the same population of water chemicals) when the regulator wants to know whether the permitted operation has an undesired negative effect on water quality; that is, why the measured concentrations have the values they do. Providing an answer that does not answer the regulator's question can have severe consequences for the permit holder.

In practice, one of the largest differences between analyzing ecological and environmental data is that many of the most appropriate models for the latter are relatively unknown or recently developed. Among these statistical models are those for quantile regression and compositional data analysis. Quantile regression measures the relationships of explanatory variables on different portions of the observed range of the response variable (not just the mean response variable as is the case with linear regression). Compositional data analysis analyzes parts of a whole; for example, some chemical constituents in a medium with many chemicals or functional feeding groups of benthic macroinvertebrates.

—

All newsletters, white papers, and other technical resources can be freely downloaded from `http://www.appl-ecosys.com/publications/`.