

**July 2005**

Donald W. Meals and Steven A. Dressing. 2005. Monitoring data – exploring your data, the first step, Tech Notes 1, July 2005. Developed for U.S. Environmental Protection Agency by Tetra Tech, Inc., Fairfax, VA, 14 p. Available online at <https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/nonpoint-source-monitoring-technical-notes>.

Through the National Nonpoint Source Monitoring Program (NNPSMP), states monitor and evaluate a subset of watershed projects funded by the Clean Water Act Section 319 Nonpoint Source Control Program.

The program has two major objectives:

1. To scientifically evaluate the effectiveness of watershed technologies designed to control nonpoint source pollution
2. To improve our understanding of nonpoint source pollution

NNPSMP Tech Notes is a series of publications that shares this unique research and monitoring effort. It offers guidance on data collection, implementation of pollution control technologies, and monitoring design, as well as case studies that illustrate principles in action.

## Monitoring Data

### *Exploring Your Data, The First Step*

Now that your monitoring program is up and running, it is time to evaluate the data. If you designed your monitoring program carefully (Tech Note #2), you will have the right kinds of data collected at appropriate times and locations to achieve your objectives. At the start, you should check your data for conformity with original plans and quality assurance/quality control (QA/QC) procedures. Use the Quality Assurance Project Plan (QAPP) you developed as a guide.

After you accept the dataset, you will still be faced with a challenge: What do I do with all these numbers? What do they mean? How do I start to make sense of them? The first step in answering such questions lies in exploratory data analysis (EDA). The purpose of EDA is to get the “feel” of your data, to begin to develop some ideas about what they can tell you and how you can draw some conclusions about them. Think of EDA as detective work—your job is to sift through all the facts, look for clues, and put the pieces together to find suggestions of meaning in the data.

This process of data exploration differs from traditional hypothesis testing. Testing of hypotheses always requires some initial assumption or prediction about the data, such as “The BMP will reduce phosphorus loads.” Although formulating and testing hypotheses is the foundation of a good monitoring program, the first pass through your data should not be too narrowly focused on testing a single idea. EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow in favor of the more direct approach of allowing the data itself to reveal its underlying structure. EDA uses a variety of techniques, both numerical and graphical, to open-mindedly search for new, perhaps unexpected, insight into the data.

The overall goal of data exploration is to uncover the underlying structure of a dataset. The principles of data exploration apply not only to water quality data but also to

meteorologic, socioeconomic, land treatment, and other kinds of data. Your specific objectives for data exploration might include the following:

- to describe the behavior of one or more variables
- to find extreme values and anomalies
- to test distribution and assumptions of independence and constant variance
- to see cycles and trends
- to find clusters or groupings
- to compare two or more locations or time periods and
- to examine relationships between variables

Data exploration is a necessary first step in analyzing monitoring data; unless initial exploration reveals indications of patterns and relationships, it is unlikely there will be something for further analysis to confirm. J. W. Tukey (1977), the founder of exploratory data analysis, said, “EDA can never be the whole story, but nothing else can serve as the ...first step.”

## Steps in Data Exploration

Data exploration is a process of probing more deeply into the dataset, while being careful to stay organized and avoid errors. Here are some typical steps in the process of EDA (Jambu 1991), although not all of them may apply to every situation.

### *1. Data decision*

This step actually refers back to the design of your monitoring program, before you took your first sample, when you defined the goals and the scope of the project. At this point you decided what data were necessary to collect to achieve the goals.

### *2. Data conception/elaboration*

In this step you chose the procedures for collecting the data. You may have determined acceptable values for the measured variables so that data could be validated after collection. The preparation of a QAPP fulfills this step and provides a basis for assessing the quality of your data.

### *3. Data input*

Building a database and establishing procedures for data input and storage are important steps in ensuring the accuracy and integrity of your dataset. Spreadsheets and relational databases are useful tools for this task.

### *4. Data management*

In the process of working with your data, you will create files; update, check, and validate the files; and sort, merge, and aggregate the data in different ways. The importance of

data screening and validation cannot be overemphasized! **This should always be done before you embark on EDA.** Be as sure as possible that your data are free from entry errors, typos, and other mistakes before you proceed further. Few things are more frustrating than finding a misplaced decimal point, reversed digits, or a lab error **after** you have completed your data analysis!

### 5. *One-dimensional analysis*

Often, the first step in really exploring your data is simply to describe or summarize the information one variable at a time, independent of other variables. This can be done by using basic statistics on range, central tendency, and variability, or with simple graphs like histograms, pie charts, or time plots. This kind of information is always useful to put your data in context, even if you pursue more intensive statistical analysis later.

### 6. *Two-dimensional analysis*

Relationships between two variables are often of great interest, especially if you suspect that there is some meaningful connection, such as between suspended sediment and phosphorus, or some cause and effect process, such as between rainfall and streamflow. Graphical techniques like scatter plots and numerical techniques like correlation are often used for this purpose.

There are more advanced steps in data exploration, including analysis of multiple variables, factor analysis, and cluster analysis, but these techniques are beyond the scope of this publication. Consult some of the references listed at the end of this publication for additional information.

## Decisions

The goals of your project and the type of monitoring you are doing should guide your exploration. If you are monitoring at a single point while upstream best management practices (BMPs) are implemented gradually, you may be most interested in looking for trends. If you are sampling for phosphorus above and below a land treatment area, you might want to compare phosphorus concentrations at the two stations. If you are monitoring an erosion problem, you may want to see if there is a relationship between streamflow and suspended solids concentrations or look at how that relationship changed following land treatment. In most cases, you will want to look at the distribution of the data to determine whether the data satisfy statistical assumptions or whether a transformation is needed before further analysis.

There are two general approaches to EDA—quantitative (numerical) techniques and graphical approaches. The following sections present some specific techniques for exploring your data.

## Quantitative (Numerical) Techniques

When deciding how to analyze any dataset, it is essential to consider the characteristics of the data themselves. These characteristics will determine your choice of appropriate analysis procedures. Some common characteristics of water quantity and quality data (Helsel and Hirsch 1992) are as follows:

- A lower bound of zero—no negative values are possible
- The presence of outliers, extreme low or high values that occur infrequently but usually somewhere in the dataset (outliers on the high side are common)
- Skewed distribution due to outliers
- Nonnormal distribution
- Censored data—concentration data reported below some detection limit
- Strong seasonal patterns
- Autocorrelation—consecutive observations strongly correlated with each other
- Dependence on other uncontrolled or unmeasured variables (values strongly vary in concert with such variables as streamflow, precipitation, or sediment grain size).

Describing and summarizing your data in a way that conveys their important characteristics is one purpose of EDA. Evaluation of characteristics like nonnormal distribution and autocorrelation will help you determine the appropriate statistical analysis down the road.

## Evaluation of One Variable

Some numerical approaches to evaluate single variables are described below. Because these quantitative techniques are readily calculated by most spreadsheet programs and statistical software, this publication will not go into detail on the calculation of these statistics.

### Measures of Central Tendency

- The **mean** is computed as the sum of all values divided by the sample size. The mean is probably the most common data summary techniques in use, but it has one serious problem: an outlier (either high or low) has much greater influence on the mean than does a more “typical” value. Because of this sensitivity to extremes, the mean may not be the best summary of the central tendency of your data.
- The **median**, or 50th percentile, is the central value of the distribution when the data are ranked in numerical order. The median is the data value for which half of the observations are higher and half are lower. Because it is determined by the order of observations, the median is only slightly affected by the magnitude of a single extreme value. When a summary value is desired that is not strongly influenced by a few extremes, the median is preferable to the mean.

## Measures of Spread

- The sample **variance** and its square root, the **standard deviation**, are the most common measures of the spread (dispersion) of a set of data. These statistics are computed using the squares of the difference between each data point and the mean, so that outliers influence their magnitudes dramatically. In datasets with major outliers, the variance and standard deviation may suggest a much greater spread than exists for the majority of the data.
- The **coefficient of variation (CV)**, defined as the standard deviation divided by the mean, is a relative measure of the variability (spread) of the data. The CV is sometimes expressed as a percent, with larger values indicating high variability around the mean. Comparing the CV of two data groups can suggest their relative variability.
- The **interquartile range (IQR)** is defined as the 75th percentile minus the 25th percentile. Because it measures the range of the central 50 percent of the data, it is not influenced at all by the 25 percent of the data on either end and is relatively insensitive to outliers.

## Skewness

Water resources data are usually skewed, meaning that the data values are not symmetric around the mean or median, as extreme values extend out further in one direction. Streamflow data, for example, are typically right-skewed because of occasional high-flow events (Figure 1). When data are skewed, the mean is not equal to the median, but is pulled toward the long tail of the distribution. The standard deviation is also inflated by the extreme values. Because highly skewed data restrict the ability to use hypothesis tests that assume the data have a normal distribution, it is useful to evaluate the skewness of your data. The **coefficient of skewness (g)** is a common measure of skewness; a right-skewed distribution has a positive g and a left-skewed distribution has a negative g.

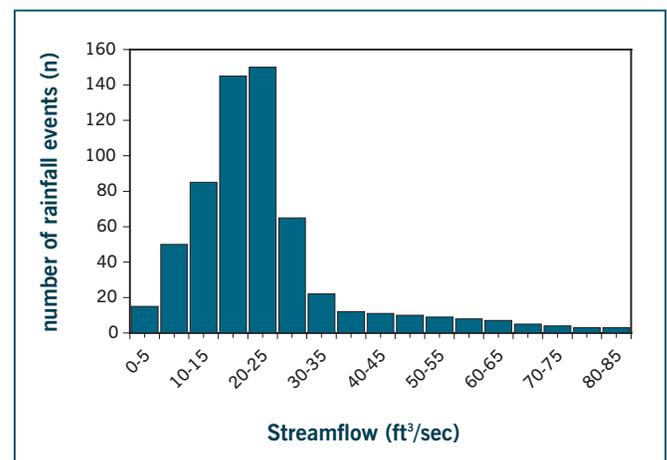


Figure 1. Right-skewed distribution

## Outliers

Outliers, data points that differ greatly from most others in the dataset, may cause some concern, but they can be the most important points in the dataset and always demand additional investigation. Of course, outliers can result from measurement or recording errors, and this should be the first thing you check. **If no error can be found, an outlier**

should never be rejected just because it appears unusual or extreme! Outliers may contain important information. Extreme high flows strongly influence stream morphology. Occasional extreme high concentrations of a toxic substance or low levels of dissolved oxygen may control the fish or invertebrate community of a stream. Graphical methods discussed later in this publication are helpful in identifying outliers.

## Distribution

Many common statistical techniques for hypothesis testing (parametric tests) require, among other characteristics, that the data be normally distributed, that is, the classic bell-shaped curve that is symmetrical around the mean. There are a number of statistical tests to evaluate how closely your data match an ideal normal distribution, generally referred to as goodness-of-fit tests. The  $W$  statistic of Shapiro and Wilk and the Kolmogorov-Smirnov  $KS$  test are two examples available in many statistical software packages.

## Transformations

Your evaluations so far may suggest that your data do not conform to an ideal normal distribution; this situation is quite typical of water quality data. If you plan to apply parametric statistical techniques such as the Student's  $t$ -Test, analysis of variance, or regression, you may wish to look at data transformations at this point. Transformation consists of applying the same mathematical operation (e.g., taking the logarithm or square root) to all the data points. Transformations are often used to make a data distribution more symmetric or more constant in variance, or to obtain a straight-line relationship between two variables. Because water resource data are often highly right-skewed, a  $\log_{10}$  transformation can often achieve these goals; such a distribution is referred to as **lognormal**. Subsequent analysis of log-transformed data must be done with care, as quantities such as mean and variance calculated on the transformed scale are often biased when transformed back to the original scale. The geometric mean (the mean of the log-transformed data back-transformed to the arithmetic scale), for example, differs from the mean of the untransformed distribution. Furthermore, results of statistical analysis can be difficult to understand or interpret when expressed on the transformed scale.

You should not assume that a transformation will solve all the problems with your data distribution; always test the characteristics of your transformed dataset again. Violations of the assumption of a normal distribution can lead to incorrect conclusions about the data when parametric tests are used in subsequent hypothesis testing. If transformed data cannot satisfy the assumptions of parametric statistical analysis, you will need to consider nonparametric techniques for data analysis. Although not generally as powerful as parametric tests, nonparametric methods do not require the data to come from a particular distribution. Nonparametric techniques are often suitable for analysis of nonpoint source data that include extreme outliers, missing values, irregularly spaced observations, and censored data.

## Evaluation of Two Variables

Many numerical techniques are available to examine and test the relationship between two or more variables. In EDA, the simplest technique is **correlation**, which measures the strength of an association between two variables. The most common measure of correlation is Pearson's  $r$ , also called the linear correlation coefficient. If the data lie exactly on a straight line with positive slope,  $r$  will equal 1; if the data are entirely random,  $r$  will equal 0. Other measures of correlation that are less sensitive to outliers include the nonparametric Kendall's *tau* and Spearman's *rho*. Measures of correlation are easily calculated by most statistical software packages.

It must be cautioned that whenever a numerical correlation is calculated, the data should be plotted in a scatterplot. Many different patterns can result in the same correlation coefficient. **Never compute a correlation coefficient and assume that the data follow a simple linear pattern!**

## Graphical Approaches

Because graphs summarize data in ways that describe essential information more quickly and completely than do tables of numbers, graphics are important diagnostic tools for exploring your data. There is no single statistical tool as powerful as a well-chosen graph (Chambers et al. 1983). Enormous amounts of quantitative information can be conveyed by graphs and the human eye-brain system is capable of quickly summarizing information, simultaneously appreciating overall patterns and minute details. Graphs will also be essential in ultimately conveying your project results to others. With the computers and software available today, there is simply no excuse for not graphing your data as part of EDA.

## Graphical Evaluation of One Variable

### Categorical Data

For categorical data such as the watershed area in different land uses or the number of aquatic macroinvertebrates in certain taxonomic groups, showing the data as frequencies in histograms or pie charts can effectively summarize data. Figure 2<sup>1</sup> shows a pie chart of the percent composition of orders of macroinvertebrates in a Vermont stream, clearly indicating that *Dipterans* dominate the community.

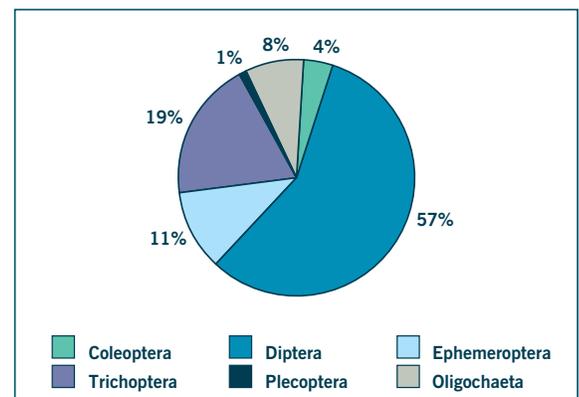


Figure 2. Percent composition of the orders of macroinvertebrates, Godin Brook, 2000.

<sup>1</sup> Data plotted in Figures 2 through 12 are taken from the Vermont NMP Project, *Lake Champlain Basin Agricultural Watersheds Section 319 National Monitoring Program Project: May 1994–November 2000* (Meals 2001).

## Histograms

Histograms are familiar graphs, where bars are drawn whose height represents the number or fraction of observations falling into one of several categories or intervals. Histograms are useful for depicting the shape or symmetry of a dataset, especially whether the data appear to be skewed. However, histograms have one major problem—their appearance depends strongly on the number of categories selected for the plot. For this reason, histograms are most useful to show data that have natural categories or groupings, such as fish numbers by species, but are usually inappropriate for data measured on a continuous scale such as streamflow or phosphorus concentration. Histograms can be useful to illustrate exceedances of standards, targets, and goals.

## Quantile Plots

Quantile plots (also called cumulative frequency plots) show the percentiles of the data distribution. Many statistics packages calculate and plot frequency distributions; instructions for manually constructing a quantile plot can be found in Helsel and Hirsch (1992) and other statistics textbooks. Quantile plots show many important data characteristics, such as the median or the percent of observations less than some critical threshold or frequency. With experience, an analyst can discern information about the spread and skewness of the data. Figure 3 shows a quantile plot of *E. coli* bacteria in a stream; the frequency of violation of the Vermont water quality standard can be easily seen (the standard was exceeded ~65 percent of the time).

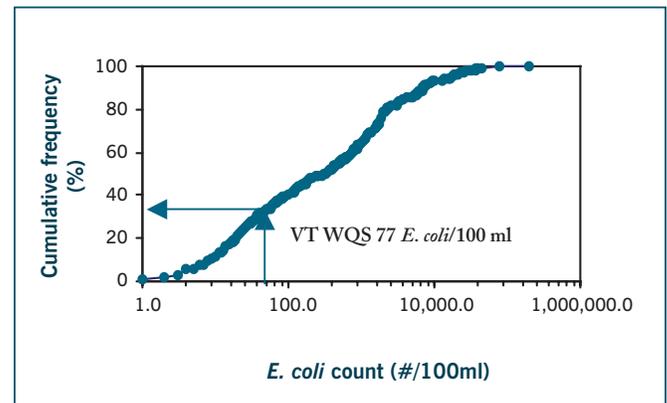


Figure 3. Cumulative frequency plot of *E. coli* data, Berry Brook, 1996.

## Boxplots

A boxplot presents a schematic of essential data characteristics in a simple and direct way: central tendency (median), spread (interquartile range), skewness (relative size of the box halves), and the presence of outliers. There are many variations and styles of boxplots, but the standard boxplot (Figure 4) consists of a rectangle spanning the 25<sup>th</sup> and 75<sup>th</sup> percentiles, split by a line representing the median. Whiskers extend vertically to encompass the range of most of the data, and outliers beyond this range are shown by dots or other symbols.

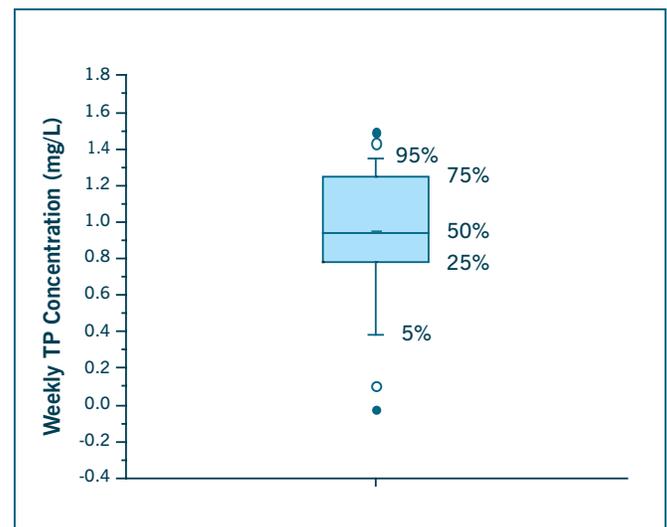


Figure 4. Boxplot of weekly total phosphorus (TP) concentration, Samsonville Brook, 1995.

of outliers may differ among graphing programs; standard definitions can be found in statistics textbooks (e.g., Cleveland 1993; Helsel and Hirsch 1992).

## Time Series

Monitoring data often consist of a series of observations in time, for example weekly samples over a year. One of the first—and the most useful—things to do with any time-series data is to plot them. Plotting time-series data can give you insight into seasonal patterns, trends, changes, and unexpected events more quickly and easily than tables of numbers. Here are some examples.

Figure 5 shows a time-series plot of *E. coli* counts in a Vermont stream. The extreme range of the counts (five orders of magnitude) and the pronounced seasonal cycle are readily apparent, with the lowest counts occurring during the winter. It is easy to see when the stream violates water quality standards for bacteria. Figure 6 shows a time-series plot of weekly total phosphorus (TP) concentration data from three stream stations. It is clear that around the middle of the year, something occurred that led to dramatic spikes in P concentration at Station 2, a phenomenon demanding further investigation. In this case, field investigation revealed concentrated overland flow from a new concentrated animal feed operation (CAFO) source. Finally, Figure 7 shows a time-series plot of total Kjeldahl nitrogen (TKN) data collected from three Vermont streams. Obviously, something happened around May 1996 that caused a major shift in TKN concentrations in all three streams. In addition, it is clear that after October, no values less than 0.5 mg/L were recorded. This shift was not the result of some activity in the watersheds but was an artifact of a faulty laboratory instrument, followed by the establishment of a lower detection limit of 0.50 mg/L. Discovery of this fault, while it invalidated a considerable amount of prior data, led to correction of the problem in the laboratory and saved the project major headaches down the road.

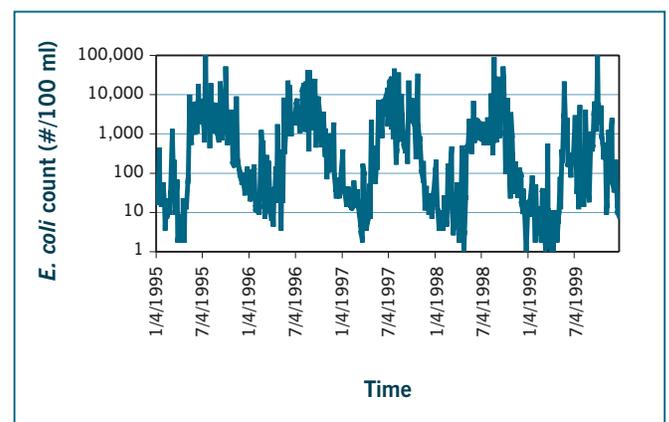


Figure 5. Time-series plot of weekly *E. coli* counts, Godin Brook, 1995–1999.

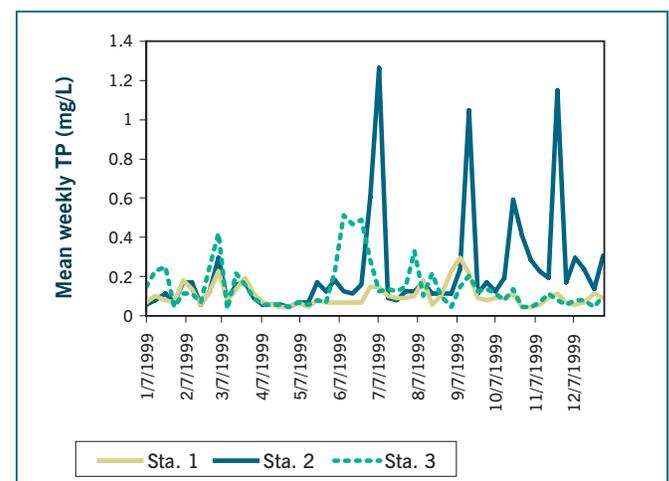


Figure 6. Time plot of weekly TP concentration, Godin Brook, 1999.

## Autocorrelation

Water resources data often exhibit autocorrelation (also called serial correlation) where the value of an observation is closely related to the observation immediately before it. Streamflow data often show autocorrelation, as high values tend to follow high values and low values follow low values. Because many hypothesis-testing statistical techniques require that data points be independent, it is useful to check your dataset for autocorrelation during EDA.

There are numerical techniques to test for autocorrelation, but a simple graphical method can suggest whether your data have significant autocorrelation—the **lag plot**. A lag plot is a graph where each data point is plotted against its predecessor in the time series, i.e., the value for day 2 and the value for day 1 are plotted as an x,y pair, then day 3/day 2, and so on. Random (independent) data should not exhibit any identifiable structure or pattern in the lag plot. Non-random structure in the lag plot indicates that the underlying data are not random and that autocorrelation may exist. Figure 8 shows a lag plot of weekly streamflow data, suggesting that moderate autocorrelation needs to be addressed. Autocorrelation can sometimes be reduced by aggregating data over different time periods, such as weekly means rather than daily values. Use of weekly means preserves much of the original information of a daily data series, but separates data points far enough in time so that serial correlation is reduced.

## Graphical Evaluation of Multiple Variables

Comparison of two or more variables is where EDA gets really interesting. This can mean comparing different datasets, such as stream nitrogen concentrations above and below a feedlot or phosphorus concentrations from a control and a treatment watershed, or comparing data from the same site over two different time periods, such as phosphorus loads from control versus treatment periods.

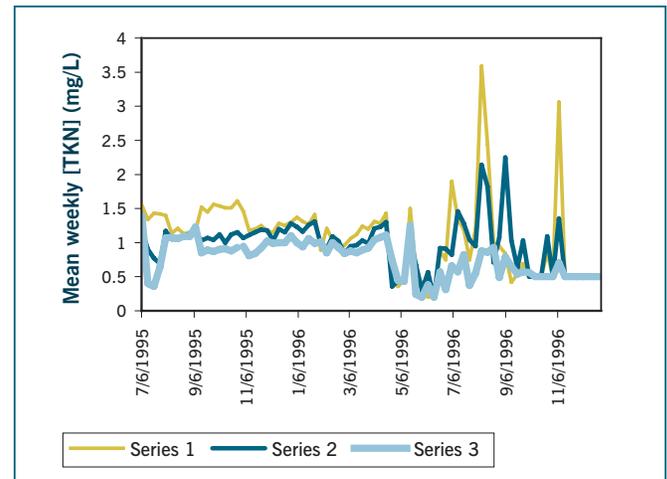


Figure 7. Time plot of TKN data from three stream stations, 1995–1996.

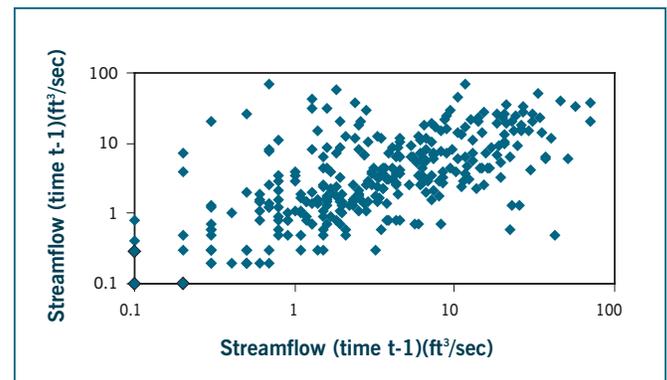


Figure 8. Lag-one plot of streamflow observations, Samsonville Brook, 1994.

Looking at how variables relate to each other, you can begin to think about causality, that is, is the behavior of one variable the result of action by another. For example, if variable B (e.g., total suspended solids) goes down as variable A (e.g., acres of irrigation management) goes up, has your BMP program improved water quality? While more rigorous statistical tests will be necessary to prove that hypothesis, graphical techniques of EDA can point you in the right direction.

## Boxplots

The characteristics that make boxplots useful for summarizing and inspecting a single dataset make them even more useful for comparing multiple datasets. The essential characteristics of numerous groups of data can be shown in a compact form. Boxplots of multiple data groups can help answer several important questions:

- Is a factor (location, period) significant?
- Does the median differ between groups?
- Does variation differ between groups?
- Are there outliers? Where?

Boxplots are helpful in determining whether central values, spread, symmetry, and outliers differ among groups. If the main boxes of two groups, for example, do not substantially overlap, you may have a reason to suspect that the two groups differ significantly. Interpretation of boxplots can help you formulate hypotheses about differences between groups or time periods that can be tested by other means later on. Figure 9 shows a box plot of total suspended solids (TSS) concentrations in three Vermont streams. The plot suggests that TSS concentrations may tend to be slightly lower at Station 3 compared to the other two stations.

## Scatterplots

The two-dimensional scatterplot is one of the most familiar graphical methods for data exploration. It consists of a scatter of points representing the value of one variable plotted against the value of another variable from the same point in time. Scatterplots illustrate the relationship between two variables. They can help reveal if there appears to be any association at all between two variables, whether the relationship is linear, whether different groups of data lie in separate regions of the scatterplot, and whether variability is constant over the full range of data. As mentioned above, associations between two variables

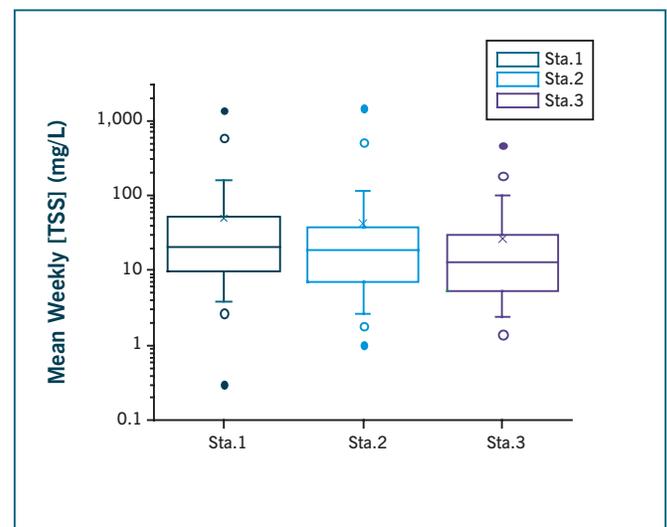


Figure 9. Boxplots of TSS concentration for three stream stations, 1998.

suggested by correlation coefficients should always be visualized using scatterplots.

Figure 10 shows a scatterplot of phosphorus export in a control and a treatment watershed in Vermont. Clearly, there is a strong positive association between total phosphorus (TP) levels in the two streams. This simple scatterplot indicates that it is worth proceeding with more rigorous statistical analysis to evaluate calibration between the two watersheds in a paired watershed design.

Figure 11 shows another scatterplot examining the relationship between streamflow and *E. coli* counts in another Vermont stream. In a nonpoint source situation, one might expect a positive association between streamflow and bacteria counts, as runoff during high flow events might wash bacteria from the land to the stream. Unfortunately, it does not require advanced statistics to conclude from Figure 11 that there is no such association (in fact, the correlation coefficient  $r$  is close to zero). However, this is not the end of the story. Remember that EDA involves an open-minded exploration of many possibilities. In Figure 12, the data points have been distinguished by season. The open circles represent data collected in the summer period, and there still appears to be little or no association between streamflow and *E. coli* counts. The solid circles, representing winter data, now appear to show some positive correlation ( $r = 0.45$ ) between streamflow and bacteria counts, with high bacteria counts associated with high flows. This picture suggests that something different is happening in winter compared to summer with respect to streamflow and *E. coli* in this watershed, a subject for further investigation.

In looking for correlations in scatterplots, choose your variables carefully. One common mistake is the comparison of variables that are already related by measurement or calculation. An example of such

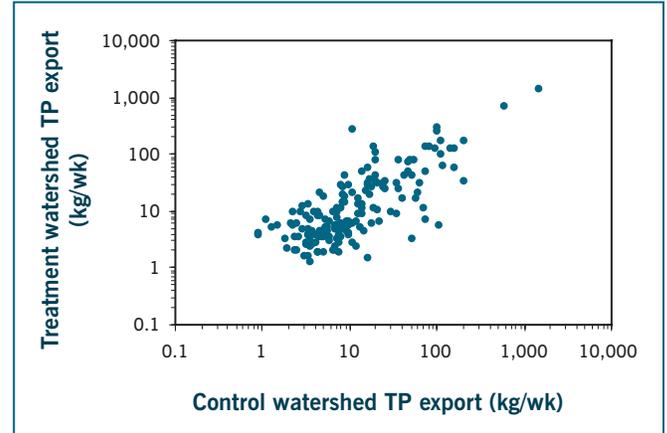


Figure 10. Scatterplot of weekly TP export from control and treatment watersheds, calibration period.

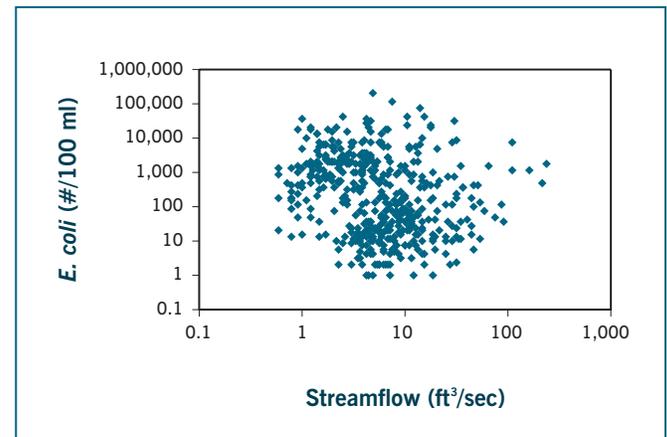


Figure 11. Scatterplot of *E. coli* vs. streamflow, Godin Brook, 1995–1998, all data combined.

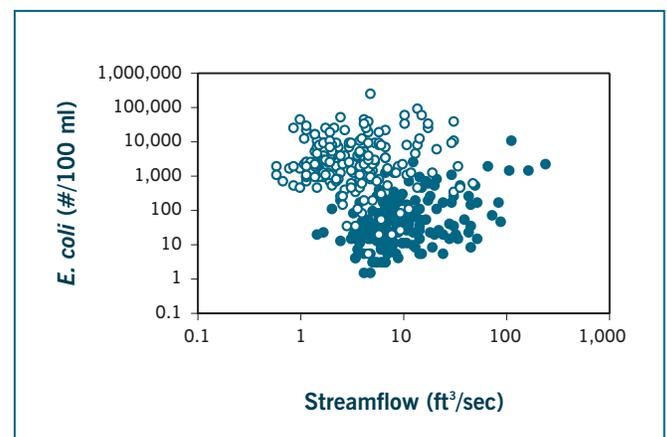


Figure 12. Scatterplot of *E. coli* vs. streamflow, Godin Brook, 1995–1998. Solid circles = winter, open circles = summer.

spurious correlation is the comparison of streamflow with load. Because load is calculated as concentration multiplied by flow, a scatterplot of flow versus load has a built-in correlation that means very little, even though it looks good in a scatterplot.

## Next Steps

The results of your data exploration—knowledge of how your data are distributed, their characteristics, and their relationships—will help you formulate hypotheses that can be tested with more advanced statistical techniques. Procedures like the Student's t-Test, analysis of variance (ANOVA), analysis of covariance (ANCOVA), and regression can be used to draw conclusions about your data and its meaning (USEPA 1997).

## Tools/Resources

With modern computers and software packages, there is really no excuse for avoiding EDA. Most standard spreadsheet programs include basic statistical functions and graphing capabilities. For more sophisticated and flexible EDA, you may wish to use more powerful statistical software. Many statistical packages are available; the table below lists some examples and Web sites to visit for more information.

Package Name	Web Site URL
Analyse-It (add in for MS Excel)	<a href="http://www.analyse-it.com">http://www.analyse-it.com</a>
DataDesk	<a href="http://www.datadesk.com">http://www.datadesk.com</a>
JMP	<a href="http://www.jmp.com/product/jmp_intro.shtml">http://www.jmp.com/product/jmp_intro.shtml</a>
MINITAB	<a href="http://www.minitab.com/products/minitab/default.aspx">http://www.minitab.com/products/minitab/default.aspx</a>
SAS/Stat, SAS/Insight	<a href="http://www.sas.com/technologies/analytics/statistics/index.html">http://www.sas.com/technologies/analytics/statistics/index.html</a>
SPSS	<a href="http://www.spss.com/spss/">http://www.spss.com/spss/</a>
SYSTAT	<a href="http://www.systat.com/products/Systat/">http://www.systat.com/products/Systat/</a>
WINKS	<a href="http://www.texasoft.com/">http://www.texasoft.com/</a>

## References and Recommended Texts

There is far more to EDA than is covered in this publication. For more details on the techniques discussed and information on other approaches, you may wish to consult these sources.

Chambers, J.M., W.S. Cleveland, B. Kleiner, and P.A. Tukey. 1983. *Graphical Methods for Data Analysis*. Wadsworth International Group, Belmont, CA. Boston: Duxbury Press.

Cleveland, W.S. 1993. *Visualizing Data*. Murray Hill, NJ: AT&T Bell Laboratories.

Friendly, M. 1995. *Exploratory and Graphical Methods of Data Analysis*. Statistical Consulting Service, York University, Toronto, Ontario, [www.math.yorku.ca/SCS/Courses/eda/](http://www.math.yorku.ca/SCS/Courses/eda/) [accessed 2/24/05].

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. New York: Van Nostrand Reinhold.

Helsel, D.R., and R.M. Hirsch. 1992. *Statistical Methods in Water Resources*. Studies in Environmental Science 49. New York: Elsevier. (available online as a pdf file at: <http://water.usgs.gov/pubs/twri/twri4a3/> [accessed 3/30/03]).

Jambu, M. 1991. *Exploratory and Multivariate Data Analysis*. Academic Press, Inc., Orlando, FL

Meals, D.W. 2001. *Lake Champlain Basin Agricultural Watersheds Section 319 National Monitoring Program Project, Final Project Report: May, 1994–November, 2000*. Vermont Department of Environmental Conservation, Waterbury, VT.

*NIST/SEMATECH e-Handbook of Statistical Methods*, [www.itl.nist.gov/div898/handbook/](http://www.itl.nist.gov/div898/handbook/) [accessed 2/24/05].

StatSoft, Inc. 2005. *StatSoft Electronic Textbook*. [www.statsoftinc.com/textbook/stathome.html](http://www.statsoftinc.com/textbook/stathome.html) [accessed 3/11/05].

Tufte, E.R. 1983. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley Publishing Co.

US Environmental Protection Agency, 1997. *Monitoring Guidance for Determining the Effectiveness of Nonpoint Source Control Projects*. EPA 841-B-96-004. U.S. Environmental Protection Agency, Office of Water, Washington, DC.

Velleman, P., and D. Hoaglin. 1981. *The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury Press.