Applied Ecosystem Services, LLC
Integrity • Credibility • Innovation

2404 SW 22nd Street
Troutdale, OR 97060-1247
Voice: 503-667-4517
Fax: 503-667-8863
E-mail: info@appl-ecosys.com

Make environmental data a key part of your success

Introduction to the importance of correct environmental data analyses

Richard Shepard, PhD

November 14, 2023

Introduction

Environmental laws, statutes, and implementing regulations have two specific purposes: forecasting future conditions and determining cause and effect. The analytical models applied to existing environmental data sets do not provide results that explicitly address the two purposes. They leave decision-makers confused, uncertain how to justify a decision, and suffering paralysis by analysis as they ask for more results that prove to be ineffective.

This document explains how environmental data are analyzed (wrongly and correctly), why the analytical approaches differ, and why knowing the details makes you more effective in managing environmental data for your operations. legal case, or litigation.

Future conditions are defined by forecasting environmental conditions based on available data sets. These forecasts frequently are the basis for approval of, for example, environmental assessments and impact statements, storm water discharge permits, incidental take of threatened or endangered species of fish and wildlife, restoration of streams, rivers, fish, and wildlife populations, and whether Superfund restoration activities are on track to fulfill expectations.

Every appeal and challenge of a proposed (or existing) environmentally-permitted activity is based on the data presented to decision-makers. It's all about the data and how they are analyzed, interpreted, and communicated.

Becoming better educated about how environmental data affect your activities helps you avoid environmental issues, more quickly and effectively resolve those that arise, and succeed in litigation. These benefits are especially valuable in changing and unstable conditions.

While the subject is highly technical and based on complex mathematics, statistics, and spatial science there's no math, stats, or other eye-glazing details in this document. I explain the issues in plain English so you understand the what and the why. How to ensure your analytical results are technically sound and legally defensible is my job.

Environmental data differ from other types of data

Environmental data differ in many ways from business, financial, engineering, and public data such as unemployment rates and the cost of living. Environmental data need analytic models that accommodate those differences and that allow the analyst to fit the model to the data. When the applied models do not support these differences the results, and the decisions made based on those results, will not be correct. That means they are not technically sound and legally defensible. Success in you job increases when you understand these differences and your data's analytical results are technically sound and legally defensible.

As an aside, remember that permit holders must meet permit requirements. Remember that while water quality standards, all aquatic biota, and animals who drink the water, are exposed to many more chemical constituents than are measured for permit issuance and compliance and too often are the core of environmental litigation. All water constituent concentrations are constantly in flux, especially in flowing water ecosystems and do not adequately represent the ambient water quality at that location and time.

Geochemical data

Geochemistry is the collective term for chemical constituents of rocks, soils, sediments, and water. No geochemical ion or organic molecule can have a concentration less than zero. The statistical models most commonly used for these data assume that the frequencies of concentration values follows a Gaussian (i.e., normal) distribution, the familiar bell curve. The Gaussian distribution, by definition, has a mean value (the most common or expected one) of zero and a standard distribution of 1. Because chemical concentrations cannot be less than zero the results of analytical models based on the normal bell curve distribution of concentration frequencies yield incorrect results.

In geochemical data the constituent amounts of most concern are toxic metals and organic compounds. Toxin concentrations affect the health and viability of fish, wildlife, livestock, and humans so they are regulated under several federal laws; e.g., the Clean Water Act, the Toxic Substances Control Act, and the Federal Insecticide, Fungicide, and Rodenticide Act.

The concentration of these chemicals are usually very low, commonly less than can be detected and quantified by the analytical laboratory's equipment, methods, and analyst's skill. These non-detected data are also called censored data because their value, event their presence, is unknown. These censored data commonly constitute 5--80 percent of the whole data set.

What should you do with data that cannot be quantified with certainty nor even know whether they are present in the collected sample? There are four ways to mishandle these data to produce wrong results, and the EPA has mandated the use of each wrong method in various water quality programs.

The first way is to ignore these data. That they cannot be detected as present in the sample or measured quantitatively if they are present is very important environmental information. If there's no proof of toxins in the water that's good news for all consumers, aquatic biota who live in the water as well as wildlife, livestock, and humans who drink it. Don't ignore these data.

The second way is to set all non-detected values to zero. This distorts the value:frequency distribution and lowers the mean value. And, it's unrealistic for many data points to have the same value. Don't set all non-detected data to zero; you'll get the wrong results.

The third way is to set all censored data to the analytical laboratory's reporting threshold, the lowest concentration they can quantify with certainty. This distorts the value:frequency distribution and raises the mean value. And, as above, it's unrealistic to have many data points with the same value. Don't set all non-detected values to the reporting level; your results will be wrong.

The fourth way is to set all censored data set to 0.707 of the reporting value, an arbitrary value. Any arbitrary value is wrong; unknown values are unknown and any number replacing them is wrong and really distorts the value:frequency distribution.

These four ways commonly used on censored data means that describing the distribution of these concentrations cannot be done correctly with models based on the normal distribution of concentration values. Lab result analyses need to use models that are based on value frequency distributions that are skewed to one side (higher concentrations at lower frequencies), and all values are positive, greater than zero.

Statistical models that meet these criteria exist and are commonly used in non-environmental applications. There's no reason not to apply then to environmental data, too. The results are provably robust, technically sound, and legally defensible.

Biological data.

Biological data (animals and plants) have their own complexities when we need to estimate population sizes, use of habitats, and their interactions with humans (particularly environmentally permitted industrial activities.)

When surveying animal populations they may not be observed at that time, even when they've been observed in prior visits to that location. This does not mean they are not present, only that they were not seen or heard. Conventional statistical and mathematical models are not able to incorporate this uncertainty in their results. Bird watchers, hunters, fishers, and fish and wildfire managers know that animals might be seen (or heard) at specific locations some times and not at other times.

Aquatic biota are difficult to survey even in wadeable streams with clear water. Fish scare easily and quickly move away. Benthic dwellers such as juvenile insects are present in different live stages so some will always escape capture regardless of net mesh size. Others, such as the large stonefly species Pteronarcys californica, can dig into the stream substrate so quickly we cannot capture them. It has been documented that immature aquatic insects can survive deep in the sediments as long as there is sufficient well-oxygenated water. In the northern Intermountain west they've been known to come out of kitchen faucets several hundred meters from a stream when wells supply the residence's water.

Plant surveys conducted by following transect lines or randomly tossing 1-meter square frames and counting numbers of each species along the transect of within the square miss the most important aspect of their distribution: whether the plants are clustered, randomly dispersed, or in a regular pattern (e.g., single species farm) over a broad area. Such surveys can also miss important plants because they are outside the searched area.

Environmental data are collected at locations and times mandated by regulatory agencies. They differ from ecological research data that are collected at regular time intervals and cover the entire area of research interest. They also differ from financial, business, and public data that lack spatial and temporal attributes of environmental biological data. These differences must be accommodated by the models used to analyze them to get robust, technically sound, and legally defensible results.

The two types of models applied to environmental data

With the previous background I can now show you how environmental data are almost always analyzed, and why they fail to both provide valid results and address decision-makers' concerns about forecasting and causation.

There are two types of analytical models applied to ecological and environmental data: mathematical differential equations and statistical.

Mathematical models

When ecologists began to transition from observational to quantitative in the 1950s and 1960s they adopted or modified available engineering models. Mathematical models also were the tools of choice when environmental statutes and regulations were introduced, perhaps because they were successfully applied to static components of the built environment such as buildings and bridges. While their limitations for highly variable natural ecosystems were accepted then, there is now no benefit to not replacing them with statistical models. However, too many are still in use today, sometimes required to be used by regulators.

Mathematical models grow out of equations that define how a system changes from one state to the next (differential equations) and/or how one variable depends on the value or state of other variables (state equations). They also can be divided into either numerical models or analytical models. The model structure is determined by creating equations that express what is believed to be the relationships between a response variable and explanatory variables. Therefore, mathematical models require input data be fit to the fixed equations of the model. Often, these complex models require many variables and a large amount of data for each variable as inputs. Acquiring sufficient data is expensive (in time and money), or not possible to acquire, so it is common for modelers to estimate or assume values for unknown rates and constants. This is often called "best professional judgment."

For geochemical and aquatic ecosystem concerns four mathematical models are commonly used to inform operational, regulatory, and policy decisions. These are HSPF (Hydrologic Simulation Program-FORTRAN), QUAL2E (Enhanced Stream Water Quality Model), Pit Lake Hydrodynamic and Water Quality Model (PITLAKQ), and Biotic Ligand Model (BLM). The complexity and comprehensive inclusiveness of these models require very large amounts of data if they are to incorporate inherent natural variability. These models are useful research tools to increase understanding of the mechanisms and dynamics of the systems they model, but they are inappropriate for operational, regulatory, or policy use because time and cost constraints limit the quantity of input data and because the output is determined by the structure of the equations.

For biological data the focus has been on presence/absence of habitat use with two main models: HSI (Habitat Simulation Index) used for terrestrial and aquatic animals and Instream Flow Incremental Methodology (IFIM) for fish. The latter is a subset of the PHABSIM (Physical Habitat Simulation) set of mathematical models. Both were developed in the 1970s and sometimes still required to be applied today. HSI models were (and perhaps still are) based on subjective assignment (on a scale of 1--4) of several habitat attributes such as size and quality. The numbers are then entered into a computer program that crunches them into a HSI value. Determining whether that value is good or bad is also subjective. The IFIM and PHABSIM models are no more quantitative than are HSI results.

Statistical models

Statistical models address concerns such as characterization of the data distribution, that is estimates of the expected value, variance, skewness, etc.; estimation of the probabilistic future behavior of a system based on past behavior (forecasting); extrapolation or interpolation of data based on probability (predicting); error estimation of observations; or spectral analysis of data.

Unlike mathematical models, statistical models are fit to existing data. It is common to apply several such models and mathematically determine which one best fits the data. This approach is applied to both geochemical and biological data; with biota the model tends to prioritize presence values over absence values rather than treating both equally.

There are two main statistical paradigms (approaches) in the analysis of environmental data.

Frequentist

The frequentist paradigm is the most familiar because it is taught in all basic statistics courses as part of a science or business curriculum. This paradigm is based on the expected frequencies of values if the population is described by a Gaussian (normal) probability distribution. There are two types of frequentist analysis: null hypothesis significance testing (NHST) and information theoretic.

The history of the frequentist paradigm began in the early twentieth century when R.A. Fisher, a British statistician and geneticist applied statistics to biological experiments. In 1919 Fisher became the statistician for the Rothamsted Agricultural Experimental Station in England and developed models for plant-breeding experiments.

Because the structure of agricultural experiments could produce biased results he introduced the concept of randomizing experimental variables. Rather than using only a single explanatory variable in an experiment, as chemists always do, he created the concept of analysis of variance (ANOVA) which he applied to a set of subexperiments that allowed him to examine the effects of multiple explanatory variables, treatments, in a single experimental design. By treating the experimental outcomes as caused by the different variables, and measuring how much each variable contributed to the results he greatly advanced how biological experiments were conducted. Fisher's statistical approach was the foundation for the content of almost all basic statistics courses: the null hypothesis significance test (NHST.)

Another statistician at that time, Karl Pearson, developed the concept of the "p-value," the probability of accepting the null hypothesis that that the results of two explanatory variables do not significantly differ. Fisher argued strongly that Pearson's threshold criterion, 95%, was arbitrary and no more meaningful than, for example, 80%. While Fisher's argument is true, Pearson's value became the defacto standard.

There are several mathematical and logical reasons why applying null hypothesis significance test statistical models to environmental data fail to produce technically sound and legally defensible results. One reason is the arbitrary nature of the 95% decision criterion which means that the probability of the two (or more) samples being different is less than 5% due to chance alone. Another reason is that the null hypothesis (that the samples are not from the same population) is the only one tested;

the alternative hypothesis (that the samples come from the same population) is not tested, only assumed by default. Most importantly, the reason to not apply any NHST models is that environmental data are observational, not experimental.

A modified frequentist paradigm is Maximum Likelihood Estimation (MLE.) These models do not assume any specific result (such as null or alternative) but examine several results and estimate the one having the maximum likelihood of describing the true relationship. The MLE modification of the NHST still applies to only experimental data, not observational environmental data.

Bayesian

The second statistical paradigm is the Bayesian paradigm. It incorporates existing knowledge (prior experience) into the prediction of future conditions. While this might seem counter-intuitive or flaky it is a robust approach well suited to many environmental data sets, particularly biological ones. Perhaps it is more easily accepted when we understand that we very frequently apply this approach to make everyday decisions. We tend to fish and hunt where we have been successful in the past, we select commuting routes and times to avoid congestion and delays we have experienced in the past, and we buy our foods from markets we know have the quality and prices we want. What the British theologian Dr. Thomas Bayes did was to apply rules of logic to this prior knowledge using mathematical probabilities.

Robust Prediction, Forecasts, and Causation

You learned how environmental data differ from other types of data, and how the models applied to them affect the quality of their analyses. Now you will learn that robust analyses answer the questions that regulators and finders of fact ask about past, current, and future environmental conditions.

While predicting and forecasting are often used interchangeably, they mean different things in environmental science. A prediction is an estimated value in a location that has no measured value. Predictions are most commonly used in water quality assessments where chemical constituent concentration measurements are available from sparse locations. A forecast is the prediction of a geochemical concentration or biological condition at a future time.

Predicting the geochemical concentration of a component of interest at an unmeasured location is done by applying an interpolation model. A common approach uses the weighted mean value of nearby measurements, with the closer measurements contributing more to the interpolated value than those further away. It's a very handy tool when there are sufficient data surrounding the location for which a value needs to be calculated. There are more robust methods applied but they all have the same purpose of estimating a value at an unmeasured location based on nearby measured values.

As that great philosopher Yogi Berra said, "It's tough to make predictions, especially about the future." This applies to the need of regulatory decision-makers when presented with a new permit application and with finders of fact in environmental litigation presented with opinions of competing expert witnesses.

Forecasting and causation are closely related, although too often forecasts do not include understanding of the causes. Extrapolating a time series of an environmental variable is simple to do, and frequently abused.

The US EPA has an established protocol for ecological risk assessments. Their web page tells us that,

> An ecological risk assessment is the process for evaluating how likely it is that the environment might be impacted as a result of exposure to one or more environmental stressors, such as chemicals, land-use change, disease, and invasive species.

They explain that the process has three phases following planning: problem formulation, analysis, and risk characterization. They have a series of tools (mathematical models) to be applied to different situations; for example, the Water Quality Analysis Simulation Program (WASP7), the Stochastic Human Exposure and Dose Simulation (SHEDS), and the Community-Focused Exposure and Risk Screening Tool (C-FERST). I've attended meetings with regulators and other consultants who apply these models and extrapolate the results for several hundred years. That is highly unrealistic, yet can be accepted by the regulator. The problem, of course, is that all differential equation models fix environmental state transitions and other processes in fixed equations that do not reflect ecosystem differences by location, time, and changing climate conditions and weather patterns. The results can effectively be challenged.

Environmental impact statements (EIS) are required for the Bureau of Land Management (BLM) and US Forest Service (USFS) to make a decision about permitting a major project on public lands. The process is long and costly. Collecting baseline data usually requires at least two years and the analytical results are commonly appealed or challenged because the models could not be proven appropriate and robust. Applying the appropriate statistical models, and interpreting the results using established ecological theory benefits all involved interests. With climate warming and weather patterns changing existing permitted projects can benefit from collecting more of their baseline data and determining if the forecasts were accurate. These additional data analyses also allow for operational modifications which adjust for changing conditions.

Storm water discharge permits are probably the most numerous environmental permits issued by regulators. They affect almost all industries and permit compliance is the responsibility of the permit holder to demonstrate that their operations are not adversely impacting receiving waters by either point or nonpoint source discharges. The two actions required to be undertaken by permit holders to better secure their future are to collect more data than required by their permit and have them analyzed correctly to document cause-and-effect of their discharges on the receiving waters.

Explaining current environmental conditions, and forecasting future conditions, requires defining and quantifying their causes. Not only what those various causes are, but how much each contributes to the observed results. While multivariate regression analyses do this with many types of data they are not appropriate for geochemical and biological environmental data.

Remember that geochemical collections always consist of a subset of all the chemical constituents in a water, sediment, soil, or rock sample and biological collections, for example of a macroinvertebrate community at a stream or river location, are not representative of all organisms and their numbers at that location and time. Non-environmental data summaries of mean, variance, standard distribution, and similar measures assume that the data represent their values in the entire unmeasured population. We know that environmental data represent collections, not statistical samples, because the populations and unknown. This means we need to analyze each collection as a composition, a set of components considered as the population, and assess change based on the relative abundances of each component in the composition. These relative abundances are almost always within a limited range of values and the causes can be more easily identified and quantified based on the ecosystem in which they are located.

When the causes have been identified the paths from sources to results can be defined and the relative contribution of each path calculated. Now you have the means to make informed forecasts and explain the causes of the observed effects.

In environmental litigation it is easier for finders of fact to accept data analyses when they understand how the constraints on those data are addressed by the analytical model.

When it's important to you that your environmental data analytical results be technically sound and legally defensible, contact me[1]. My strong record of providing clients and decision-makers with results they understand and can use to justify operational, regulatory, and legal decisions can help you to sustainable success.

[1] See the top of the first page.