How To Maximize Data Value*

Richard B. Shepard, Ph.D.†

1 Introduction

For executives in the highly regulated natural resource industries, it is appropriate to look at a usually overlooked aspect that affects costs and the bottom line. You know that your business thrives, survives, or fails based on the information you have. Information comes from raw data, and that is the topic of this article: the real cost and value of data, information, and usable knowledge.

A quick review

Data are the raw numbers collected from the appropriate sources at the proper times using the correct methods. This is true for all data: geological, mineralogical, financial, chemical, biological, climatalogical, ecological, hydrological, sociological, and economical. These data are analyzed numerically, spatially, visually, or statistically and converted into information. We turn the information we receive into knowledge based on our experience, training, and intuition, and then we use this knowledge to make decisions. Obviously, informed decisions cannot be made without the highest quality data. It does not matter how sophisticated are the analytical tools we use, how much jargon we speak, or the number of initials after our name, we cannot turn low quality or missing data into useful information. Unfortunately, too many times this futile attempt is made. Such attempts come out of academia, regulatory and resource agencies at all levels of government, consultants to business and industry, and employees within an organization.

The real cost of data is more than what you paid someone to collect it. Sometimes we cannot calculate the real cost until it is too late: we have made wrong decisions and we cannot calculate the cost of lost business. Sometimes we can calculate the real cost because our applications are rejected by permitting agencies and we have to start all over again. It makes much more sense to reduce the real cost of data as much as we can while maintaining—or increasing—the value we get for our money.

†President, Applied Ecosystem Services, Inc., Troutdale, OR 97060

Basic Principals

Three truisms apply here:

- We cannot return at a later time to collect missing data we need.
- We must know the ultimate decisions to be made before we know how to begin.
- The real world is highly variable in place and time, and we very rarely can estimate the amount of variation.

2 How to Guarantee Data Value

Know just why you want to collect data

If you are establishing baseline conditions against which future measurements will be compared, you need different techniques than if you are collecting data for compliance with permit conditions.

Know what attribute of the collection technique is important

If you are measuring settling rates of fill, you need samples which can be analyzed by time series methods. If you are measuring discharge water temperature, you need samples in the receiving water, too. If you are measuring water quality downstream from your project site, you need samples which are time-consistent with distance from the discharge point.

If you are evaluating potential impacts on an ESA-listed animal species, you need to sample when the animal is present in the area of concern. If you need to determine chemical contamination on sediments, then whether the chemical of concern is organic or inorganic determines what sediment types need to be collected. If you are measuring toxicological effects, then you need to know whether thresholds and other standards were determined in a laboratory or in the real world; in the latter case, where (geographically) were they determined and whether those conditions are relevant to your system.

Know where data need to be collected

In many situations the spatial distribution of values is more important than the magnitude of those values. The spacing of the sampling points (evenly distributed, clustered or randomly located) determines the interpolation method which translates the data points into contour lines or a 3D surface.

The importance of spatial knowledge is understood by all hunters and fishermen, but is sometimes forgotten when collecting scientific data for regulatory purposes. For example, the distribution of sediment types in an aquatic system is not uniform, so taking grab samples across transects is rarely appropriate. When samples are likely to be spatially autocorrelated (common in

surface and ground water pollution studies), the sampling locations must take this autocorrelation into account.

If you are looking to determine tributary effects (or the effects of a point source discharge on the receiving water body), then samples must be taken upstream of the confluence/point source, at the confluence/point source, and downstream over a sufficient distance to determine mixing dynamics (and determine if the values measured downstream come from further upstream or from the tributary/point source).

Know when the data need to be collected

Terrestrial plants are best identified during the height of their growing season, not in the middle of winter. Aquatic sampling must be done on a regular schedule over at least a year to produce meaningful data (because the insects and other invertebrates have many different life cycles throughout the year). Bird presence depends on their migratory patterns, so different sampling times produce different results.

Salmonids and other cold-water fish species hide in off-channel refugia or deep pools during hot summer days, but they come out to feed about a half hour after sunset when the benthic invertebrates start drifting down river. Migration of anadromous fish occurs in comparatively narrow time periods in tributary rivers and streams; measuring water quality and physical parameters when no fish are in that reach yields nothing that could add to our knowledge of the fish.

Any time the purpose of the data collection is to detect trends or other changes over time, samples *must* be collected at fixed intervals. You cannot begin collecting data monthly, then change to quarterly, semi-annually, or annually. Well, truthfully, you certainly can do this, but then the data cannot be correctly analyzed using the correct statistical techniques.

Know how to collect the data

Using the incorrect tool or technique may give you numbers, but they will not be useful numbers. Having a survey crew locate wetland boundary flags or animal nests within centimeters is a waste of money; nothing in the natural world is that precise. Most natural features (ore or mineral bodies, wetlands, forest edges) are transition zones over a broad area. Usually 1 to 5 meters of accuracy are sufficient. Elevation measurements are a different story. For these, take measurements accurate to less than 1 meter.

When collecting biological data, the method used *always* has biases that cannot be accurately measured. For example, the nets used to collect samples of aquatic invertebrates have a fixed mesh size. If the mesh size is large then small individuals pass through and are not part of the sample. If the mesh size is too small, then back-pressure of the water washes individuals out of the net.

The quality of water quality samples also is highly dependent on method. In a flowing water system, it is important to know if the water came from the

surface, near the bottom, or somewhere in between; near one bank or from the center. There are many standard samplers that can be lowered to a particular depth, then closed to keep that sample representative of the location. These samplers are designed to open horizontally or vertically. Each type has an appropriate application.

Sediment samples are fairly easy on land: dig a hole or fill a bucket with a standard volume or weight of material. In aquatic environments, however, it is much more complicated. The sampler must be appropriate to the particle size of the sediments (from silts and muds through sands and gravels to cobbles and boulders), and to the analytical procedure to be applied to the sample. Sometimes metal samplers are appropriate, sometimes they will interfere with low levels of chemicals to be analyzed so various plastics need to be used as the material in contact with the samples.

Follow the above guidelines and you will have the maximum value possible from each sampling dollar spent.

3 The Next Step: Analyzing Data

Sitting someone down at a computer and showing him how to run a word processor does not guarantee a well-written document will come out of the printer. Similarly, using computer-based statistical software without understanding the assumptions and requirements of each test (and the nature of the data being tested) does not produce well-analyzed information. This step of converting raw data to useful information is critically important. It must be done correctly if you need results that are technically sound and legally defensible.

Parametric and Non-Parametric Numerical Statistics

Like a soufflé, quality results from your collected data are dependent on gentle handling in the right way. Statistics is the branch of mathematics that lets us calculate useful summaries of population features from a sample (or collection). This means that we do not need to measure or count every individual to say something meaningful about the whole population.

For example, animal counts and other population abundance estimates need to be analyzed by what are called *non-parametric* statistics rather than by the more familiar *parametric* statistics. The reason for this is *parametric* statistics are based on the parameters of distributions used to estimate total populations from samples. This means that the sample must be Representative of the population as a whole. With animals and plants, this requirement is rarely met. Sampling men's heights using basketball players does not reflect the distribution of heights of all men. Similarly, the animals in our sample cannot be assumed to truly represent the population as a whole, but we can get very useful summary statistics by applying the appropriate *non-parametric* test which is based on less restrictive assumptions.

For other questions, spatial statistics must be used rather than numeric statistics. The distribution of water or sediment chemical content is one of the most common examples of this; another is the distribution of plants or animals. We need to know where the animals are, not the average number per unit area. Similarly, chemicals are not uniformly distributed in sediments. Organic chemicals are found on organic sediments (small leaves, other wood fragments) and on silts and clays. Inorganic chemicals are most commonly found on inorganic sediments such as sands and gravels. The forces of the moving water (in lakes and reservoirs as well as in rivers and streams) sort the particles by size, with the smaller sizes settling on the bottom and remaining only in the slackest water and the larger particles in faster waters. Actually, suitably rigorous aquatic sediment sampling is quite complex to do correctly, but that's the subject of a separate article.

And, if numerical or spatial modeling is applied to the data, you must know the model's assumptions and number crunching techniques to understand just how it is massaging your carefully collected data. If you are looking at changes over time, remember that there is a limit how far from your last data point you can validly extrapolate your prediction. Think of this as balancing a long board on a table: the further the board hangs off the edge of the table, the less stable it is. Eventually, the board falls off because there is not enough support for it. Time series data suffer the same fate.

4 Interpreting the Information

The analyzed data is information on the state of the systems you measured. To be useful, this information must be transformed into knowledge and insight. Technically sound and legally defensible decisions are made using knowledge and insight generated through the proper analysis of correctly collected raw data.

Many tools and techniques are available to interpret information. Two of the most common are the plot (graph in business terms) and the map. Plots and maps reveal patterns in the information: changes over time and space and the variability within the data on which the information is based. There are many types of plots and maps, and choosing the most appropriate type needs understanding of the questions about which decisions are to be made, the technical knowledge of the decision-makers, and the amount of information that needs to be integrated for full understanding.

5 It's Your Call

As scientific consultants, it is our responsibility to collect high quality data, use the appropriate analytical techniques and present the information using the best combination of words, graphs and maps. Part of our role is to share with you our experience and knowledge by suggesting alternative courses of action.

We offer the pros and cons to each alternative, and we offer guidance as you request. But, in the end, it is your business and your responsibility to make the decisions. When we do our job properly, your decision is easier to make. You are responsible for the success of your business, and you know where you want to take your company or project. We have given you the maximum value possible from your data when you make that decision with full appreciation for the foundation on which it is based, and with full confidence that no one could make a better decision than you could.

6 About the Author

Richard B. Shepard, Ph.D. is a stream/watershed ecologist and fluvial geomorphologist with more than 30 years experience across the US and overseas. After stints in academia and state government, he turned to consulting in the private sector. In 1993 he formed Applied Ecosystem Services, Inc. to serve the natural resource industries by accelerating environmental permitting. The company's business is strictly with the private sector on permitting issues. He can be reached at 2404 SW 22nd Street Troutdale, OR 97060-1247; telephone 503-667-4517, fax to 503-667-8863, or send e-mail to info@appl-ecosys.com.