Collecting and Analyzing Data (Newsletter)*

December 20, 2010

Data on natural ecosystems are collected and analyzed to inform business and permitting decisions, and to monitor compliance with permit conditions. Too often decisions on what to measure or observe, where, and how frequently are made without consideration of analytic constraints. The results do not answer the questions asked; the money and time has been wasted and decisions are made with less certainty than desired by operators, project proponents, and regulators.

Natural world data come in two types: continuous and discrete. Continuous data such as temperature and chemical concentration in water or rocks have values that can be represented by mathematical distributions (such as the familiar Bell curve). Because these distributions represent all possible values in the entire population of interest we can calculate statistical representations of the population from samples we collect. These representations are known as parametric statistics because they are parameters of the entire (unmeasured) population of values. Discrete data such as numbers of animals and plants are not continuous; there is no fractional animal in the population. These data need to be analyzed using the family of non-parametric statistics which make different assumptions about the populations as wholes and remain equivalent to the familiar parametric statistics (such as mean, variance, standard deviation).

The most common and frequent mistake made in the interpretation of data is confusing correlations (that is, association of two or more things) with cause and effect. Associations are common with independent parameters. For example, nice weather during the work week and rain on the weekends is a correlation, an association. Weekends do not cause rain, even if you planned a picnic for Saturday. Cause and effect must be established by demonstrating that one parameter is directly dependent on the value of the second, independent parameter.

Temporal and spatial analyses give ecosystem data much greater value, an important factor for decision-makers. To predict what might be found in the near future we look at data changes in the past. The statistical tool used is time series analysis and it requires the intervals between values to be the same over the time span being analyzed. Changing data collection frequency violates an assumption required by the underlying mathematical model. While you can plot the points and see the resulting line you cannot determine if any slope is different from a horizontal line. Time series analyses reveal meaningful trends up or down, but only if they have consistent time intervals.

Spatial statistical analyses are especially valuable because they can explain observed behaviors and relationships. When you want to know why water quality differs in adjacent or nearby watersheds, why fish are in one stream but not in another that appears the same, why hydrology and sediment transport varies within and between river systems, and why trees are found in clumps rather than evenly distributed, spatial analyses can provide answers. Because all ecosystem processes have a strong spatial component the appropriate statistical analyses and models can reveal the underlying processes that produce the observed results. Situations where spatial analyses provide answers that numeric analyses cannot include wolf/cattle interactions, why certain properties flood in heavy rain storms, where to site facilities to avoid water quality degradation, and optimal paths for transmission or pipe lines. Beyond mapping, spatial analyses provide robust cause-and-effects explanations with results presented as maps so they are easily understood by non-technical audiences.