7. METHODS AND TOOLS FOR ANALYZING SPATIALLY EXPLICIT INFORMATION

Susan Cormier, U.S EPA Office of Research and Development, Cincinnati, OH
Jeff Hollister, U.S EPA Office of Research and Development, Cincinnati, OH

What is in this chapter? Several common approaches are described for analyzing geographical information from field or remotely sensed data. The chapter is organized by the three common elements of assessment: planning, analysis, and synthesis. Highlights include general analytical considerations for descriptive and associative analyses, example methods and interpretations for *in situ* data used to understand causal relationships, and analysis of spatial data using overlays, map algebra and other spatial analytical methods. Statistical and spatial analysis tools and their potential use in water quality programs are available in the Toolbox.

Environmental decisions are not made using raw data. Data must be analyzed to make it useful and informative. Descriptive statistics characterize the population or area from which data is obtained and can be used to track changes over space or time. Extrapolation from first principle and empirical models can be used to estimate stressors and their sources, target monitoring, project likely outcomes of remediation, and help target and prioritize remedial effort. In a nutshell, analyses that demonstrate causation can be used to understand why a watershed has changed and predict how it can change in the future and decide what to do to protect and rehabilitate environmental quality.

7.1. INTRODUCTION

Spatially explicit data can be used in any assessment and any facet of an assessment. As previously discussed in Chapter 3, assessments are composed of three activities: planning, analysis, and synthesis. Furthermore, assessments often depend on other assessments to be successful. A condition assessment that discovers a problem is simply a witness to environmental decline. It does little to inform action unless there are subsequent assessments. Similarly, a remedial action plan that is directed at the wrong cause or source will probably fail. Seeing the bigger picture from a landscape, watershed, or regional perspective helps to remind us that integrating assessments is often necessary. Also, the imagery provided by maps, charts, and other

outputs are wonderful communication tools that can make environmental problem solving a community affair.

This section expands on planning, analysis, and synthesis with a greater focus on using the data and combining the data from the field with remotely sensed data. Because of the breadth of this topic, we refer to other sources for more detailed explanations of statistical, modeling, and spatial analysis. Instead, we try to provide a basic primer for different types of statistical analysis using spatially linked data.

The steps recommended here emphasize an iterative approach to analysis and data gathering, the importance of paired stressor and response data, and focus on decisions relevant to regulatory monitoring and assessment. Most examples presented here are related to the Clean Water Act (CWA), however, successful Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) programs also depend heavily on very similar assessment approaches.

As more layers of data and information are brought together, there is more effort in the activities involved in the assessment. These include careful planning for more intensive geographic information system (GIS) and statistical analyses, gathering landscape and *in situ* data specifically tailored to the project and analysis methods, and deriving a wider array of landscape factors. Additional statistical analyses are needed to reduce the number of variables and to develop robust stressor/response relationships, and more formal analysis of the statistical power (and likelihood of false negatives and false positives—alpha and beta) of the results. Successful assessments often require peer reviewing and publishing the results, and wider and more formal use of extrapolations to make targeting, priority, and other water quality management decisions.

7.2. PLANNING AND PROBLEM FORMULATION

7.2.1. What New Information is Sought?

Four major types of assessments are described in detail in Chapter 3. Here we remind you that it is essential to know what you are trying to do and the sequence of activities you might need to perform. Some activities and products are listed in Table 7-1. Organization is essential. If you know the cause and the source and want to

TABLE 7-1

Activities Associated with Different Assessment Types

Type of Assessment	Activity Performed by the Assessment
Condition	Identify impaired waterbodies or other ecosystems
	Identify impaired waterbodies or other ecosystems that were not sampled *in situ*
Causal Pathway	Determine the cause of a biological impairment
	Find the sources of a pollutant or stressor
	Estimate the amount of stressor contributed by each source
Predictive	Estimate the risk to a valued resource from a stressor
	Estimate the amount of stressor that can occur while also maintaining designated uses
	Evaluate options for preventing degradation or remediating ecosystem quality
	Develop protective or remedial benchmarks such as water quality criteria and standards
	Target areas to prevent degradation
	Prioritize among options and targeted areas for management action
Outcome	Evaluate the performance of management actions and BMPs to reduce stressor loads
	Evaluate the effectiveness of actions to protect or rehabilitate ecosystem services
All	Develop communication materials for education and partnership programs

BMP = best management practice.

Note: Also see Table 1-2 Spectrum of Uses for Landscape and Predictive Tools.

regulate the release of the substance, you do not need a condition or causal pathway assessment. Instead, you need a predictive assessment to develop mitigation or protective benchmarks or perhaps even water quality criteria. If you want to cost-effectively place a limited number of on-the-ground controls in a watershed, that is also a predictive assessment; however, it is done differently from a criterion assessment (Suter and Cormier 2008b).

Also, beware that many costly remediation efforts have fixed the wrong problem because an ecosystem condition was natural or because the cause or source was assumed. Even more studies have been done that never informed a single environmental decision. Being acutely aware of the objectives and the implementation team is the difference between work that is interesting and work that is interesting *and* guides environmental protection and rehabilitation.

7.2.2. What is the Regulatory Authority or Social, Political, Economic Driver?

Assessments are necessary to implement all other types of regulations. For a review, see U.S. Environmental Protection Agency (EPA, 2010a) and Chapter 2 for some of the CWA sections and related programs. However In some cases, the analysis is prescribed and others they are not. In all cases, a sound scientific argument is allowed, but it must be more thoroughly defended. Therefore, using some of the newer, more powerful statistical and spatial analysis tools requires a bit more explanation in reporting.

7.2.3. What Needs Protecting or Rehabilitating?

All environmental assessments require that the assessment endpoint be defined. The more precisely defined the assessment endpoint, the more likely that the analyses will meet the needs of the needs of the assessment (Suter, 2007, Cormier and Suter, 2009). Assessment endpoints might be physical, chemical, or physiological attributes of entities such as organisms, water, habitat, ecosystem functions, or ecosystems in their entirety. Some typical examples are listed below:

- Habitat/channel/geomorphology
  - e.g., sinuosity, embeddedness, clarity, depth, suitability
- Chemistry
  - e.g., concentrations relative to criteria
- Hydrology
  - e.g., removal, flashiness, low flow, overland flow, ground water
- Biology
  - e.g., species richness, abundance, reproduction, presence/absence, invasive species
- Temperature
  - e.g., extremes, climate change

7.2.4. What Type of Analyses Need to be Performed and How Good do They Need to be to Make a Decision?

Aim for scientific elegance, parsimony. A precise value might not be needed to make an environmental decision. Qualitative information or a range often can get things moving in the right direction. Easier, faster, cheaper assessments mean that more assessments can get done (Suter and Cormier, 2008a). Consider the type of analyses and the data that is absolutely necessary versus what would be nice to have. Some examples of useful data are:

- Gradient of sites covering full range of stressors
- Laboratory studies
- Probability survey data: biological response and stressors
- Before/after and control/impact designs
- Remotely sensed data
- Preprocessed spatial information

Will a screening assessment suffice? Can adaptive management approaches be reasonably used? Are existing data sufficient to answer questions with required sensitivity? If not, is additional data needed? The next step is to draw up a project plan and assemble the data sets and tools (see Geospatial Toolbox).

7.3. ANALYSIS

Because of the range of potential types of assessments, places, and objectives, we provide some general and common methods and tips but must rely on the many statistical, spatial analysis, and decision support publications and Web sites for greater detail. The examples in Section III illustrate different types of analysis, but the associated reports also provide much more detail. The intent is to provide a primer and show the connections needed to complete a wide variety of assessments and analyses. The list of methods in Chapter 7, while not exhaustive, will provide a basis for conducting assessments that use spatially distributed data to inform assessments. The statistical methods were largely taken from the Analyzing Data section of the Casual Analysis Diagnosis Decision Information System (CADDIS) Web site (http://cfpub.epa.gov/caddis/analytical_tools.cfm) and are associated with CADStat (http://cfpub.epa.gov/caddis/analytical_tools.cfm?Section=144), an interface to easily perform these analyses in R, which is an open-source statistical package.

7.3.1. Software

There are numerous sources of software available for conducting statistical analyses. Many of the software tools are general use tools that provide access to a range of statistical tools, while others are designed for a specific suite of analytical methods. The list below discusses both and presents examples of commercial and open-source solutions.

7.3.1.1. *Commercial*

Some commonly available commercial software includes S-plus, SAS, SPSS, and Matlab. However, even commercial spreadsheets such as Microsoft Excel have some statistical computing capability.

7.3.1.2. *Open Source*

R (http://www.r-project.org) is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows, and MacOS. It is capable of running all the statistical analyses offered by commercial products; however, it is somewhat more difficult to master. For commonly performed analyses described in this chapter, a user-friendly front end can provide the statistical code thus simplifying analysis.

One is CADStat (http://www.epa.gov/caddis/da_software_cadstat.html), a menu-driven package of several data visualization and statistical methods. It is based on a Java Graphical User Interface to R (http://www.xmarks.com/site/jgr.markushelbig.org/JGR.html). Methods in this package include scatter plots, box plots, correlation analysis, linear regression, quantile regression, conditional probability analysis, and tools for predicting environmental conditions from biological observations. Another is CProb 1.0, which is a tool written using R as the back-end statistical processor and Microsoft Excel as the front end interface to calculate conditional probabilities and bootstrapped estimates of confidence intervals. CProb 1.0 (http://www.epa.gov/emap/nca/html/regions/cprob/) is a Microsoft Excel Add-in, developed with the R language and environment for statistical computing, the R(D)Com Server and Visual Basic for applications, intended to aid in conditional probability analysis. It offers options to generate scatter plots, cumulative distribution functions, and conditional probability plots (Hollister et al., 2008).

7.3.2. Geographic Information System Tools

7.3.2.1. *Commercial*

ArcInfo and ArcView, marketed by Environmental Systems Research Institute, are the dominant commercial GIS analysis software systems. These high-quality systems are heavily licensed. IDRISI was one of the first raster-based systems and was developed by Clark University, Worcester, Massachusetts, to provide access to GIS tools in less developed countries. As such, it is simple to use and master. ERDAS (Earth Resource Data Analysis System) Imagine® is proprietary software for GIS and image processing. ENVI is often used to interpret remotely sensed imagery. For more

commonly used proprietary software, see
http://en.wikipedia.org/wiki/List_of_GIS_software.

7.3.2.2. *Open Source*

There are many open-source GIS analysis and visualization software packages. They all require time and effort to use and master. Geographic Resources Analysis Support System, commonly referred to as GRASS, is a GIS that has been in existence for many years and provides many of the capabilities most environmental assessors can use such as geospatial data management and analysis, image processing, graphics/maps production, spatial modeling, and visualization (http://grass.itc.it/). However, there are many other systems, and users should consider the objectives and the capabilities of these systems before investing the time it takes to organize a spatially explicit database. Chameleon (http://en.wikipedia.org/wiki/Chameleon_%28GIS%29) provides a front end for building applications with MapServer (http://en.wikipedia.org/wiki/MapServer), a very commonly used Web-based mapping server, developed by the University of Minnesota. For an extensive list, see http://opensourcegis.org/.

7.3.3. Analysis of Field-Collected Data

As a part of planning, the waterbody or region should already be chosen. In the analysis phase, the appropriate geographic frameworks and temporal scale are explicitly described and selected for a variety of field and remotely sensed data and analysis. They can include mapped areas such as stream reach, catchments, ecoregions, or other appropriate classifications to establish realistic areas for analysis, extrapolation, and assessment. Existing frameworks have been described in Chapter 5, but in some cases, the data will need to be classified and normalized before analysis can be done. This is described in Sections 7.3.3.2 and 7.3.3.3 of this chapter. The order can vary. For example, a general geographical framework can be chosen, data sets assembled, and then the geographic framework revisited so that data sets can be matched to the proper temporal and geographic scale or classified and normalized for other reasons.

It might be possible to prepare preliminary *wall-to-wall* landscape and other data coverages to document natural and stressor gradients and exposure. These can be helpful for other steps in the analyses and are useful to refine and plan analyses. Some activities are the delineation of watershed boundaries and riparian proximity buffers for sites and other appropriate landscape factors. Early in the analysis it is useful to map landcover, sources, waterbody, road, and other relevant information. For some predictive assessments, it is helpful to model patterns of movement and migration of biological entities.

7.3.3.1. *Linking Data and Scale*

As mentioned previously, assessments involve analyzing data that are based on causal relationships (Suter and Cormier, 2008a). Evidence of causal relationships is developed by demonstrating an association between two or more variables that include the cause and the effect (Cormier et al., 2010). The data that are used must be appropriately matched in time and space. The process for matching data and interpreting results also must be documented to ensure quality and transparency. However, measures from biological assessments cover a spectrum of temporal and spatial possibilities, and the relevance of that variability must be taken into account when matching data. The mechanisms by which environmental parameters affect organisms and how organisms can respond to stressful conditions influences the relevance of variability and how it is addressed. Relevant spatial and temporal scales should be considered when deciding how data should be matched. Consider the logic of the situation and possibly use a GIS for modeling spatial relationships. For details and perspectives on scale in ecological analyses, see Wu et al. (2006).

Because matched data and the resultant data sets are usually synthesized from various sources, it is of vital importance to explicitly address data management concerns. These concerns range from choosing a database management system (e.g., simple flat text files, spreadsheets, relational databases) to documenting details about the synthetic data sets (i.e., with metadata). Without adequate documentation of all data management decisions, it would be difficult, if not impossible, to repeat or defend

the analysis. For more information on ecological and environmental data management and metadata standards, see Michener (2000; 2006).

General considerations when developing a data set for analysis include:

- Matching data for relevant spatial and temporal scale, quality, and comparability.
- Identifying inherent assumption about the appropriateness of measurements, such as mean, extremes, rates, or other endpoint.
- Exercising caution when merging datasets especially when sampling schemes or methods are different.

7.3.3.2. *Classification*

During the Analysis Phase, field data are analyzed by assessors to classify ecosystems into ecologically similar classes in a way that reduces the influence of natural variables (see Chapter 5). In this way, the remaining differences among locations are more likely to be due to the variable of interest rather than other factors. For ways to deal with potential confounding that may not be related to geographically related variables, see Appendix B in EPA (2011). Aquatic systems that have comparable exposure-response relationships are used to classify system. For example, low-gradient coastal systems have different substrate compositions, flow regimes, climate, soils, and landcover than high-gradient, montane streams. Classification is also used to assure that the assessment endpoints are comparable across the region being assessed and to assure that the region used to develop models is representative of the local area being studied. For example, species native to Alaska would not be used to develop temperature tolerances for Florida. Assessors might also classify the data sets for analysis on the basis of tiered aquatic life uses (U.S. EPA, 2005; Davies and Jackson, 2006).

7.3.3.3. *Normalization*

Normalizing data allows different types of data to be compared and analyzed (U.S. EPA, 2010a). For data that is strongly influenced by a known natural factor, the

association is quantified and then removed from the signal of the variable of interest. For example, stream size affects the diversity of fish. Regressing basin area against species richness generates a fairly linear curve. The residual can be removed so that the effect of stream size is normalized.

For measurements that are qualitatively different, the values can be converted to proportions or placed in categories with or without weighting. Another option is to convert values to utilities (Efroymson et al., 2004; Hanley and Spash, 1993; Linkov et al., 2006).

7.3.3.4. *Descriptive/Association*

After the data set is characterized, information can be extracted by discovering possible patterns and associations between factors that can interact or that act as surrogates and are measured from satellite imagery (Senay et al, 2001). This is done with univariate or multivariate procedures. The methods indicate how to perform and interpret exploratory analysis for collocation, co-occurrence, correlation, range, and stressor-response associations.

Some can be used directly for assessments or combined with spatial coverages in more complex analyses. The resulting information can be used to refine the analysis and scope and types of assessments. They can be used to identify gaps and plan next steps for gathering additional *in situ*, remotely sensed, or composite data for analysis.

7.3.3.4.1. Scatter plots

Scatter plots are graphical displays of matched data plotted with one variable on the X-axis and the other variable on the Y-axis. Data are plotted with measures of an influential parameter on the X-axis (independent variable) and measures of an attribute that can respond to the influential parameter on the Y-axis (dependent variable).

Scatter plots are a useful first step in any analysis because they help the analyst to choose which relationships to model and to select models. A scatter of points that suggests the attribute responds to changes in the independent variable can be analyzed further using correlation (http://cfpub.epa.gov/caddis/analytical_tools.cfm?section=147&step=22&parent_section=143) or regression methods

(http://cfpub.epa.gov/caddis/analytical_tools.cfm?section=149&step=22&parent_section=143). However, a scatter of points without any apparent relationship is unlikely to provide insights into relationships, even using multivariate analyses. The distribution of points in a scatter plot can suggest whether the relationship is, for example, (A) linear, (B) a higher-order polynomial (quadratic shown), (C) exponential, or (D) logarithmic (see Figure 7-1). The distribution of points also can reveal apparent thresholds or discontinuities in the relationship.

7.3.3.4.2. Correlation

Correlation is a method for measuring the degree of association between two variables in a matched data set. The Pearson product-moment correlation coefficient (r) is a unitless value between –1 and 1 measuring the degree of linear association between variables. The corresponding nonparametric analysis calculates a Spearman rank-order correlation coefficient (ρ, rho—pronounced "row") which is computed using the ranks of the data and does not assume that the relationship is linear. Kendall's tau (τ) has the same underlying assumptions as Spearman's rank-order correlation coefficient but represents the probability that the two variables are ordered nonrandomly.

A value of r, ρ or τ is interpreted as follows:

- A coefficient of 0 indicates that the variables are not related.
- A positive coefficient indicates that as one variable increases the other also increases (see Figure 7-2, D).
- A negative coefficient indicates that as one variable increases, the other decreases (see Figure 7-1, A).
- Larger absolute values of coefficients indicate stronger associations (e.g., see Figure 7-2, A vs. C).

However, such correlations do not prove causation and could be due to confounding or error. Thus, correlation coefficients are only suggestive. In addition, small Pearson product-moment coefficients can be due to nonlinearity (see

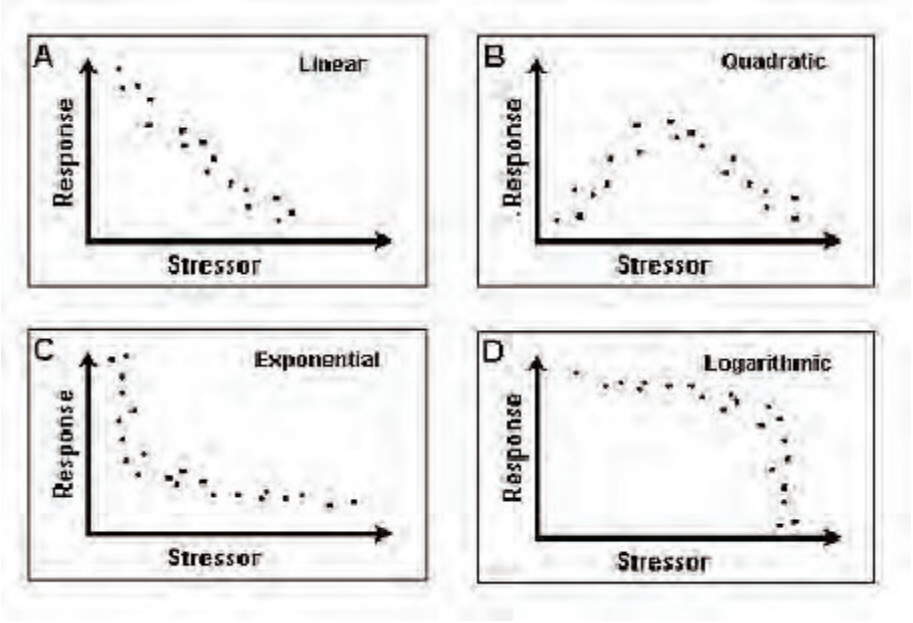


FIGURE 7-1

Scatter Plots Illustrating Different Patterns Suggesting the Underlying Form of the Stressor-Response Relationship. (A) linear, (B) quadratic, (C) exponential, (D) logarithmic.

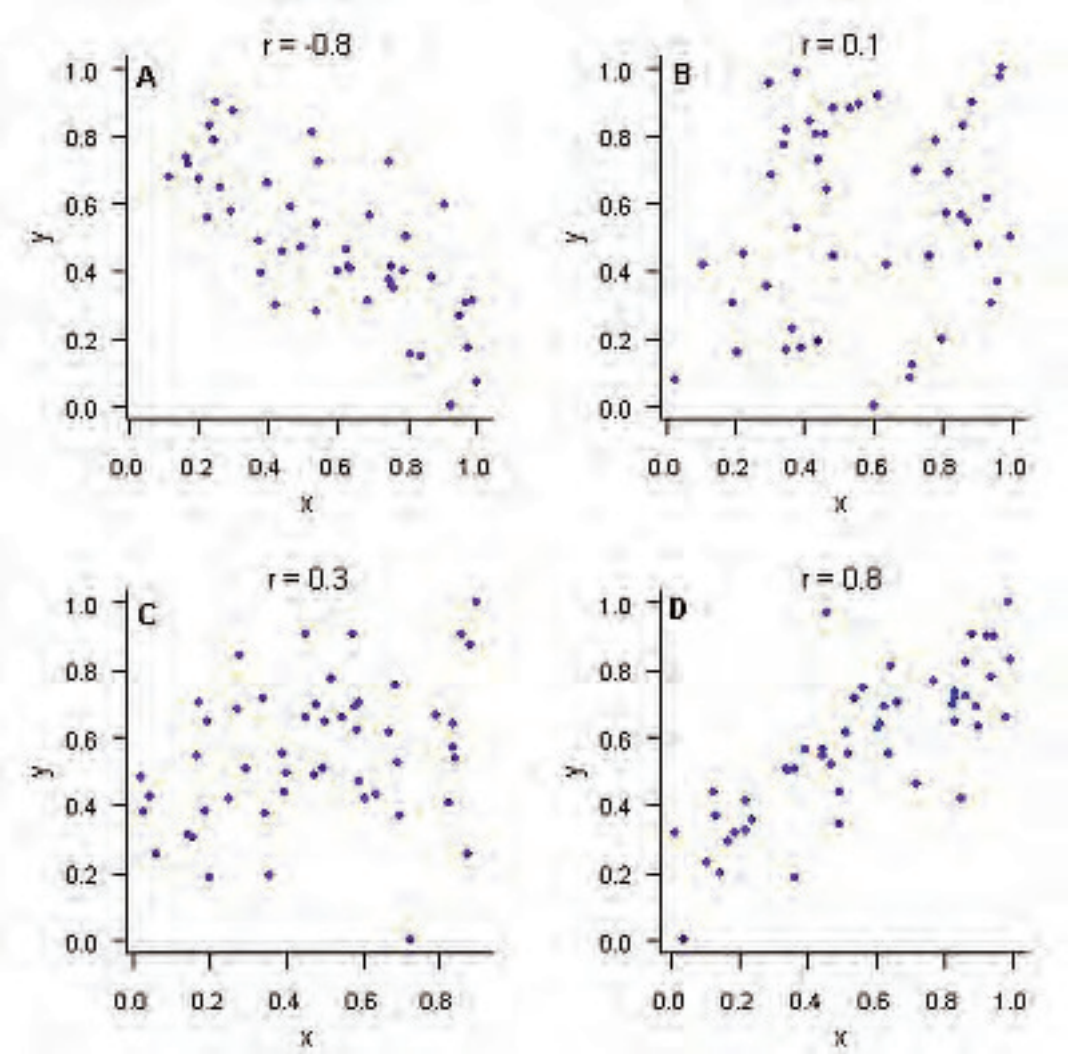


FIGURE 7-2

Examples of Different Correlations between Two Variables, x and y

Notes: (A) with an *r* value of −0.8, the band of points indicates a decrease in y with an increase in x; (B) with an *r* value of 0.1, points are diffusely scattered throughout the plot area; (C) with an *r* value of 0.3, the points indicate a weak increase in y with an increase in *x*, or perhaps a nonlinear relationship; and (D) with an *r* value of 0.8, the band of points indicates an increase in y with an increase in *x*.

Figure 7-2, C) rather than to a lack of association (see Figure 7-2, B). Therefore, scatter plots should be examined for nonlinearity and to identify outliers or unduly influential data.

7.3.3.4.3. Box plots or box-and-whisker plots

Box plots, or box-and-whisker plots, depict the distribution of observations within a data set by dividing it into four sections (see Figure 7-3). The box indicates the spread of the central 50% of the data; the median is denoted by a horizontal line through the box. The portion of the box above the median line denotes the 50^{th}–75^{th} percentile range. Likewise, the portion of the box below the median denotes the 25^{th}–50^{th} percentile range. If all data lie within 1.5 times the interquartile range (75^{th} percentile minus the 25^{th} percentile) from either end of the central box, the whiskers represent the full range of the data. If not, the whiskers extend to 1.5 times the interquartile range, and more extreme data are plotted as points. These conventions are not always followed. Box plots generated by different software can differ in the percentiles used to denote the box-and-whiskers and other features.

Because box plots depict the distribution of observations, they can be useful for identifying appropriate statistical analyses and deciding whether data should be transformed (see Figure 7-4). For example, box plots can show whether the shape of the data distribution is symmetrical or skewed. If the upper box and whisker are approximately the same length as the lower box and whisker (see Figure 7-4, A), the data are distributed symmetrically. If the upper box and upper whisker are longer than the lower box and whisker (see Figure 7-4, B, and C), the data are skewed to the right. If the upper box and upper whisker are shorter than the lower box and whisker (see Figure 7-4, D, and E), the data are skewed to the left. Box plots also reveal the kurtosis, or relative spread, of a distribution. The smaller the length of the box is relative to the whiskers and points, the tighter the distribution (see Figure 7-4, B, and D). Skewed distributions indicate that the data are not normally distributed and that the variances might not be homogeneous (see Figure 7-4, B, C, D, and E). When analyzing such data, it is generally recommended that nonparametric methods and regression models are used that accommodate nonlinear data. If parametric methods or linear

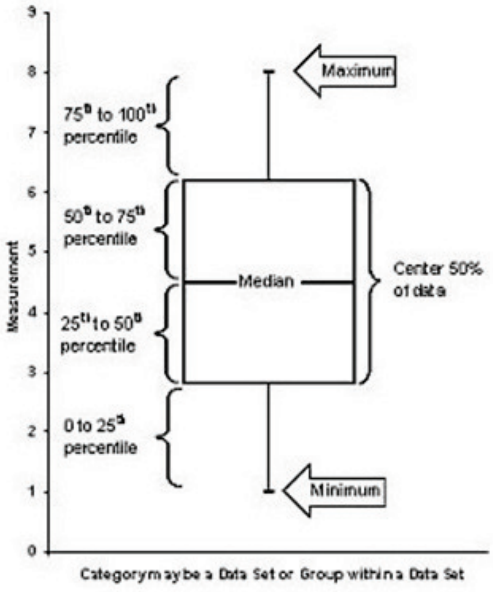


FIGURE 7-3

A Sample Box Plot with Different Components of the Plot Labeled

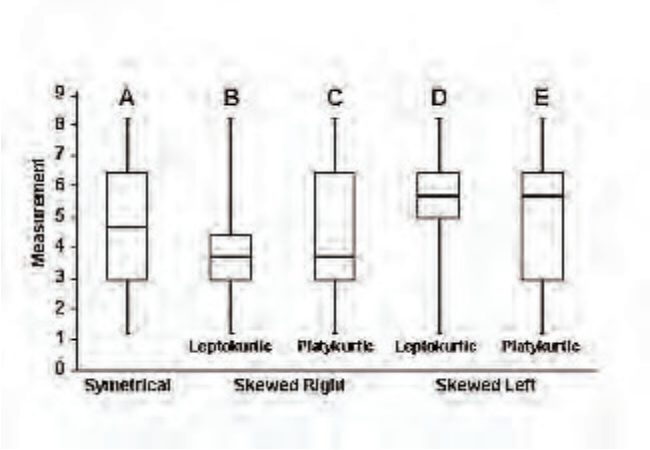


FIGURE 7-4

Box Plots Showing Symmetrical or Skewed Data Distribution and Different Types of Kurtosis, or Relative Spread

regression are used anyway, the data transformation approach should accommodate the type of data (continuous, count, proportion), the skewness of the distribution, and any zero or negative values in the data set.

7.3.3.4.4. Regression

Linear regression quantifies the relationship between a dependent (response) variable and one or more independent (explanatory) variables by minimizing the sum of the squared residuals (the difference between the predicted and observed values). The resulting line models the average value of the dependent variable for each level of the independent variable. For example, if the resulting line models the average relative losses of Ephemeroptera, Plecoptera, and Trichoptera species richness for each level of percent fines, the model can be used to predict how many species are likely to be present at 5, 10, 20, or 30% fines (see Figure 7-5).

Quantile regression models the relationship between a specified quantile of a dependent (response) variable and an independent (explanatory) variable (U.S. EPA, 2006). The resulting line represents the upper value of dependent variable that would

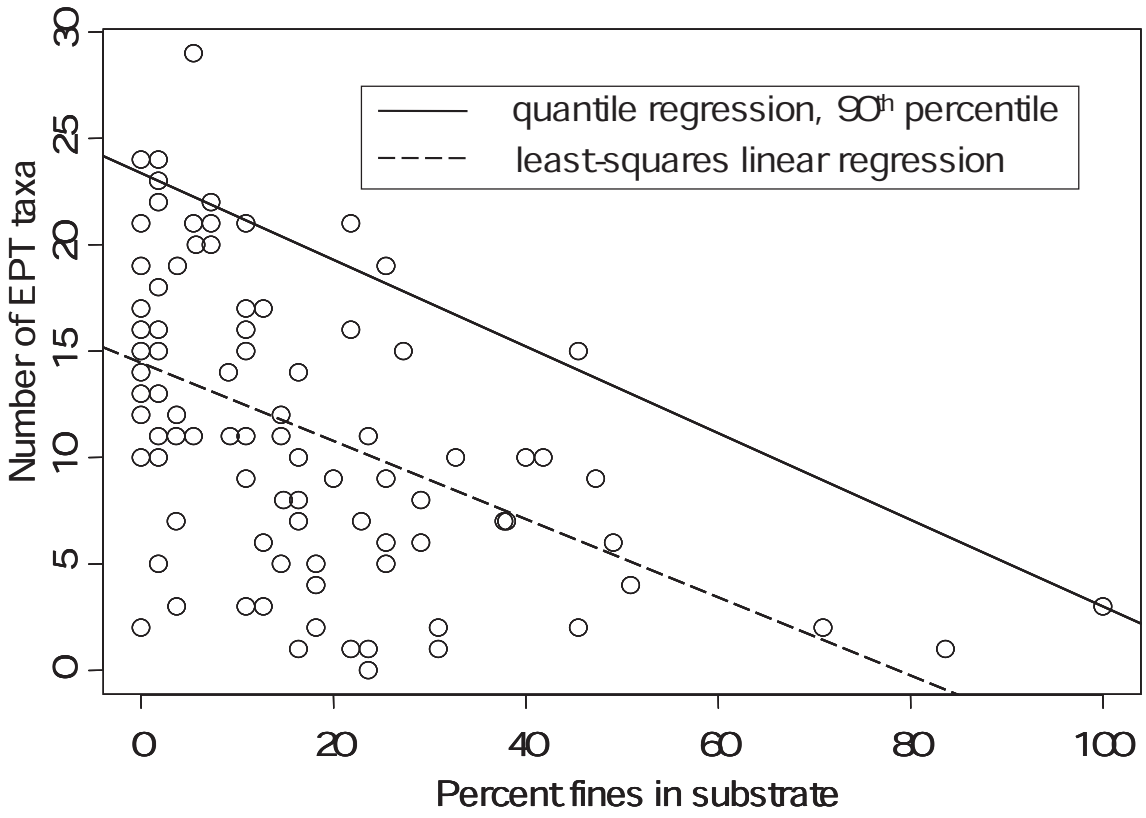


FIGURE 7-5

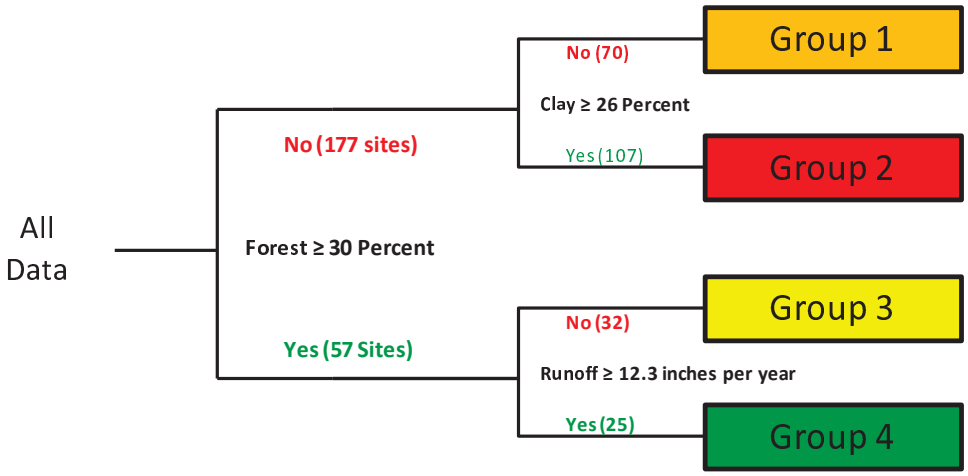
Scatter Plot with 90th Percentile Quantile Regression (solid line) and Ordinary Least Squares Regression (dashed line)

be expected for each level of the independent variable, which is an estimate of the effect when the independent variable is the limiting factor (12.5) (see Figure 7-5).

Classification and regression tree (CART) recursively partitions a matched data set of categorical variables (for classification trees) or continuous variables (for regression trees) into progressively smaller groups, using binary splits based on single independent or predictor variables (De'ath and Fabricius, 2000; Prasad et al., 2006).

CART analysis constructs a set of decision rules with the independent variables. During each recursion, splits for each independent variable are examined, and the split that maximizes the homogeneity of the two resulting groups with respect to the dependent variable is chosen. A typical output from these analyses is shown in Figure 7-6.

A. Including all environmental factors



B. Excluding land-use characteristics

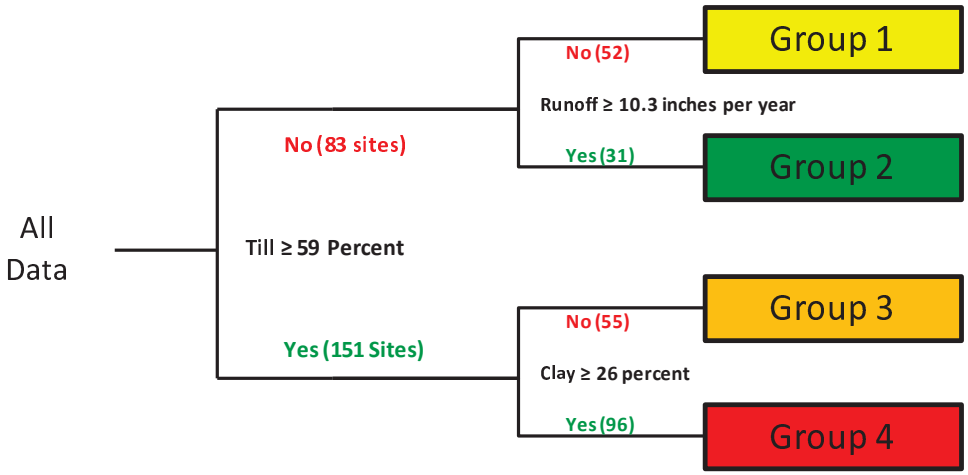


FIGURE 7-6

Results of CART Analysis of Total Phosphorus Resulting in Different Phosphorus Groups Using (A) All Environmental Predictors, or (B) Excluding Land-Use Predictors

Source: Robertson et al. (2001; Figure 19).

In general, CART can be applied effectively to classification or normalization of geographical data and in the development of stressor-response relationships from other field study data. Be aware that sampling bias can be an important classifying variable and responsible for some splits. Of course, exposing these biases sooner rather than later is best. Splits in the biological response variable can also identify inflection points or nonlinearities in a stressor-response relationship. Generally, the first few splits of the data are the most reliable.

7.3.3.5. *Applications of Statistical Models*

Data from remotely sensed imagery can also be analyzed using statistical methods. Effect levels and exposure levels can be extracted from these statistical analyses and used to estimate risks, estimate benchmarks or goals, or provide evidence of one sort or another. They can be used to construct relatively simple empirical relationships or models linking landscape to *in situ* stressor or response metrics or used to develop *wall-to-wall* landscape coverages and other data to document natural and stressor gradients and exposure levels.

7.3.4. Common Spatial Analysis Methods

7.3.4.1. *Landscape Metrics*

Landscape metrics are used to quantify the spatial structure of categorical landscape data, primarily land use/landcover (LU/LC) data. Landscape metrics are most commonly applied to landscapes made up of homogenous patches where a patch is defined as a polygon representing the boundary of a given LU/LC class (e.g., forest, urban, agriculture). From these data, landscape metrics can be calculated for each individual patch, summarized for each LU/LC class, or summarized for the landscape as a whole. In the context of water quality applications, class level metrics (e.g., total urban area or proportion of stream buffer that is forested) tend to have the widest use. As such, much of the following discussion focuses on class level metrics. For a detailed review of all types of metrics and background on landscape pattern metrics, see McGarigal et al. (2002).

Class level landscape metrics quantify two aspects of landscape structure: composition and configuration. Composition refers to the amount of a given class and measures aspects such as class abundance (e.g., total area or total proportion) or class diversity (e.g., total number of patches). Configuration describes the spatial character of individual patches and the spatial arrangement of those patches. Metrics of configuration measure aspects such as the distribution of patch size, core area, isolation, and connectivity (McGarigal et al., 2002).

7.3.4.2. *Specific Tools*

Many tools exist that help calculate hundreds of different landscape metrics. Two good examples of these software tools are FRAGSTATS 3.3 and ATtILA (Analytical Tools Interface for Landscape Assessments).

- FRAGSTATS 3.3—FRAGSTATS is an open source, stand-alone tool that calculates hundreds of different metrics and accepts raster input data sets in many formats including Arc Grid, ASCII, binary, and ERDAS.
- ATtILA—is an ArcView 3.1 extension that calculates many common metrics via a graphical user interface. ATtILA requires ArcView 3.1 and the Spatial Analyst extension.

7.3.4.3. *Important Considerations*

Calculating these metrics requires a basic level of skill with some software tools or GIS experience; however, the most important consideration in using landscape metrics is selecting appropriate metrics and interpreting the ecological and biological meaning of the metric. For a comprehensive list of caveats, see McGarigal et al. (2002).

Selecting appropriate metrics—Although it is possible to calculate hundreds of metrics, it is not necessarily desirable to use all those metrics because many of the metrics are highly correlated. Many papers have been devoted to methods for evaluating and choosing appropriate landscape metrics (e.g., Riitters et al., 1995; Gustafson, 1998).

Interpreting the ecological/biological meaning of the metrics—When calculating landscape metrics and associating those metrics with ecological or biological endpoints, it is imperative to have a sound conceptual basis for making those associations. As is often the case with water quality applications, landscapes, and the patterns of those landscapes are not the causes of impairment or even the source of the cause. Rather, sources of pollutants and other causes of impairment tend to be associated with certain land uses. For instance, total urban land itself might not necessarily increase sediment loads to streams, yet the configuration of the urban lands (e.g., proximity to streams), the intensity of the urbanization, typical current practices, infrastructure such as storm drains, and physiographic setting interact to determine the amount of sediment that ultimately reaches a stream.

7.3.4.4. *Spatial Interpolation*

Spatial interpolation is a technique that predicts values at unsampled locations on the basis of the values at sampled locations. Our ability to interpolate is based on Tobler's law, which simply states that features close to one another are more alike than features far apart (for an interesting review of Tobler's law, see Miller, 2004). There are numerous techniques for both vector and raster formats that span a range of sophistication. Some of the more commonly encountered vector interpolation methods are briefly described below.

- **Thiessen/Voronoi Polygons**—Thiessen/Voronoi polygons are relatively simple vector based methods to estimate a polygon surface from point data. Point data with a sampled value is required to create Thiessen/Voronoi polygons. Boundaries for each polygon are the halfway point to all other neighboring points and the value for the resulting polygon is assumed to be that of the central point (i.e., Thiessen/Voronoi polygons make the assumption that any point on a surface will have the value of the closest sampled point). For an example of a Thiessen/Voronoi polygon surface, see Figure 7-7.

- **Triangulated Irregular Networks (TINs)**—TINs are vector-based data models that capture information about a surface with points, connections of points and the resultant triangular face (see Figure 7-8). The entire TIN is represented by a complex network of triangles. The most common usage of TINs has traditionally been capturing elevation data, but any continuous surface can be represented by a TIN. The triangular faces (i.e., facets) provide information such as the slope,

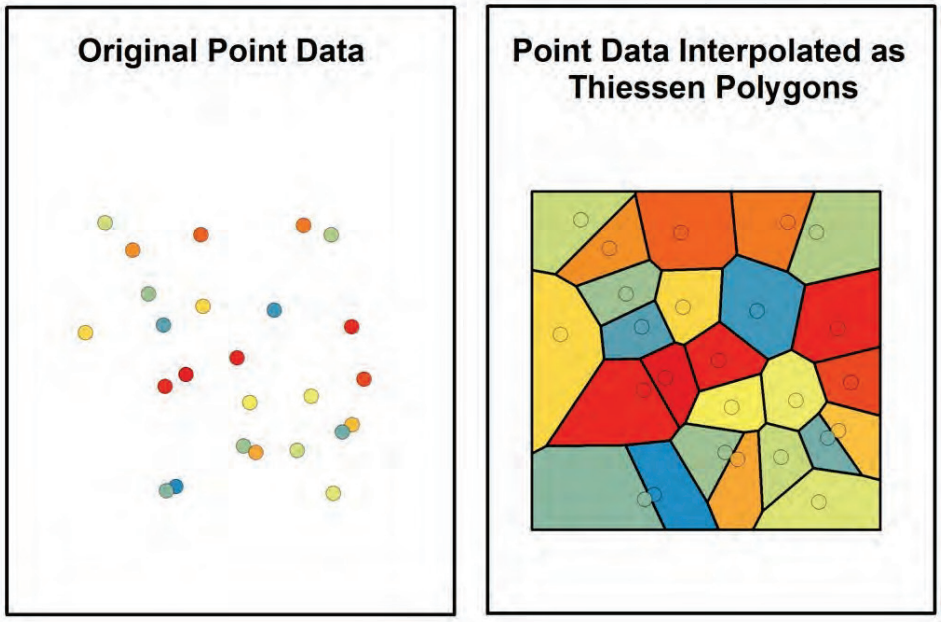


FIGURE 7-7

Example Thiessen/Voronoi Polygon Surface

Source: ESRI (2006).

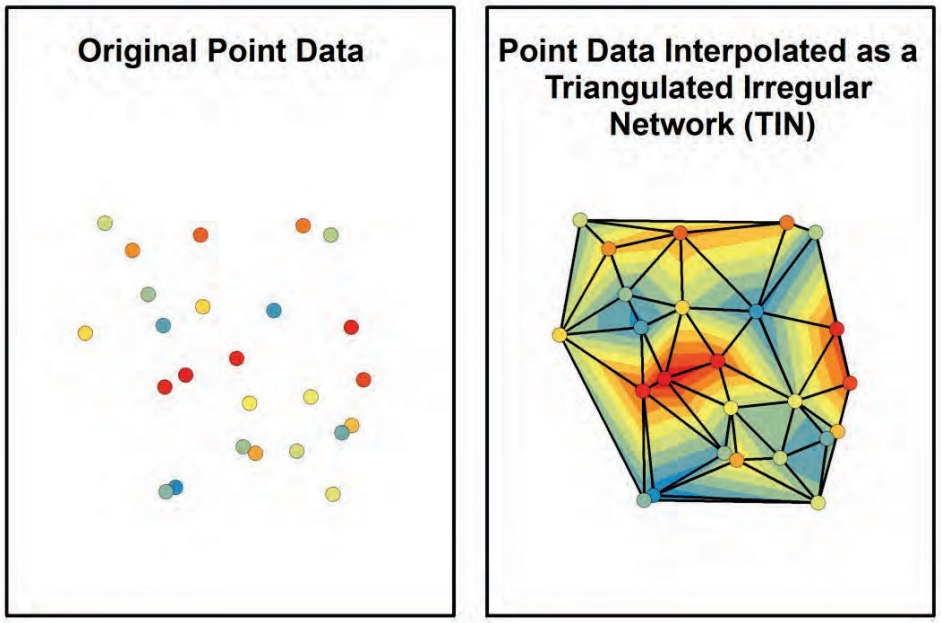


FIGURE 7-8

Example Triangulated Irregular Network

aspect, and surface area of the facet as well as interpolated values of any point within the facet.

Raster-based interpolation techniques are generally classified into two broad categories:

- **Deterministic**—Deterministic functions use values from sampled points and predict at unsampled locations on the basis of smoothing functions, similarity between sample points, or the relative locations of those points. Common deterministic techniques are Inverse Distance Weighting, Radial Basis Functions (e.g., spline), and polynomial functions.
- **Geostatistical**—Geostatistical methods, also known as Kriging, use statistical properties (i.e., spatial autocorrelation) of the sampled points to predict at unsampled locations. Furthermore, because geostatistical methods are based on statistical models, it is possible to generate prediction error surfaces that give an indication of accuracy across the predicted surface.

7.3.4.5. *Specific Tools*

Although many tools exist to facilitate different interpolations, two of the more commonly used are ArcGIS and Geostatistical Analyst.

- **ArcGIS**—ArcGIS contains very widely used commercial GIS data and provides tools for conducting both vector and raster interpolation techniques. Additional extensions (e.g., Geostatistical Analyst) provide more complete and robust interpolation capabilities.
- **Geostatistical Analyst**—Geostatistical Analyst is an ArcGIS extension that was created specifically for raster interpolations and focuses on the family of geostatistical methods (e.g., ordinary kriging, universal kriging, indicator kriging, cokriging). In addition, Geostatistical Analyst includes numerous deterministic methods and useful tools for exploratory data analysis and evaluating interpolated surfaces.

7.3.4.6. *Important Considerations*

Spatial interpolations are useful methods for extending the spatial coverage of sampled data and for visualizing variation in spatial data. In a water quality context, spatial interpolations might be used to provide an estimated coverage of precipitation on

the basis of rain gauge locations or to predict certain water quality parameter throughout a waterbody (e.g., a lake or estuary) on the basis of sampling locations. It is important to remember that conducting spatial interpolations correctly depends on a number of very important considerations.

Appropriate sampling designs are imperative—Although any data set with continuous data can be interpolated, not all sampling designs provide sufficient information for the resultant interpolation to have much meaning. Any sampling design can be used, but uniform sampling designs are often preferred. However, the most important aspect of the design is not necessarily the points' selection but the density of those points. The density must adequately capture the spatial variation of the phenomena being measured.

Interpolated surfaces are predictions and have error—Interpolated surfaces are relatively easy to create and display with modern geospatial tools. The display and use of these products can unintentionally imply a certain degree of certainty about the data they represent. In some cases, that might be fair, and in other cases, not. It is imperative that users and analysts of interpolated data understand the data used to generate the surface and, even more importantly, understand that there is error associated with the final predicted surface.

7.3.5. Hydrologic Analysis

Hydrologic analysis is the analysis of topographic data, usually with raster digital elevation models (DEMs), to delineate and quantify hydrologic features and networks. Hydrologic modeling represents a broad set of techniques that examine elevation data to determine how water flows across a landscape, delineate watershed boundaries, and identify likely flow paths.

7.3.5.1. *Specific Tools*

Tools and methods for conducting terrain analysis are standard features of most GIS software. There are numerous examples of applications using commercial software (i.e., ArcGIS) and open source software (i.e., GRASS GIS).

A related set of tools analyze networks. A primary example of a hydrologic network is the National Hydrography Dataset (NHD). The NHD is a vector data set that combines information about hydrology with the NHD vector data to form networks that follow hydrologic flow paths and allows for complex hydrologic modeling.

7.3.5.2. *Important Considerations*

Hydrologic analysis is based on analysis of DEMs; thus, the quality of the resultant models and features is directly tied to the resolution of the elevation models and the variability of the terrain that is being modeled. For instance, modeling hydrology in mountainous regions with significant relief might successfully be accomplished with coarser-resolution DEMs because slope, aspect, and flow direction might be adequately estimated. In relatively flat areas (e.g., coastal plains), coarse-resolution DEMs will fail to capture the fine-scale variation that drives hydrology in those regions.

7.3.6. Overlays and Proximity

Overlay and proximity analyses are good examples of traditional GIS analytical approaches. Overlay analysis examines spatial relationships between features represented in multiple layers. For instance, an overlay analysis could be conducted to determine how ecoregions and watersheds spatially interact. An analyst could determine which watersheds occur in only a single ecoregion or which watersheds span multiple ecoregions (for more information on ecoregions and watersheds, see Chapter 5). Proximity analysis simply examines the proximity of one feature to another. A common use of proximity analysis is to identify nearest neighbors or to conduct a buffer analysis. For example, an analyst could use proximity analysis to identify point sources of a pollutant that are closest to a water quality sampling station.

7.3.6.1. *Specific Tools*

Many overlay and proximity analyses can be conducted with either vector or raster data. Although most GIS applications will have the capability to conduct these analyses, the tools listed below are specific to ArcGIS (ESRI, 2006).

Overlay and Proximity Tools—There are numerous types of overlay and proximity methods. The most common vector methods are the following:

- Union: Merges all features of two input data sets into one output data set. Features from both data sets are preserved.
- Intersect: Merges intersecting features of two input data sets into one output data set. Only features common to both data sets are preserved.
- Identity: Merges intersecting features of two input data sets into one output data set. All features of the input data set and intersecting features of the identity data set are preserved.
- Buffer: Generates a polygon that represents the area contained within a given distance from an input feature.
- Near/Point Distance: Calculates the distance from point features in one data set to features in another data set.

Some common raster methods are the following:

- Map Algebra: Raster overlay methods are mostly accomplished by using map algebra. Map algebra is a general language that allows the cell-by-cell analysis of multiple input rasters. A common example would be the addition of multiple rasters to create an output raster representing the sum of the two raster data sets.
- Distance tools: Most raster-based proximity tools are contained in the broad class of distance tools that include Euclidean distance and cost distance. Euclidean distance calculates the straight line distance of each pixel in the output data set to all feature or pixels in an input data set. Cost distance is similar, but distances are constrained by a cost surface that results in paths that are not necessarily straight lines (e.g., topography as a cost surface would define paths along hydrological flow paths).

7.3.6.2. *Important Considerations*

Two considerations with conducting overlays are the quality of the data and error propagation.

Data quality—As with all spatial analysis techniques, the quality of the analysis is a function of the quality of the input data; however, this is especially true with overlay

techniques because spatial error in the location of line and point features and polygon boundaries will have a profound effect on how features overlay. On a similar vein, the scale of each input data set will affect the result and should be as closely matched as is possible. Poor spatial accuracy or mismatched scales would preclude conducting any overlay analyses.

Error propagation—Related to data quality is how error in multiple input data sets propagates to the final output data set. For example, in studies of LU/LC change, the general rule is that the output change data set can be no more accurate than the product of the input LU/LC data sets. For instance, two data sets with accuracies of 90 and 85%, would create an output LU/LC change data set with accuracy no better than 76.5%. Furthermore, it would be unlikely that the accuracy of the output data would be consistent across the full extent of the data and that the error would also vary spatially. It is often difficult to fully quantify how error propagates, but the magnitude of error in the results is certainly a concern in interpreting data resulting from one or more overlay analyses.

7.3.6.3. *Applications of Statistical and Spatial Models*

Alone or in combination, spatial coverages can be used in many ways as illustrated in the examples in Section III of this document and the many Web sites provided in Appendices A and B. A causal relationship is at the core of each one. The spatial and statistical models attempt to characterize that relationship to better understand what is going on in the environment. The information can be used to do the following:

- Develop evidence of a cause or source.
- Predict effects or exposures that are estimated to yield desired results.
- Extrapolate to areas lacking *in situ* data.
- Evaluate performance and effectiveness.

7.4. SYNTHESIS

Synthesis brings together the results from the analysis to generate the findings of the assessment in a useful form for the decision. Synthesis is devoted to producing a coherent output that integrates all evidence and endpoints to inform the decision maker. This includes deriving endpoint estimates and associated uncertainties from the results of the analysis, integrating multiple forms of evidence, comparing the management alternatives, and deriving overall results. The methods and criteria for syntheses vary with the type of assessment, but they share similar processes of estimation, integration, comparison, and characterization (U.S. EPA, 2006; 2011, Cormier et al, 2008; 2010; Suter and Cormier, 2008a; Suter and Cormier, 2011).

7.4.1. Decision Support Systems (DSSs)

Decision support systems (DSSs) can help characterize the exposure or effect that occurred, is occurring, could occur, or is desired to occur. The can help organize information and thinking so that the resulting relationships extrapolate to sites lacking *in situ* data or to inform decision making. They help combine and interpret finding to answer questions posed in planning and formulation that assess ecosystem condition; causes of impairment; sources of stressors; risks from known or expected exposures; exposure that will be protective; optimization of type, placement, and deployment of management actions and best management practices (BMPs); performance of BMPS and controls; and effectiveness of management actions and controls to resolve the environmental problem. They can also help to evaluate and articulate uncertainties and potential for false negatives and false positives (Type 1 and Type 2 errors). They are particularly helpful for reporting findings and ultimately compelling decisions and actions. They can be used to perform stakeholder and peer review of the products but these activities should also occur at the beginning of a project or program. DSSs can be used to communicate results to the decision maker and publish results.

A DSS is a system for helping to choose among alternatives. An environmental DSS can be described as providing a complete project management system for environmental decision problems that is composed of the following components:

- Guidance for each aspect and function of the DSS. This includes interpreting results and explanation of technical terms and methods.
- Access to and integration of project-specific knowledge bases with further access to the wealth of information available on the Web.
- Database management including SQL queries and GIS access.
- Environmental modeling capability using risk assessment, and statistical and decision analysis tools.
- Expert system components that help the user navigate the technical choices available within the DSS analysis tools (e.g., risk assessment, financial and social options, or statistical and decision analysis options).
- A presentation system that can be tailored to the specific needs of the users.
- A document production system that can be tailored to any form of computational output (e.g., Web-based, PDF, Office products).
- Quality assurance (QA) that is continuously measured and evaluated through user supplied feedback as well as more traditional QA techniques.
- Interactive training in each aspect of the DSS (Black and Stockton, 2009).

In addition to DSS targeted for environmental assessment there are many open source DSS software. GIS can be used in DSS to facilitate integrating different types of analyses and different types of environmental assessments thus leading to environmental problem solving.

7.4.2. Some Specific Decision Support Systems (DSSs)

The EPA's Regional Vulnerability Assessment (ReVA) program is a DSS for regional-scale, priority setting being developed by EPA's Office of Research and Development. Extensive effort has been made to evaluate environmental condition and known stressors within the Mid-Atlantic region, but predicting future environmental risk to prioritize efforts to protect and rehabilitate environmental quality efficiently and effectively is still difficult. ReVA is being developed to identify those ecosystems most vulnerable to being lost or permanently harmed in the next 5 to 25 years and to determine which stressors are likely to cause the greatest risk. The goal of ReVA is not

exact predictions but, rather, identification of the undesirable environmental changes expected over the coming years. The ReVA program extends environmental assessments for the region by using integrative technologies to predict future environmental risk and support informed proactive decision making and prioritization of issues for risk management.

CADDIS, is another type of DSS that provides an organization process for performing causal assessments along with information about commonly encountered causes of ecological impairments, statistical software and guidance for use and interpretation in the context of causal assessment.

Some DSSs are developed for a particular geographic area. For example, the Tennessee Valley Authority developed The Integrated Pollutant Source Identification, a geographic database and set of tools for designing and implementing water quality improvement and protection projects within a watershed. See Chapter 10 for details and Chapter 11 for an example of its implementation.

7.5. DECISIONS AND ACTIONS

At the end of each assessment, a decision is made. The decision can be (1) to stop the assessment process because there is no further problem; (2) to perform an assessment informed management action; (3) to initiate the next assessment in the sequence; or (4) to bypass the next assessment and proceed to a another type of assessment. Alternatively, although not a preferred option, a decision can be made without using the information offered by the assessment (e.g., if the assessment results suggest a politically or economically unacceptable conclusion [NRC, 2005]) (Cormier and Suter, 2008).

7.6. REFERENCES

Black, P. and T. Stockton. 2009 Basic steps for the development of decision support systems. In: Decision Support Systems for Risk-based Management of Contaminated Sites. A. Marcomini, G.W. Suter II, A. Critto, Ed. Springer, New York. pp. 1–28.

Cormier S.M, G.W. Suter II, and S.B. Norton. 2010. Causal Characteristics for Ecoepidemiology. Hum. Ecol. Risk Assess. 16(1): 53–73.

Cormier S.M., J.F. Paul, R.L. Spehar, P. Shaw-Allen, W.J. Berry, and G.W. Suter, II. 2008. Using field data and weight of evidence to develop water quality criteria. Integr. Environ. Assess. Manag. 4(4):490–504.

Cormier, S.M. and G.W. Suter II. 2008. A framework for fully integrating environmental assessment. Environ. Manage. 42(4):543–556. Available online at http://www.springerlink.com/content/h56531j12q33776t/fulltext.pdf.

Davies, S.P., and S.K. Jackson. 2006. The biological condition gradient: a descriptive model for interpreting change in aquatic ecosystems. Ecol. Appl. 16(4):1251–1266.

De'ath, G., and K.E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology. 81(11):3178–3192.

Efroymson, R.A., J.P. Nicollette, and G.W. Suter, II. 2004. A framework for net environmental benefit analysis for remediation or restoration of contaminated sites. Environ. Manage. 34(3):315–331.

ESRI (Environmental Systems Research Incorporated). 2006. ArcGIS 9.2 Desktop Help. ESRI, Redlands, CA. Available online at http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=welcome. Last modified March 15, 2007.

Gustafson, E.J. 1998. Quantifying landscape spatial pattern: What is the state of the art? Ecosystems 1:143–156.

Hanley N, and C.L. Spash. 1993. Cost-Benefit Analysis and the Environment. Edward Elgar Publishing, Cheltenham, UK.

Hollister, J.W., H.A. Walker and J.F. Paul. 2008. CProb: A Computational Tool for Conducting Conditional Probability Analysis. J. Environ. Qual. 37(6):2392–2396.

McGarigal, K., S.A. Cushman, M.C. Neel, and E. Ene. 2002. FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps. Computer software program produced by the authors at the University of Massachusetts, Amherst. Available online at www.umass.edu/landeco/research/fragstats/fragstats.html.

Linkov, I., F.K. Satterstrom, G. Kiker, et al. 2006. Multicriteria decision analysis: a comprehensive decision approach for management of contaminated sediments. Risk Anal. 26:61–78.

Michener, W. 2006. Meta-information concepts for ecological data management. Ecol. Inform. 1(1):3–7.

Michener, W.K. 2000. Metadata. In: Ecological Data: Design, Management, and Processing. W.K. Michener and J.W. Brunt, Eds. Oxford Press, Oxford, UK. p. 92–116.

Miller, H.J. 2004. Tobler's First Law and Spatial Analysis. Ann. Assoc. Am. Geogr. 94(2): 284–289.

NRC (National Research Council). 2005. Superfund and mining megasites: lessons from the Coeur D'Alene river basin. National Academies Press, Washington, DC.

Prasad, A.M, L.R. Iverson, and A. Liaw. 2006. Random forests for modeling the distribution of tree abundances. Ecosystems 9:181–199.

Riitters, K.H., R.V. O'Neill, C.T. Hunsaker et al. 1995. A factor analysis of landscape pattern and structure metrics. Landsc. Ecol. 10:23–39.

Robertson, D.M, D.A. Saad, and A.W. Wieben. 2001. An alternative regionalization scheme for defining nutrient criteria for rivers and streams. United States Geological Survey (USGS) Water Resources Investigations Report 01-4073. United States Department of the Interior, USGS, Middleton, WI. Available online at http://wi.water.usgs.gov/pubs/wrir-01-4073/wrir-01-4073.pdf.

Senay, G.B., N.A. Shafique, B.C. Autrey, F. Fulk, and S.M Cormier. 2001. The selection of narrow wavebands for optimizing water quality monitoring on the Great Miami River, Ohio using hyperspectral remote sensor data. J. Spatial Hydrol. 1(1): 1–22.

Suter, G. W., II. 2007. Ecological Risk Assessment, Second Edition. Taylor and Francis, Boca Raton, FL.

Suter, G.W. II, and S.M Cormier. 2008a. A theory of practice for environmental assessment. Integr. Environ. Assess. Manag. 4(4)478–485. .

Suter, G.W. II, and S.M Cormier. 2008b. What is meant by risk-based environmental quality criteria? Integr. Environ. Assess. Manag. 4(4) 486–489.

Suter, G.W. II and S.M Cormier. 2011. Why and how to combine evidence in environmental assessments: weighing evidence and building cases. Sci Total Environ 409(8): 1406–1417.

U.S. EPA (Environmental Protection Agency). 2005. Use of Biological Information to Better Define Designated Aquatic Life Uses in State and Tribal Water Quality Standards: Tiered Aquatic Life Uses. U.S. EPA Office of Water, Washington, DC. EPA/822/R-05/001. Available online at http://www.epa.gov/bioiweb1/pdf/EPA-822-R-05-001UseofBiologicalInformationtoBetterDefineDesignatedAquaticLifeUses-TieredAquaticLifeUses.pdf. Accessed 2 August 2008.

U.S. EPA (Environmental Protection Agency). 2006. Framework for Developing Suspended and Bedded Sediments Water Quality Criteria. Office of Water, Washington, DC. EPA/822/R/06/001. Available online at http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=164423.

U.S. EPA (Environmental Protection Agency). 2010a. CADDIS (The causal analysis/diagnosis decision information system). Available online at: www.epa.gov/caddis (accessed 6/6/11).

U.S. EPA (Environmental Protection Agency). 2010b. Integrating Ecological Assessment and Decision-Making at EPA: A Path Forward: Results of a Colloquium in Response to Science Advisory Board and National Research Council Recommendations. Risk Assessment Forum, Washington, DC. EPA/100/R-10/004. Available online at http://www.epa.gov/raf/publications/pdfs/integrating-ecolog-assess-decision-making.pdf.

U.S. EPA (Environmental Protection Agency). 2011. A Field-Based Aquatic Life Benchmark for Conductivity in Central Appalachian Streams. Office of Research and Development, National Center for Environmental Assessment, Washington, DC. EPA/600/R-10/023F. Available online at: http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=233809.

Wu, J., K.B. Jones, H. Li, and O.L. Loucks. 2006. Scaling and Uncertainty Analysis in Ecology: Methods and Applications. Springer, Dordrecht, The Netherlands.