Technical Guidance for Exploring TMDL Effectiveness Monitoring Data

1. Introduction

Effectiveness monitoring is a critical step in the Total Maximum Daily Load (TMDL) process for addressing water quality impairments in lakes, streams, and other water bodies. Since the overall goal of TMDL effectiveness monitoring is to identify water quality improvements that result from TMDL implementation, high quality data analysis is needed throughout all project phases. During project planning, water resource practitioners will find themselves making decisions that can be informed through statistical analysis of pilot data using ***Exploratory Data Analysis*** (EDA) techniques. EDA provides insight into the basic characteristics of individual datasets, as well as general relationships between two or more data groups.

EDA methods used to address project planning questions are also relevant when conducting analysis for interim and final reporting using ***Confirmatory Data Analysis*** (CDA) techniques. CDA is generally applied to confirm or reject assumed relationships between two or more data groups with a quantifiable level of confidence. A number of statistical tests are available for CDA. Although the selection of an appropriate test is largely based on the study design and objectives, characteristics of the data at hand must also be considered. Failure to do so can result in violation of basic assumptions of the statistical test and lead to conclusions that are, at best, poorly justified and, at worst, invalid. EDA techniques provide a straightforward means to evaluate such assumptions.

The primary purpose of this guide is to direct readers to EDA techniques that influence: 1) the selection of water quality parameters to include as part of TMDL effectiveness monitoring; 2) sampling frequency, timing, and duration; and 3) statistical test selection. Several EDA methods are demonstrated using stream monitoring data collected by the Ambient River and Stream Monitoring Program of the Washington Department of Ecology and screen captures from Microsoft Office Excel 2007 spreadsheet software. Specific topics covered in this guide include:

- Use of histograms to investigate data distributions and identify outliers;
- Transforming data with a skewed distribution;
- The use of time series plots to identify missing data and daily/seasonal fluctuations;
- Use of boxplots to further explore daily/seasonal fluctuations;
- Use of scatterplots to identify covariates;
- Use of lag plots to explore autocorrelation; and
- Analysis of censored data.

This guide offers just a small glimpse into exploratory and confirmatory statistical analysis methods. Readers interested in learning more about these topics are encouraged to consult the references listed at the end of this document. Additional information on TMDL effectiveness monitoring can also be found in a companion document titled *Technical Guidance for Designing a TMDL Effectiveness Monitoring Plan*.

2. Data Exploration

In any study involving data analysis, the integrity of reported results is directly tied to how thoroughly the basic characteristics of study data are explored and understood. For a TMDL effectiveness monitoring project, exploration of monitoring data that are collected following project implementation is essential. Further, exploration of data used to inform planning decisions (pilot data) can greatly enhance the informative value of the project. Pilot data can include data collected as part of a pilot water quality study, data collected from study sites under other water quality monitoring programs, or existing data collected from comparable sites.

Data exploration answers a number of questions that are relevant to the selection of water quality parameters for TMDL effectiveness monitoring; sampling frequency, timing, and duration; and statistical test selection. These questions can be answered by examining individual data points and by developing a few, readily producible graphs.

<u>Data Distributions & Histograms</u>

Consider any water quality parameter, such as a sample site's total phosphorous concentration. Over time, the concentration of total phosphorous will vary between a minimum and maximum value. Some values will occur regularly while others only in rare instances. If samples were collected continuously at all points in time (i.e., the entire population were known), a complete assessment of the ***probability distribution*** of total phosphorous could be performed. A parameter's probability distribution describes the range of values that can occur and the relative likelihood that they will occur. Since it is not feasible to measure each and every occurrence of total phosphorous, the probability distribution of total phosphorous (or any other water quality parameter) must be inferred from the distribution of a limited number of data points, the ***sampling distribution***. The sampling distribution can be thought of as an estimate of the otherwise-unknown probability distribution.

The sampling distribution of a variable can be displayed graphically using a ***histogram***. A histogram consists of individual data values, or ranges of data values (called bins), on the x-axis. The y-axis contains the number of times each data (or bin) value appears in the dataset (the frequency of the value). Step-by-step instructions for generating a histogram in Microsoft Excel 2007 are provided on page 5.

When viewing a histogram, a principal question to ask is "Do the data display the ***normal distribution***?" The normal distribution is a specific type of probability distribution, with values clustered symmetrically around a common value (the mean) (Figure 1). The normal distribution serves as the basis for ***parametric statistics***. A fundamental assumption of parametric statistical methods is that data are normally distributed, and the accuracy of these methods is greatly diminished if they are applied to data that are not normally distributed. This is a key point for TMDL effectiveness monitoring since a number of parametric statistical tests are available to evaluate monitoring data. These tests frame study questions in the form of a prior assumption, or hypothesis, about the data. As such, they are known as ***hypothesis tests.*** Alternatives to parametric tests exist in the form of ***nonparametric tests***. Nonparametric tests include no assumption of a normal (or any other) distribution (i.e., they are robust). When applied to normally distributed data however, nonparametric tests always have less power than an equivalent parametric test.

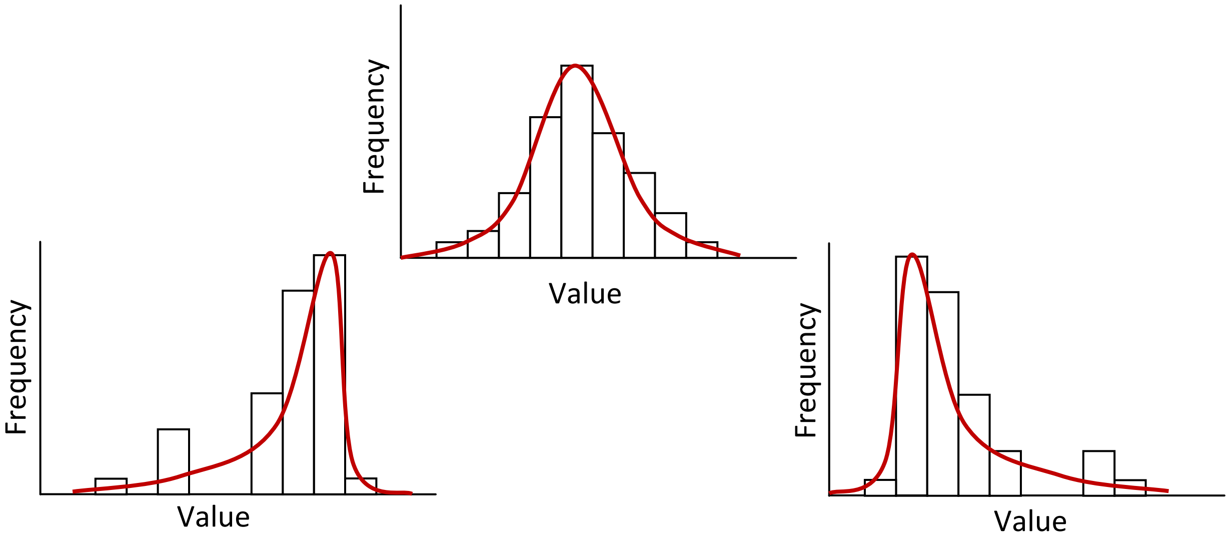
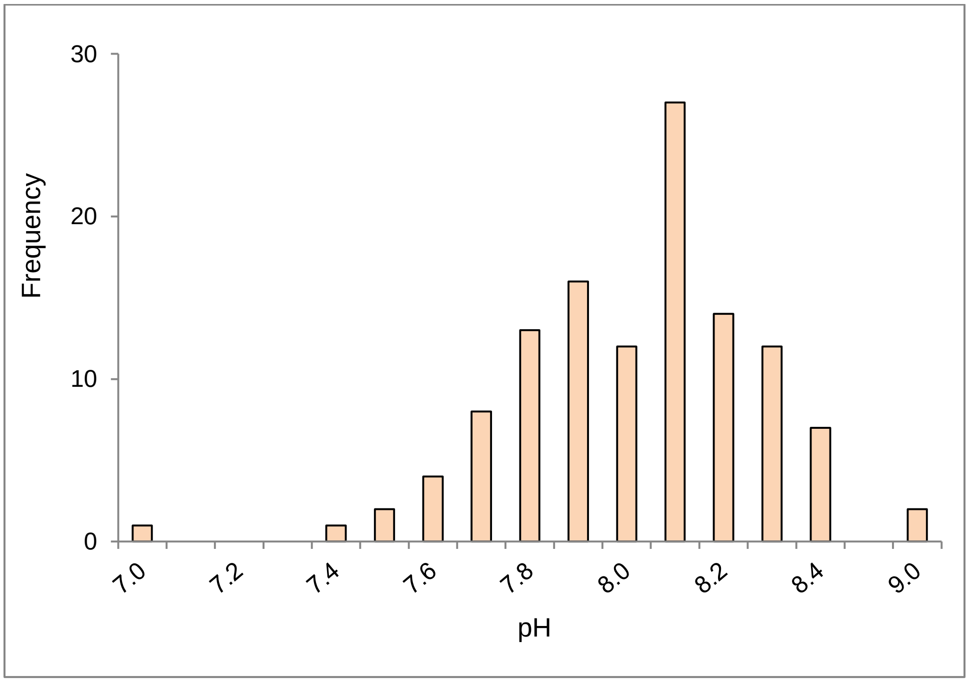


Figure 1. Example histograms demonstrating data with the normal distribution (top), left-skewed distribution (bottom left), and right-skewed distribution (bottom right).

Figure 2 illustrates histograms for two sample datasets from the Spokane River, WA. The pH histogram conveys an approximate normal distribution, while the phosphorous histogram appears right-skewed. Keep in mind that the classic, bell-shape of a histogram of normally-distributed data will rarely be observed. This should not lead one to reject the prospect of a normal distribution. Rather, characteristics such as sample size, the degree of skewness, and the parameter considered should be evaluated.



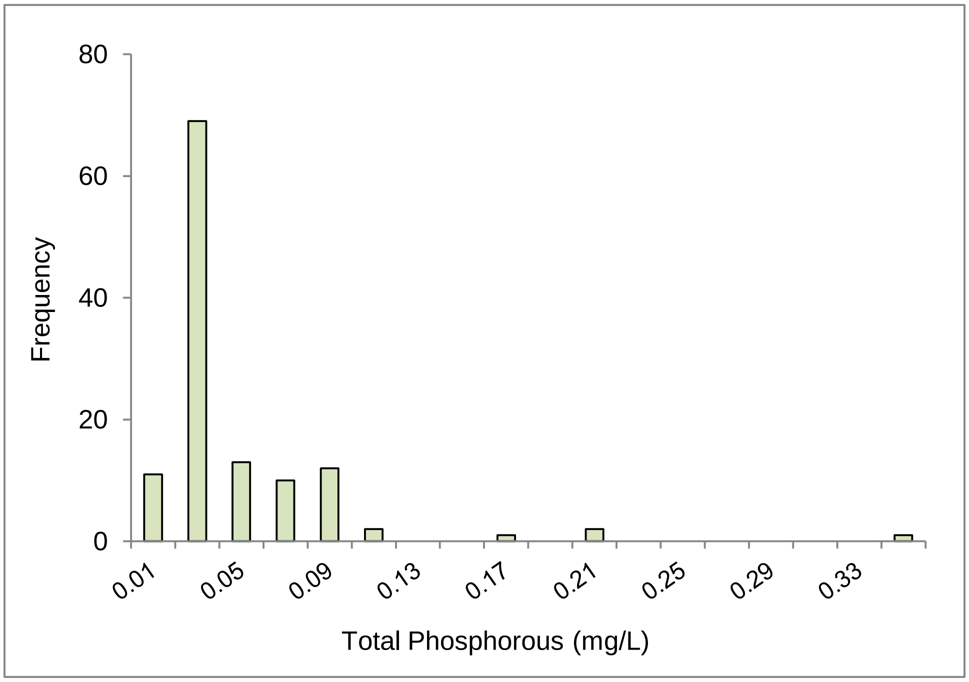
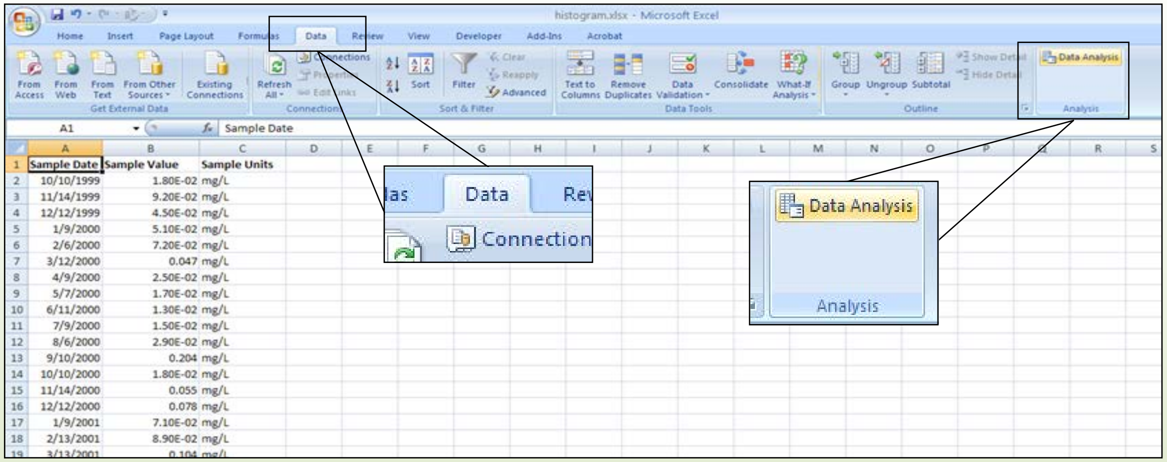


Figure 2. Histograms of Spokane River (at Riverside State Park) pH (top) and total phosphorous concentration (bottom) data collected between 10/1999 and 9/2009. Note differences in the degree of skewness (symmetry) for each parameter.

EDA in Excel. Creating a Histogram

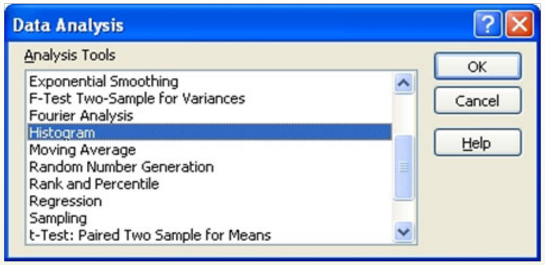
In Excel 2007, use the ***Data Analysis*** tool to create a histogram. In the example below, a histogram is created for Spokane River (at Riverside State Park) total phosphorous concentration using monthly samples collected over the period of October 1999 through September 2009.

Step 1. Arrange the data so that one column contains all measured values (see below example). Click the ***Data*** tab. In the Analysis section, click ***Data Analysis***. If the Data Analysis button is not displayed, see Appendix A for instructions on installing the Analysis Toolpak.



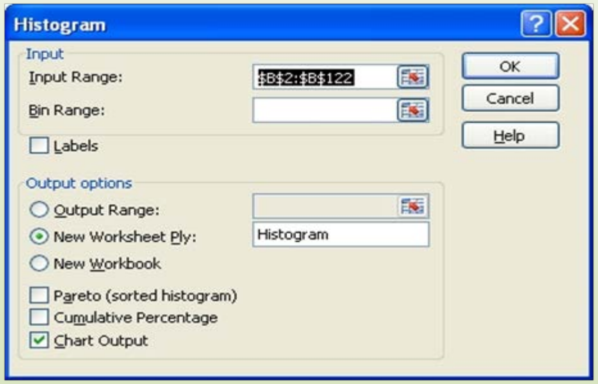
	A	B	C
1	Sample Date	Sample Value	Sample Units
2	10/10/1999	1.80E-02	mg/L
3	11/14/1999	9.20E-02	mg/L
4	12/12/1999	4.50E-02	mg/L
5	1/9/2000	5.10E-02	mg/L
6	2/6/2000	7.20E-02	mg/L
7	3/12/2000	0.047	mg/L
8	4/9/2000	2.50E-02	mg/L
9	5/7/2000	1.70E-02	mg/L
10	6/11/2000	1.30E-02	mg/L
11	7/9/2000	1.50E-02	mg/L
12	8/6/2000	2.90E-02	mg/L
13	9/10/2000	0.204	mg/L
14	10/10/2000	1.80E-02	mg/L
15	11/14/2000	0.055	mg/L
16	12/12/2000	0.078	mg/L
17	1/9/2001	7.10E-02	mg/L
18	2/13/2001	8.90E-02	mg/L
19	3/13/2001	0.104	mg/L

Step 2. Select ***Histogram*** from the Data Analysis dialog box.



Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means

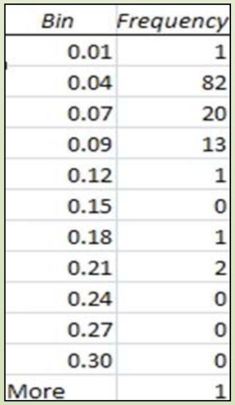
Step 3. In the histogram interface, specify the Input & Output Range. Make sure the ***Chart Output*** option is checked.



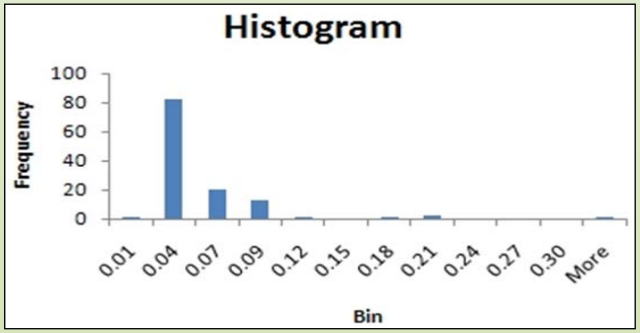
Input
Input Range: $B$2:$B$122
Bin Range:
□ Labels

Output options
○ Output Range:
● New Worksheet Ply: Histogram
○ New Workbook
□ Pareto (sorted histogram)
□ Cumulative Percentage
☑ Chart Output

Histogram output includes bin ranges and frequency data in table and graph form.
Bin ranges can be edited and input to the histogram interface by repeating steps 1 through 3.



Bin	*Frequency*
0.01	1
0.04	82
0.07	20
0.09	13
0.12	1
0.15	0
0.18	1
0.21	2
0.24	0
0.27	0
0.30	0
More	1



Data Transformation

If a sampling distribution is heavily skewed, a data transformation can often be applied to produce symmetrical data that better fit the normality assumption of parametric statistical tests. In theory, any transformation can be used as long as it is applied to all data points and symmetry is achieved. Log or power transformations are the most common transformation types. These transformations involve calculation of the logarithm of each data point or the value of each data point raised to some power (such as the square or square-root of each data point).

Figure 3 contains histograms of untransformed and log-transformed total phosphorous data collected from the Spokane River (at Riverside State Park). The untransformed data are right-skewed. Transformation using the logarithm provides a dataset that is approximately normally distributed. These data can subsequently be used for parametric statistical testing.

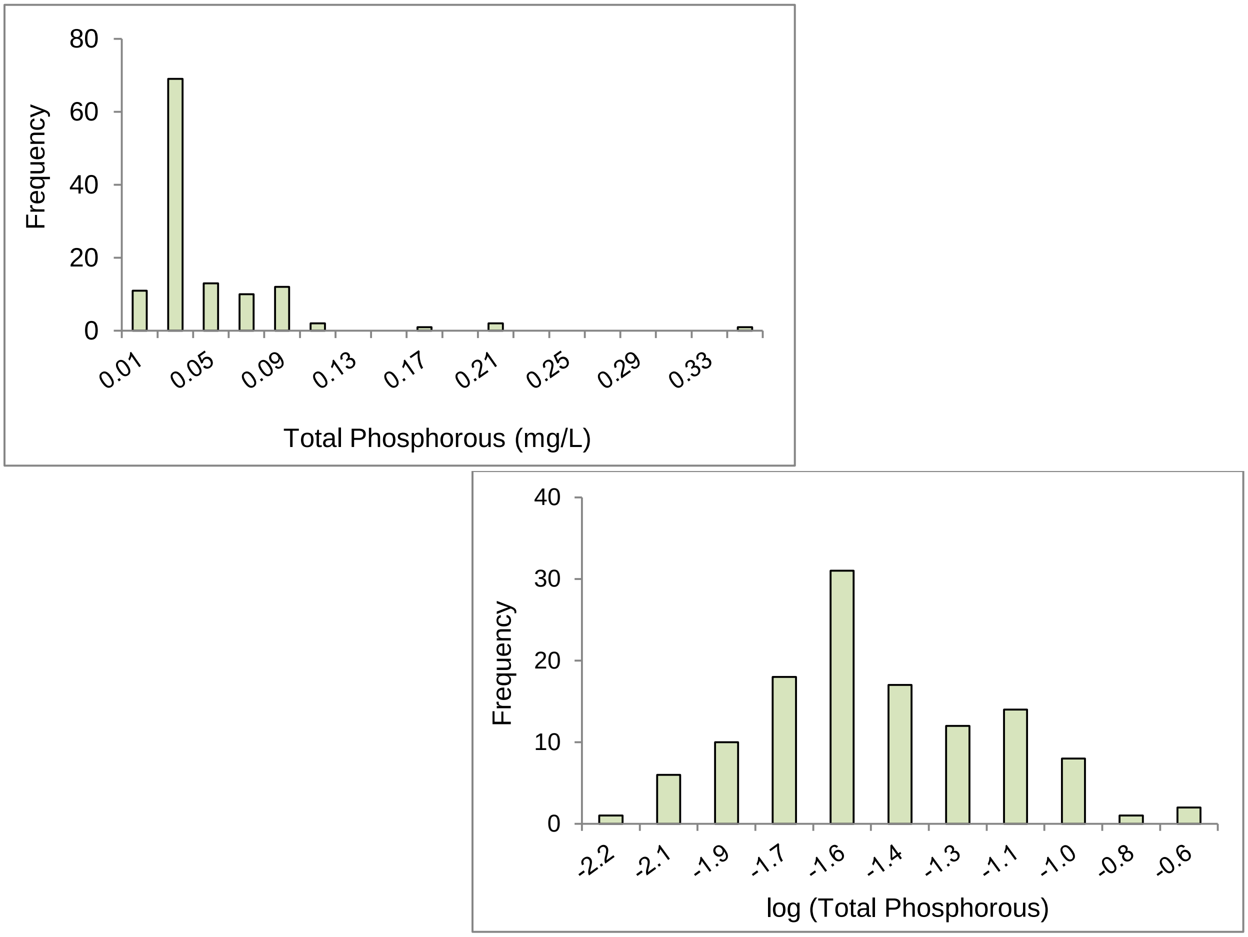


Figure 3. Histograms of total phosphorous concentration (top) and log-transformed total phosphorous concentration (bottom) data for samples collected from the Spokane River (at Riverside State Park) between 10/1999 and 9/2009. Note that log-transformed data appear normally distributed.

When dealing with skewed data, one should not become preoccupied with finding the "right transformation." Start with the common transformations (log, square, square-root, etc.) and compare histograms. If one of these does not produce symmetrical data, consult sources specific to the water quality parameter being analyzed. Keep in mind that transformations can provide values of zero and negative numbers and that some transformations cannot be applied if a dataset contains one or more zero values. Remember that if two groups of data are compared (e.g., before and after TMDL implementation), the same transformation must be applied to each group. A comparison of pre-TMDL log-transformed total phosphorous and post-TMDL untransformed total phosphorous, for example, would not be appropriate. Also, results generated using transformed data are often transformed back to original units for reporting (back-transformation). Readers are advised to use caution when back-transforming results to prevent calculation and reporting errors.

A final note to consider is that even though a transformation may provide data that meet the normality assumption of parametric statistical tests, other assumptions specific to the test(s) applied should also be explored. Refer to Helsel & Hirsch (2006) and U.S. Environmental Protection Agency (2006) for a detailed discussion of the assumptions of common statistical tests.

Outliers

It is common for a dataset to contain one or more values that are far removed from the remaining observations. These values, called outliers, can arise from a variety of sources. An outlier may simply reflect the occurrence of a rare event or be the product of a measurement or data entry error. Histograms are useful for identifying outliers (e.g., see the top histogram in Figure 3, which shows an outlier of 0.35 mg/L).

Outliers should not be removed or deleted from a dataset (unless they are known to be measurement/recording errors). Instead, the presence of outliers should be noted, as these values can influence the selection of a parametric vs. nonparametric statistical test. This is due to the use of the mean value of a dataset for parametric tests and bias in the mean that is caused by outliers. If one or more outliers are identified, data transformation options should be explored to determine if an outlier-free dataset can be produced. If transformation does not produce symmetry in the data (or if transformation is not desirable), nonparametric methods should be used for subsequent analysis. Should a parametric test be performed on a dataset that includes outliers, practitioners can evaluate the influence of the outliers by performing the test twice, once using the full dataset (including the outliers) and again on the reduced dataset (excluding the outliers). If the results are different from one another, a nonparametric test should be used.

Time Series & Missing Data Points

Water quality monitoring often involves the collection of samples at regular intervals over time (hourly, weekly, monthly, etc.). Plotting these data in sequence (in the order they were collected) can display important information, including shifts in the central tendency (mean or median) or variability over time, and the existence of large gaps in coverage over time. These plots are known as *time series* plots. Time series plots contain the sample date on the x-axis and sample value on the y-axis. Instructions for producing time series plots in Microsoft Excel 2007 are provided on page 9.

A time series plot of fecal coliform concentration in Burnt Bridge Creek, WA over a six-year period is provided in Figure 4. Readers may notice one of several data characteristics conveyed in this plot. These include the presence of outliers, seasonality of measured values (lower values during the winter months), and the large gap in data points from October 2005 to October 2007.

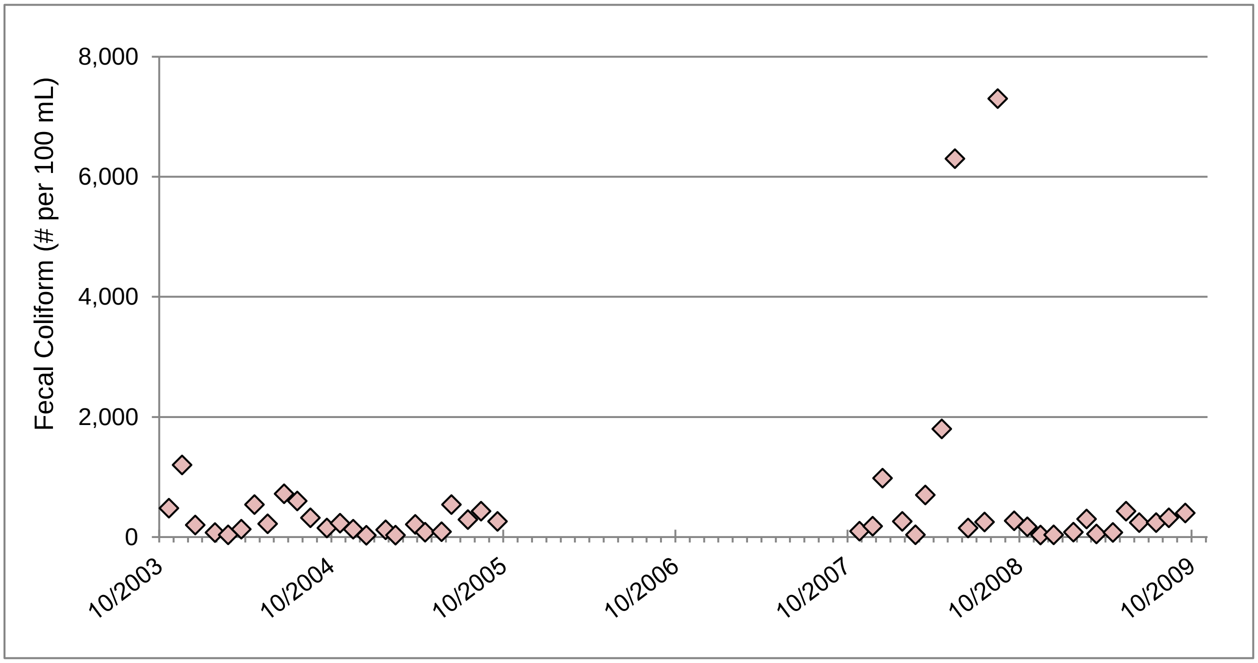


Figure 4. Time series plot for monthly fecal coliform concentration in Burnt Bridge Creek, WA. The plot displays several important pieces of information, including the data gap in 2005-06, the presence of 2 outliers in May and August of 2008, and seasonality in measured fecal coliform concentrations (lower values during the winter months).

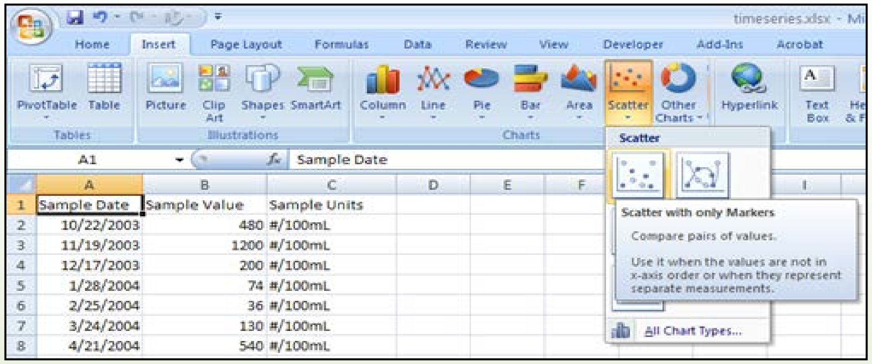
The large data gap apparent in Figure 4 can be problematic if these data are included in a trend monitoring study. Trend monitoring data are used to assess changes in a parameter over time (are the values increasing or decreasing with time?). A complete data record is ideal for such analysis. However, data gaps often occur, and researchers must decide whether the data gap will affect trend analysis results. An objective method for evaluating the significance of data gaps, proposed by Helsel & Hirsch (2002), includes 3 steps:

1. Divide the entire study period into three separate periods of equal length.
2. Calculate the percent coverage in each period (the ratio of actual observations to potential observations, as a percentage).
3. Discard the data if percent coverage is less than 20% in any of the three periods.

EDA in Excel. Creating Time Series Plots

In Excel 2007, use the ***Scatter Chart*** tool to create a time series plot. In the example below, a time series plot is created for Burnt Bridge Creek fecal coliform concentration using monthly sample data collected over the period October 2003 through September 2009.

Step 1. Arrange the data so that one column contains all measured values and one column contains ordered sample dates (see below example). Click the ***Insert*** tab. In the ***Charts*** section, click ***Scatter***. Select ***Scatter with only Markers***.



A1 · *fx* Sample Date

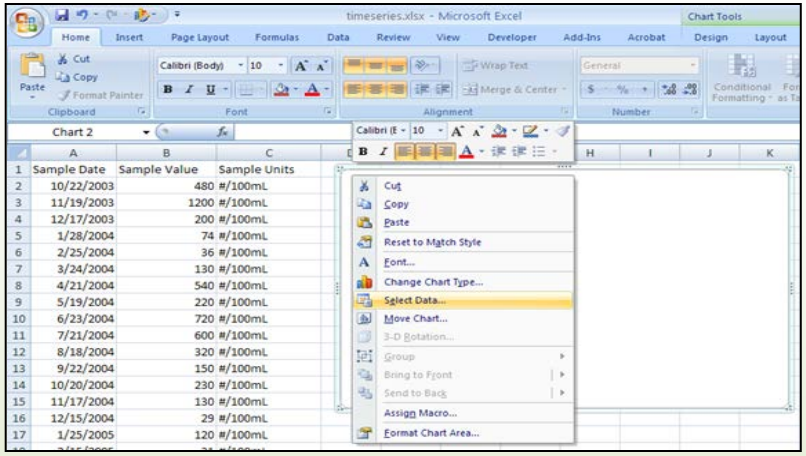
	A	B	C
1	Sample Date	Sample Value	Sample Units
2	10/22/2003	480	#/100mL
3	11/19/2003	1200	#/100mL
4	12/17/2003	200	#/100mL
5	1/28/2004	74	#/100mL
6	2/25/2004	36	#/100mL
7	3/24/2004	130	#/100mL
8	4/21/2004	540	#/100mL

Scatter with only Markers

Compare pairs of values.

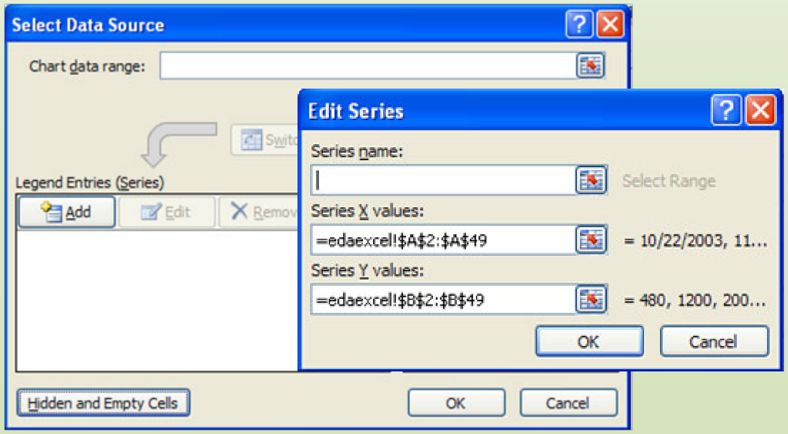
Use it when the values are not in x-axis order or when they represent separate measurements.

Step 2. A blank chart is created. Right-click on the chart area and choose ***Select Data***.



	A	B	C
1	Sample Date	Sample Value	Sample Units
2	10/22/2003	480	#/100mL
3	11/19/2003	1200	#/100mL
4	12/17/2003	200	#/100mL
5	1/28/2004	74	#/100mL
6	2/25/2004	36	#/100mL
7	3/24/2004	130	#/100mL
8	4/21/2004	540	#/100mL
9	5/19/2004	220	#/100mL
10	6/23/2004	720	#/100mL
11	7/21/2004	600	#/100mL
12	8/18/2004	320	#/100mL
13	9/22/2004	150	#/100mL
14	10/20/2004	230	#/100mL
15	11/17/2004	130	#/100mL
16	12/15/2004	29	#/100mL
17	1/25/2005	120	#/100mL

Step 3. In the ***Select Data Source*** dialog box, click ***Add***. In the Edit Series dialog box, enter the range of cells containing sample dates below ***Series X values***. Enter the range of cells containing sample values below ***Series Y values***. Click *OK*.

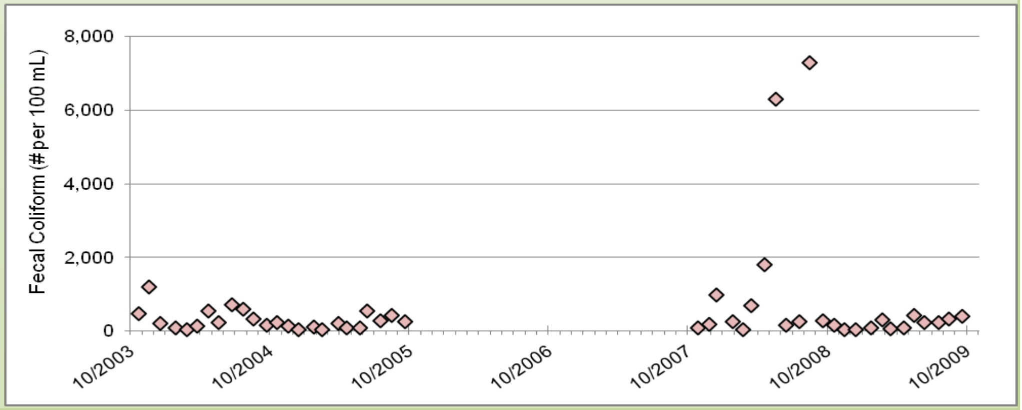


Series name:

Series X values: =edaexcel!$A$2:$A$49 = 10/22/2003, 11...

Series Y values: =edaexcel!$B$2:$B$49 = 480, 1200, 200...

Step 4. Edit axis limits, titles, and data markers as needed.



Daily/Seasonal Fluctuation and Box Plots

Just as weather data (air temperature, rainfall, etc.) regularly display distinct daily and seasonal patterns, water quality parameters often fluctuate over daily and/or seasonal cycles. Daily, or ***diel***, fluctuation refers to regular, cyclic variation in the distribution of a parameter over a 24-hour period. ***Seasonality*** refers to the phenomenon of regular, cyclic variation over a one-year period. Diel and seasonal fluctuations can be detected from time series plots and are important for analysis of TMDL effectiveness monitoring data. For example, a comparison of pre- and post-TMDL data for a parameter showing strong seasonality is of little use if pre-TMDL data consist of observations from the month of December and post-TMDL data consist of July observations only. Similarly, changes in a parameter with large diel fluctuation over time can be clouded if continuous data are not available and individual grab samples are not collected at the same time of day.

In addition to time series plots, cyclic fluctuation can be explored using ***box plots***. A box plot conveys five key pieces of information on the distribution of a parameter: the minimum, the maximum, the median (50th percentile), the 25th percentile, and the 75th percentile. A box plot contains summary statistic values on the y-axis and parameter groups on the x-axis. For each parameter group, the five summary statistics are displayed using the following format:

- A box with lower and upper limits at the 25th and 75th percentile values, respectively.
- A horizontal line or point within the box representing the median value.
- Vertical lines (often called whiskers) extending outside of the box to the minimum and maximum observed values.

A daily or seasonal boxplot can be constructed to compare summary statistics between seasons or periods of the day (day vs. night). An example is shown in Figure 5 and instructions for preparing boxplots in Microsoft Excel 2007 are provided on page 12.

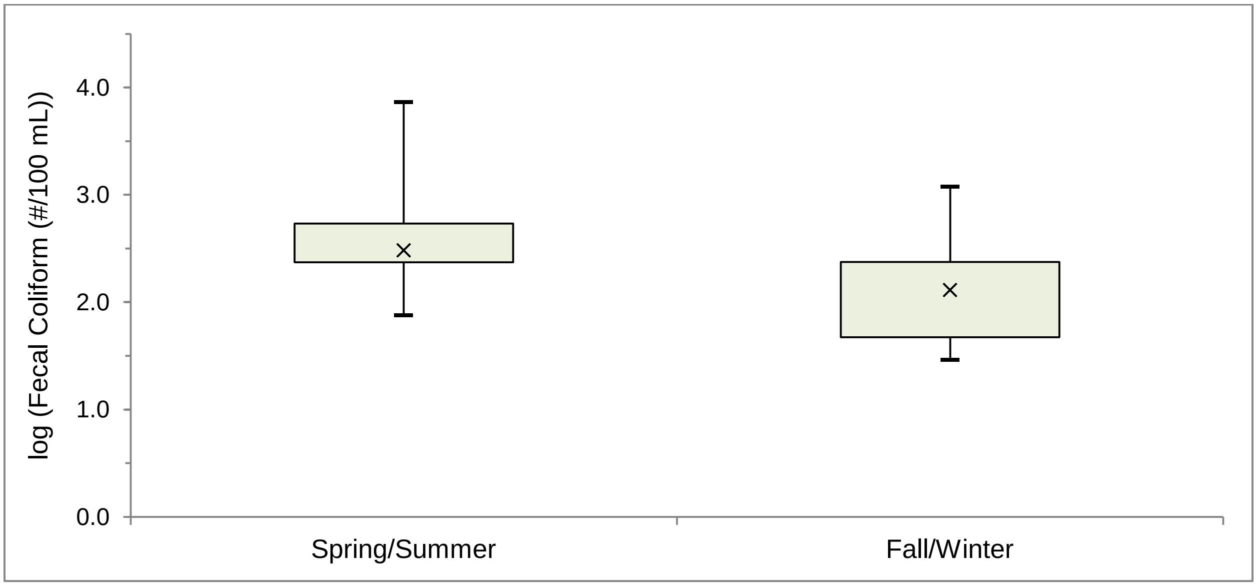


Figure 5. Box plot comparing spring/summer and fall/winter fecal coliform concentration in Burnt Bridge Creek, WA over the period 10/2003 through 9/2009. The box plot provides a graphical display of five key summary statistics: the median (the X symbol); 25th and 75th percentile values (the lower and upper sides of the box); and minimum and maximum observed values (the lines outside the box). Here, the box plot points to seasonal differences in the distribution of fecal coliform concentrations.

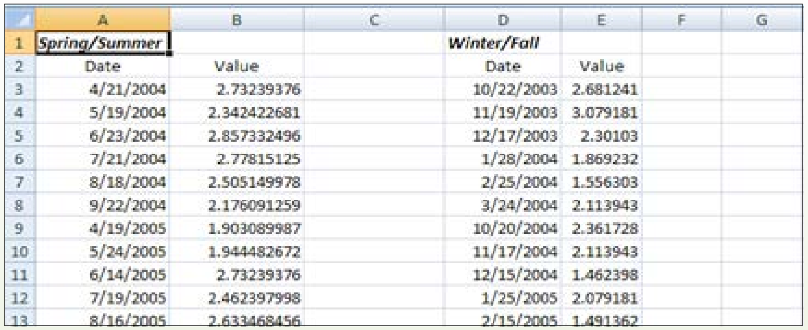
Knowledge of diel and seasonal fluctuation in parameters included in TMDL effectiveness monitoring should be integrated into project planning and analysis decisions. For example, it may be desirable to sample a parameter with large seasonal fluctuation only during the portion of the year when it is known to be at risk for exceeding water quality standards (e.g., summer bacteria counts). Alternatively, several statistical methods that take seasonal variability into account are available for analysis of TMDL effectiveness monitoring data. For example, the Seasonal Kendall test, a modified version of the nonparametric Kendall test for trend, is a common method for including the effect of seasonal variability in trend analysis. Detailed information on the Seasonal Kendall test, and other statistical methods that account for seasonality effects, can be found in Helsel & Hirsch (2002) and Helsel et al. (2006).

Parameters with a strong diel pattern may require continuous water quality monitoring, in which measurements are collected and recorded several times throughout the day. These data can be used to generate daily statistics (mean, maximum, minimum, etc.) for use in further statistical analysis. If continuous monitoring is not a viable option, parameters with large diel fluctuation should be sampled at roughly the same time of day. Examples of parameters with potential diel cycles include dissolved oxygen, pH, water temperature, and stream flow. In some water bodies, the concentration of nutrients can also vary dramatically throughout the day.

EDA in Excel. Creating a Boxplot

Excel 2007 does not include a built-in box plot tool. Box plots instead must be created using the ***Line Chart Tool***. In the example below, a seasonal box plot is created for Burnt Bridge Creek fecal coliform concentration using monthly sample data collected over the period October 2003 through September 2009.

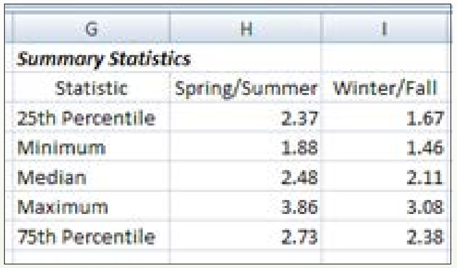
Step 1. Arrange the data so that one column contains all measured values and one column contains the sampling date for each seasonal group.



	A	B	C	D	E	F	G
1	*Spring/Summer*			*Winter/Fall*			
2	Date	Value		Date	Value		
3	4/21/2004	2.73239376		10/22/2003	2.681241		
4	5/19/2004	2.342422681		11/19/2003	3.079181		
5	6/23/2004	2.857332496		12/17/2003	2.30103		
6	7/21/2004	2.77815125		1/28/2004	1.869232		
7	8/18/2004	2.505149978		2/25/2004	1.556303		
8	9/22/2004	2.176091259		3/24/2004	2.113943		
9	4/19/2005	1.903089987		10/20/2004	2.361728		
10	5/24/2005	1.944482672		11/17/2004	2.113943		
11	6/14/2005	2.73239376		12/15/2004	1.462398		
12	7/19/2005	2.462397998		1/25/2005	2.079181		
13	8/16/2005	2.633468456		2/15/2005	1.491362		

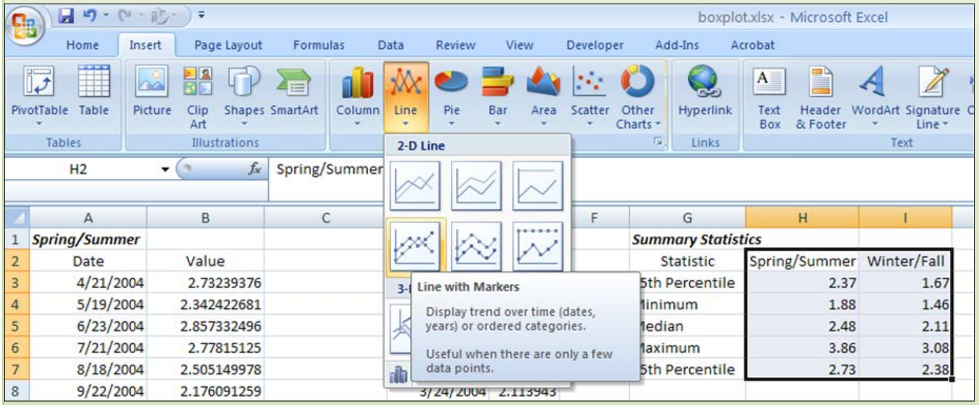
Step 2. Create a table of the five summary statistics for each seasonal group in the following order: 25th percentile, Minimum, Median, Maximum, 75th percentile. Formulas for calculating summary statistics are provided below:

- 25th percentile: =PERCENTILE(Data Range, 0.25)
- Minimum: =MINIMUM(Data Range)
- Median: =MEDIAN(Data Range)
- Maximum: =MAXIMUM(Data Range)
- 75th percentile: =PERCENTILE(Data Range,0.75)



G	H	I
Summary Statistics		
Statistic	Spring/Summer	Winter/Fall
25th Percentile	2.37	1.67
Minimum	1.88	1.46
Median	2.48	2.11
Maximum	3.86	3.08
75th Percentile	2.73	2.38

Step 3. Highlight summary data and click the ***Insert*** tab. In the ***Charts*** section, click ***Line*** and select ***Line with Markers***.



H2 Spring/Summer

Line with Markers

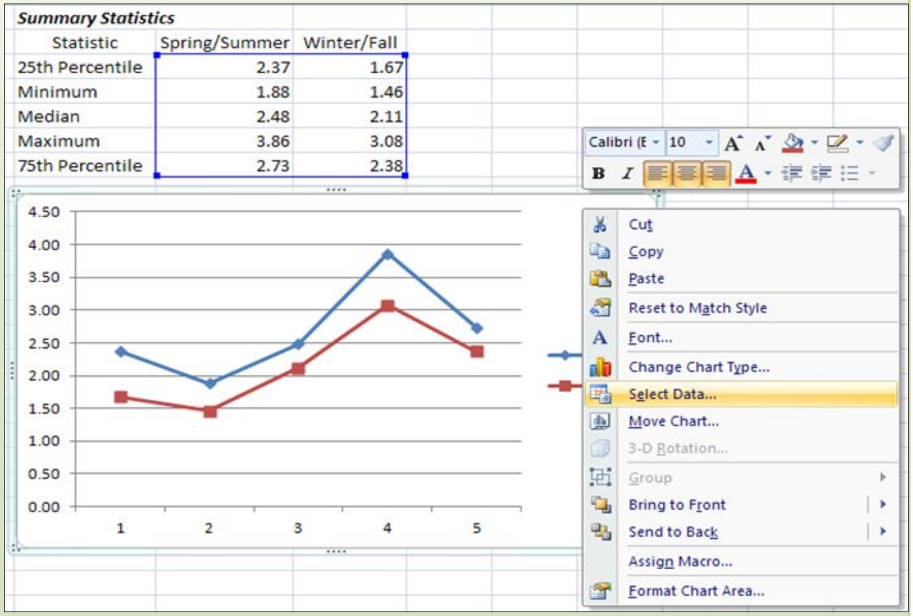
Display trend over time (dates, years) or ordered categories.

Useful when there are only a few data points.

	A	B	C	F	G	H	I
1	*Spring/Summer*				*Summary Statistics*		
2	Date	Value			Statistic	Spring/Summer	Winter/Fall
3	4/21/2004	2.73239376			5th Percentile	2.37	1.67
4	5/19/2004	2.342422681			inimum	1.88	1.46
5	6/23/2004	2.857332496			edian	2.48	2.11
6	7/21/2004	2.77815125			aximum	3.86	3.08
7	8/18/2004	2.505149978			5th Percentile	2.73	2.38
8	9/22/2004	2.176091259					

Continued on next page…

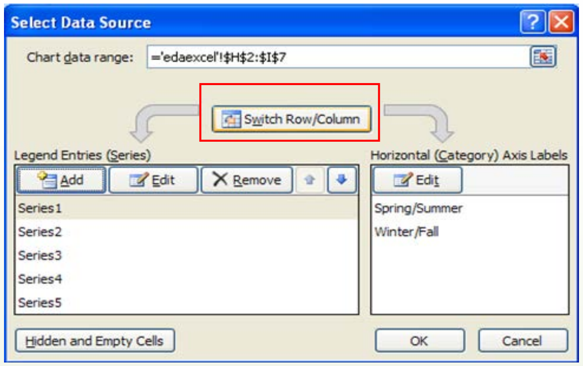
Step 4. Right-click on the new chart and go to ***Select Data…***



Summary Statistics		
Statistic	Spring/Summer	Winter/Fall
25th Percentile	2.37	1.67
Minimum	1.88	1.46
Median	2.48	2.11
Maximum	3.86	3.08
75th Percentile	2.73	2.38

EDA in Excel. Creating a Boxplot (cont'd.)

Step 5. In the Select Data Source box, click ***Switch Row/Column***. Click ***OK***.



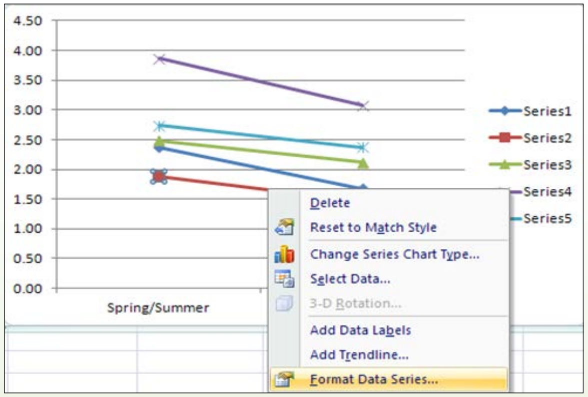
Select Data Source

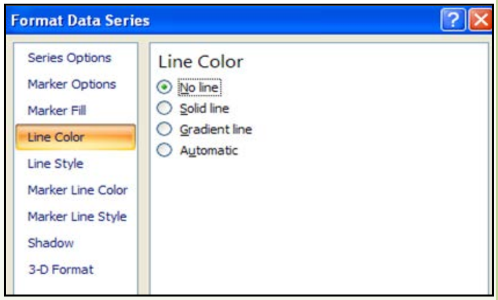
Chart data range: ='edaexcel'!$H$2:$I$7

Legend Entries (Series): Series1, Series2, Series3, Series4, Series5

Horizontal (Category) Axis Labels: Spring/Summer, Winter/Fall

Step 6. Right-click on a data series and select ***Format Data Series…***. In the Format Data Series box, select the ***Line Color*** menu. Select ***No Line*** and click ***Close***. Repeat for all 5 data series.

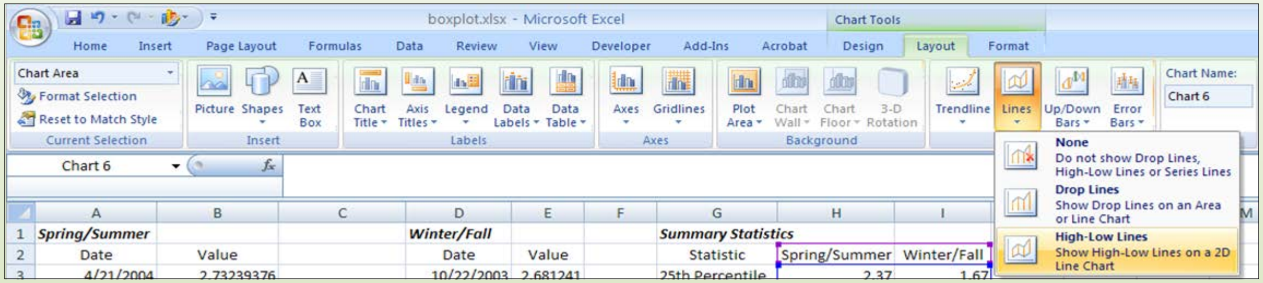




Format Data Series

Line Color: ● No line ○ Solid line ○ Gradient line ○ Automatic

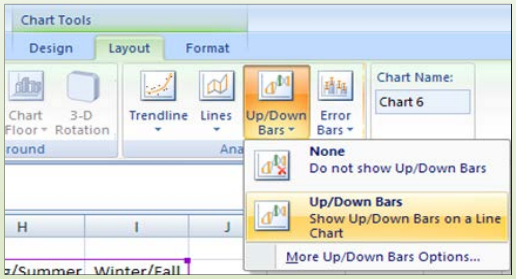
Step 7. Under ***Chart Tools***, select the ***Layout*** tab. In the ***Analysis*** section, select ***Lines***. Click ***High-Low Lines***. Repeat for ***Up/Down Bars***.



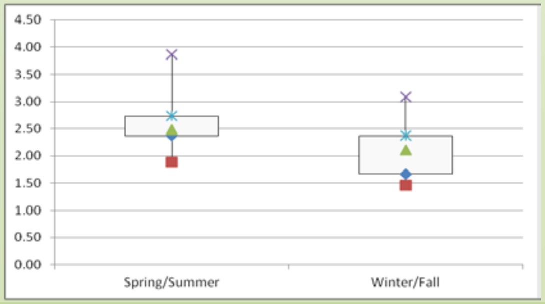
boxplot.xlsx - Microsoft Excel

Chart Name: Chart 6

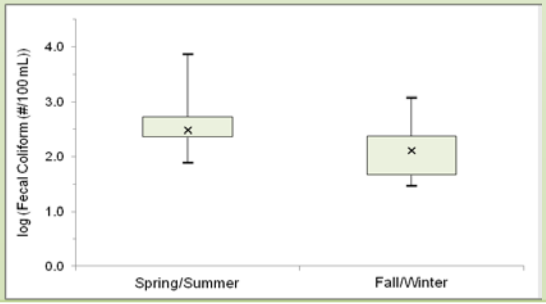
	A	B	C	D	E	F	G	H	I
1	*Spring/Summer*			*Winter/Fall*			*Summary Statistics*		
2	Date	Value		Date	Value		Statistic	Spring/Summer	Winter/Fall
3	4/21/2004	2.73239376		10/22/2003	2.681241		25th Percentile	2.37	1.67



Your chart will now contain markers for each summary statistic, a vertical line connecting the minimum and maximum value, and a box connecting the 25th and 75th percentile values. Edit chart formatting as desired.



→



Autocorrelation and Lag Plots

Seasonal and daily cycling in time series data are two specific examples that reflect the statistical phenomenon known as ***autocorrelation***. Autocorrelation, or serial correlation, is defined as correlation between the elements of a series with other elements of the same series separated by some time interval. In other words, autocorrelation describes how similar (positive autocorrelation) or different (negative autocorrelation) observed values are with past or future values. For a parameter with strong seasonal fluctuation, observations separated by one year will exhibit positive autocorrelation, and observations separated by six months may demonstrate negative autocorrelation. Similarly, a parameter with large diel fluctuation will exhibit positive autocorrelation among observations separated by 24 hours and negative autocorrelation among observations separated by 12 hours.

When exploring autocorrelation, the time interval separating correlated observations is referred to as the time lag. A common tool for evaluating autocorrelation at a particular time lag is the ***lag plot***. The lag plot is a collection of points, each representing a matched pair between the observed value at time, t, and the observed value at time, t minus lag. Random scatter in the lag plot indicates minimal autocorrelation at the selected lag time, while data with strong autocorrelation will produce a more structured lag plot. Example lag plots are shown in Figure 6 and instructions for creating a lag plot in Microsoft Excel 2007 are provided on page 16.

The importance of autocorrelation that reflects regular, periodic cycling of water quality conditions (e.g., diel/seasonal cycling) to TMDL effectiveness monitoring is discussed in the preceding section. Autocorrelation that indicates short-term persistence in water quality conditions from one observation to the next must also be reviewed. Persistence refers to the tendency of similar values to follow one another (large values follow large values and small values follow small values). Short-term persistence can be evaluated by exploring autocorrelation at small time lags, generally at the sampling frequency (i.e., time lag equal to 1 sampling interval, or lag-1 autocorrelation).

Lag-1 autocorrelation is noteworthy for analysis of TMDL effectiveness monitoring data since a basic assumption of many standard statistical tests is that the dataset is comprised of independent or random observations. The presence of strong lag-1 autocorrelation in a dataset implies that observed values are not random. Ultimately, this reduces the amount of information contained within the data, as a given sample provides minimal "new" information beyond what was already known from the previous observation. While lag-1 autocorrelation is generally strongest for samples collected at a high frequency, it can also be present in data collected at monthly or even annual intervals.

Strong lag-1 autocorrelation can affect the accuracy of long-term trend detection and increase the frequency of "false positives" reported by statistical tests (i.e., the actual confidence level is less than that specified under a classical statistical test). Methods to "correct" for autocorrelation generally involve modifying the dataset to remove the autocorrelation effect or the use of statistical tests that have been specifically developed for autocorrelated data. Simple options include performing a classical statistical test using a subset of the original dataset or grouped values (such as monthly or quarterly means). If lag plots point to the presence of strong short-term persistence, readers are encouraged to consult advanced statistical resources (such as Helsel & Hirsch (2002) or Kirchner (2001)) or a statistician to gain insight into methods specific to their analysis objectives.

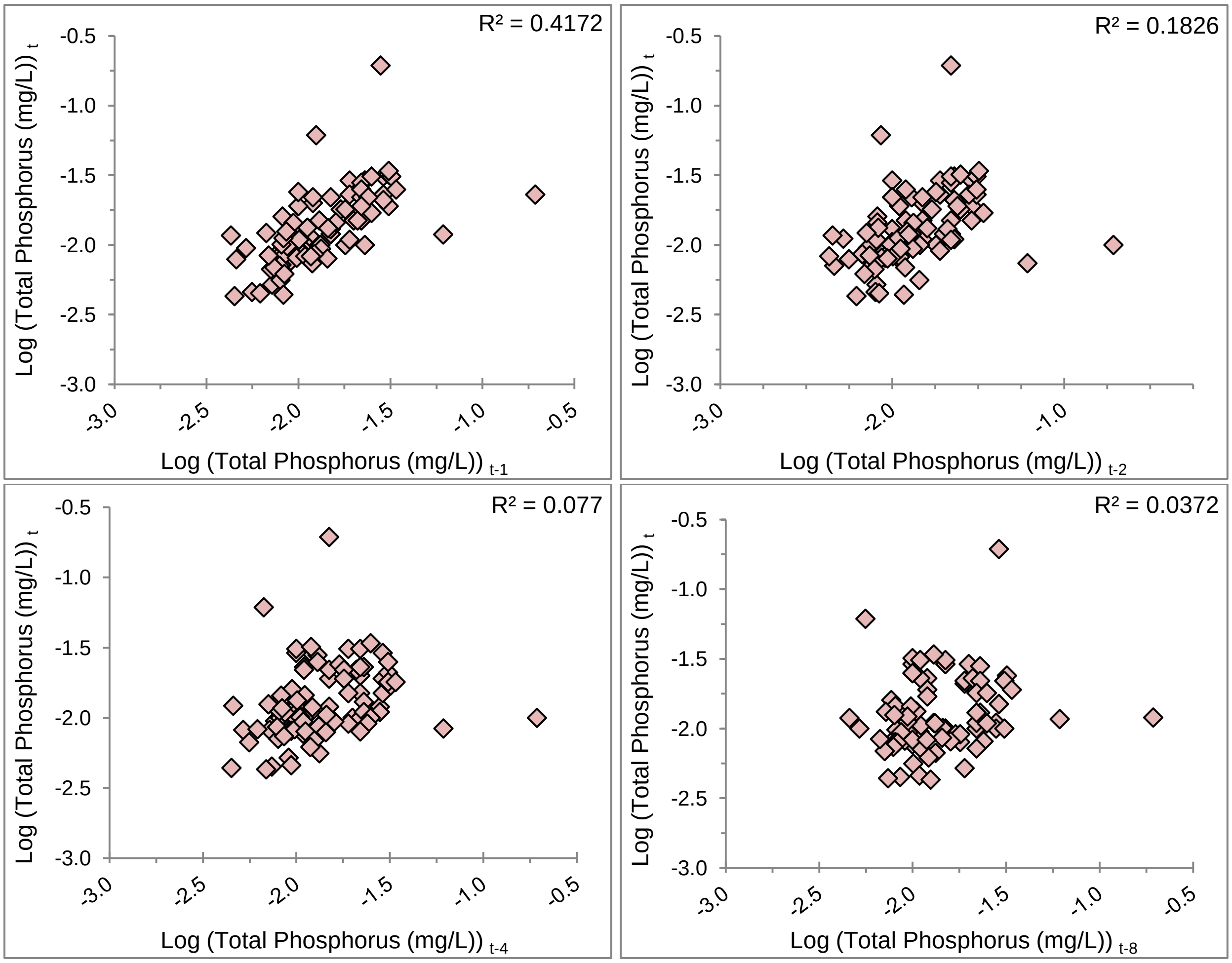
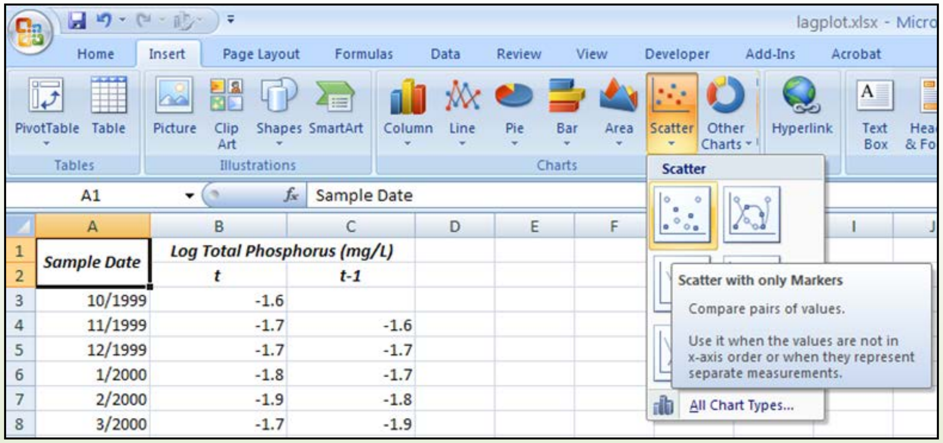


Figure 6. Lag plots of Cedar River, WA total phosphorus concentration at lag times of one month (top left), two months (top right), four months (bottom left), and eight months (bottom right). Note that the strong correlation evident at the one month time lag fades at higher lag times. This pattern is typical of data with short-term persistence.

EDA in Excel. Creating a Lag Plot

In Excel 2007, use the ***Scatter Chart*** tool to create a lag plot. In the example below, a 1-month lag plot is created for Burnt Bridge Creek fecal coliform concentration using monthly sample data collected over the period from October 1999 through September 2007.

Step 1. Arrange the data so that one column contains observed values and one column contains values offset by the selected time lag. Click the ***Insert*** tab. In the ***Charts*** section, click ***Scatter***. Select ***Scatter with only Markers***.



A1 Sample Date

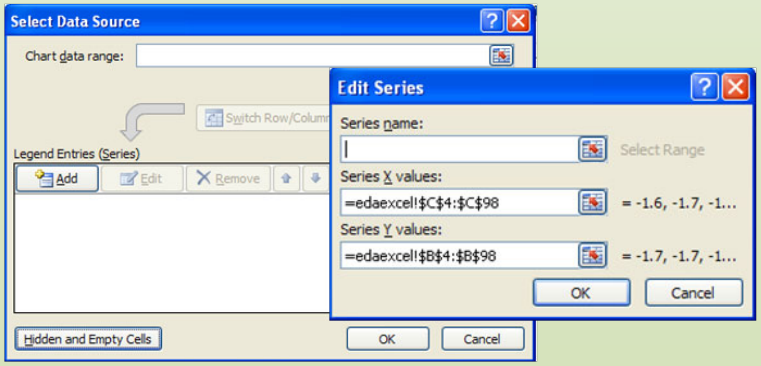
	A	B	C
1	Sample Date	Log Total Phosphorus (mg/L)	
2		t	t-1
3	10/1999	-1.6	
4	11/1999	-1.7	-1.6
5	12/1999	-1.7	-1.7
6	1/2000	-1.8	-1.7
7	2/2000	-1.9	-1.8
8	3/2000	-1.7	-1.9

Scatter with only Markers

Compare pairs of values.

Use it when the values are not in x-axis order or when they represent separate measurements.

Step 3. In the ***Select Data Source*** dialog box, click ***Add.*** In the Edit Series dialog box, enter the range of cells containing lagged values under ***Series X values***. Enter the range of cells containing unlagged values under ***Series Y values***. Click *OK.*



Series name:

Series X values: =edaexcell!$C$4:$C$98 = -1.6, -1.7, -1...

Series Y values: =edaexcell!$B$4:$B$98 = -1.7, -1.7, -1...

Step 2. A blank chart is created. Right-click on the chart area and choose ***Select Data***.

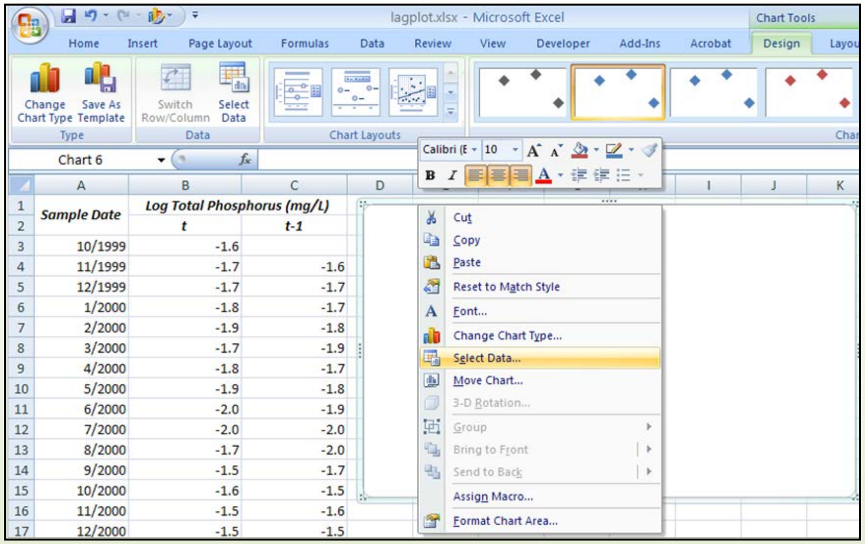
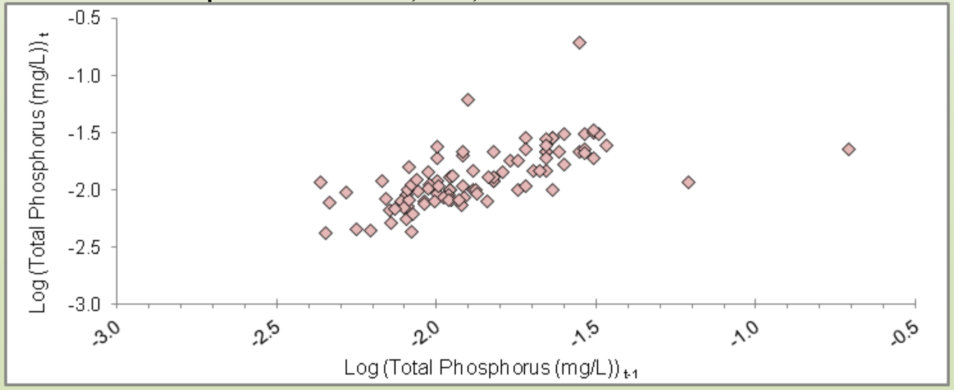


Chart 6

	A	B	C
1	Sample Date	Log Total Phosphorus (mg/L)	
2		t	t-1
3	10/1999	-1.6	
4	11/1999	-1.7	-1.6
5	12/1999	-1.7	-1.7
6	1/2000	-1.8	-1.7
7	2/2000	-1.9	-1.8
8	3/2000	-1.7	-1.9
9	4/2000	-1.8	-1.7
10	5/2000	-1.9	-1.8
11	6/2000	-2.0	-1.9
12	7/2000	-2.0	-2.0
13	8/2000	-1.7	-2.0
14	9/2000	-1.5	-1.7
15	10/2000	-1.6	-1.5
16	11/2000	-1.5	-1.6
17	12/2000	-1.5	-1.5

Step 4. Edit axis limits, titles, and data markers as needed.

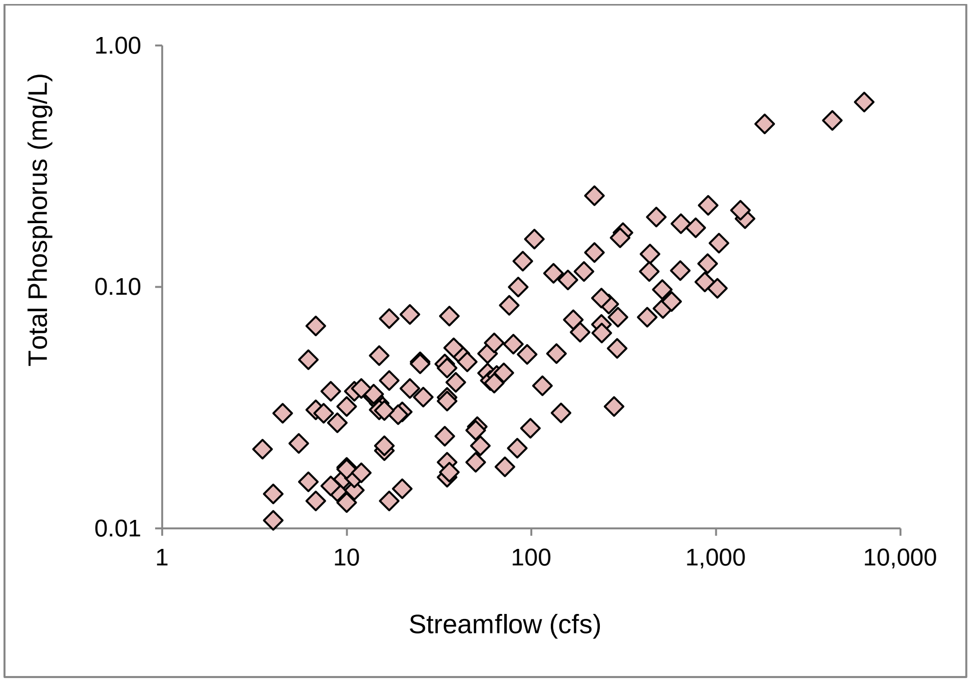


Covariates and Scatterplots

Under a TMDL effectiveness monitoring program, average water quality conditions, and variability in these conditions, are quantified at each monitoring site for the purpose of comparing sites or evaluating change over time. Oftentimes, variability in the parameter(s) of interest is tied to variability in other measurable parameters that directly or indirectly reflect the ultimate drivers of water quality change. Observations of these ***covariates*** increase/decrease with those of the parameter of interest (the data vary together). Covariate data can be used to “sort out” natural variability versus variability attributed to TMDL implementation; therefore, the inclusion of covariates in TMDL effectiveness monitoring can greatly improve the power of the statistical methods applied to detect differences between sites or change over time. For the case of common parameters addressed in TMDLs (e.g., nutrient concentrations, dissolved oxygen concentration, water temperature) common covariates include precipitation, air temperature, and stream flow.

Covariates can be identified by constructing a ***scatterplot***. Scatterplots contain the independent variable on the x-axis and the dependent variable on the y-axis. Note that time series and lag plots are specific types of scatterplots. A scatterplot can be created in Microsoft Excel 2007 by following the instructions provided on page 9 for creating a time series plot (with covariate data replacing sampling date).

A scatterplot allows for a visual review of the correlation between two parameters (do the two parameters vary together?). Figure 7 contains scatterplots of stream flow vs. total phosphorous concentration and stream flow vs. fecal coliform concentration for Hangman Creek, WA. A strong relationship between flow and total phosphorous is apparent, with higher flows corresponding to higher concentrations. Conversely, no relationship is evident between flow magnitude and fecal coliform concentration. If a scatterplot reveals correlation between two parameters, parametric (multiple linear regression) and nonparametric (Kendall trend test with LOWESS) trend analysis techniques are available to account for covariates to better identify trends. A common method for including covariates when comparing two groups (e.g., before and after TMDL implementation) is Analysis of Covariance (ANCOVA). Refer to Helsel & Hirsch (2002) for a detailed discussion of such methods.



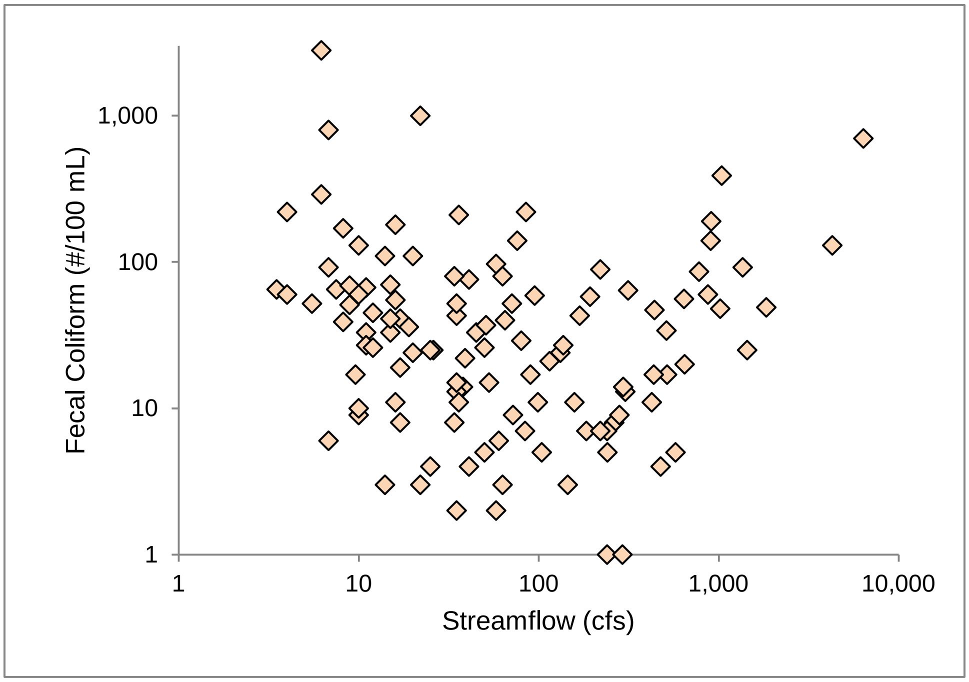


Figure 7. Scatterplots of total phosphorous (top) and fecal coliform (bottom) concentration vs. stream flow for Hangman Creek, WA from 10/1999 through 9/2009. Note the difference in the degree of correlation between stream flow and each water quality parameter.

Censored Data

Laboratory analysis of water quality samples often returns values that are less than the lower detection limit (non-detects) or greater than the upper detection limit of the analytical method applied. In such cases, the actual sample value is only partially known. Data with measured values that are only partially known are referred to as ***censored data***. A researcher may be tempted to discard censored samples to ease analysis. Though these data contain information that is limited relative to other data points, they contain information nonetheless, and therefore should not be removed for subsequent analysis. Doing so will bias measures of the central tendency (e.g., mean) of the dataset and its variability, and can lead to inaccurate interpretation of the data.

Several options are available to address censored data. A basic method is to replace the censored value with a real number value to conform to other measured data. For example, values reported as less than the detection limit can be estimated as zero, as the detection limit, or as a percentage of the detection limit. Replacement values are then used when calculating summary statistics or performing a statistical test. This approach can bias the sample mean and standard deviation and should only be used if a small number of data points are censored. More complex methods of estimating summary statistics should be applied when several data points are censored (see Helsel & Hirsh (2002); U.S. Environmental Protection Agency (2006)). A general rule of thumb is to apply simple substitution only if censored data make up less than 15% of the dataset (U.S. Environmental Protection Agency, 2006).

For hypothesis testing, substitution of censored data is not required if a nonparametric test is performed. For example, substitution is not needed for trend analysis if the nonparametric Mann-Kendall test is applied. However, results of nonparametric tests should be viewed with caution if the number of censored data points rises above 50%. If censored data are prominent, refer to Helsel & Hirsch (2002) and U.S. Environmental Protection Agency (2006) for a detailed review of estimation and analysis of censored data.

3. Conclusions

This guide presents several graphical methods for exploring and understanding the basic characteristics of TMDL effectiveness monitoring data. The information acquired through data exploration is highly valuable for multiple phases of a TMDL effectiveness monitoring project. During project planning, it is recommended that practitioners undertake exploration of pilot data. At minimum, pilot data should include observations of those water quality parameters that are the focus of TMDL effectiveness monitoring. Pilot data collected from study sites is preferred (though data from comparable sites can also be explored) and should include potential covariate data whenever possible. Exploration of pilot data allows planners to make data-driven decisions regarding which parameters to include in TMDL effectiveness monitoring, and sampling frequency, timing, and duration. Pilot data exploration also provides preliminary direction for post-monitoring confirmatory data analysis and informs selection of a study design (see *Technical Guidance for Designing a TMDL Effectiveness Monitoring Plan*). To facilitate the exploration of pilot data, an Excel-based TMDL effectiveness monitoring planning tool is provided as a complement to this document.

Following project implementation, exploration of TMDL effectiveness monitoring data is required to finalize the selection of methods used for confirmatory data analysis. While highly dependent on data characteristics, confirmatory data analysis decisions should always be made with the analysis objective in mind. Potential analysis objectives for TMDL effectiveness monitoring data are outlined in Table 1 and include:

1. Compare two independent data groups;
2. Compare two data groups with matched pairs;
3. Compare two data groups while adjusting for covariates;
4. Evaluate the relationship between one data group and time;
5. Evaluate the relationship between one data group, time, and covariates.

Analysis objectives 1 and 2 deal with two distinct groups of observations of a single water quality parameter (e.g., dissolved oxygen concentration before and after TMDL implementation). The difference between the two relates to the presence of paired observations between groups. For example, a monitoring program that includes concurrent sampling of upstream and downstream locations provides a group of “upstream” observations with a logical matched pair in the “downstream” group. Alternatively, a study of water quality before and after TMDL implementation provides “before” observations that have no matched pair in the “after” group (the groups are independent). Analysis objective 3 is common for a paired watersheds study, where observed values of a parameter during pre-treatment and post-treatment periods are compared while adjusting for natural variability using observations from the control watershed (the covariate). Analysis objectives 4 and 5 result from trend monitoring, with objective 5 applied if covariates are included in monitoring activities.

A number of confirmatory tests are available to address the analysis objectives discussed above. Alternative tests for a given analysis objective can generally be grouped as parametric or nonparametric tests. Remember that the fundamental difference between the two is that parametric tests assume that data are normally distributed, while nonparametric tests include no assumptions about the distribution of data. Other data characteristics identified through data exploration (e.g., the presence of outliers or censored data) will also influence the selection of a parametric vs. nonparametric test, and the assumptions of each alternative test should be evaluated before proceeding.

Table 1 contains several common parametric and nonparametric tests available to address the potential analysis objectives for TMDL effectiveness monitoring data discussed above. Detailed information on these and other statistical tests is provided in Helsel & Hirsch (2002) and U.S.

Environmental Protection Agency (2006). A number of these tests can be carried out manually using spreadsheet software such as Microsoft Excel (or using Excel's built-in Data Analysis ToolPak). Others require advanced statistical software (such as the free R statistical package).

Table 1. Some useful statistical tests for analysis of TMDL effectiveness monitoring data.

Objective	Study Design	Statistical Test	
		Parametric	Nonparametric
Compare two independent data groups	Before/After; Upstream/Downstream	Two-Sample (Unpaired) t-Test	Rank Sum Test
Compare two data groups with matched pairs	Paired Watersheds; Upstream/Downstream	Paired t-Test	Signed Rank Test
Compare two data groups while adjusting for covariates	Paired Watersheds	Analysis of covariance (ANCOVA)	
Evaluate the relationship between one data group and time (without seasonality)	Trend Monitoring	Linear Regression	Mann-Kendall Test
Evaluate the relationship between one data group and time (with seasonality)	Trend Monitoring	Linear Regression with Seasonal Term	Seasonal Kendall Test
Evaluate the relationship between one data group, time, and other variables (without seasonality)	Trend Monitoring	Multiple Linear Regression	Mann-Kendall Trend Test with LOWESS
Evaluate the relationship between one data group, time, and other variables (with seasonality)	Trend Monitoring	Multiple Linear Regression with Seasonal Term	Seasonal Kendall Test with LOWESS

4. References

Helsel, D. R. (2006). *Computer program for the Kendall family of trend tests: U.S. Geological Survey Scientific Investigations Report 2005–5275.*

Helsel, D. R., & Hirsch, R. M. (2002). Statistical Methods in Water Resources. In *Hydrologic Analysis and Interpretation*. Avaiable online at: http://water.usgs.gov/pubs/twri/twri4a3/.

NIST/SEMATECH , www.itl.nist.gov/div898/. (2003). *Engineering Statistics Handbook*. Available online at: http://www.itl.nist.gov/div898/handbook/.

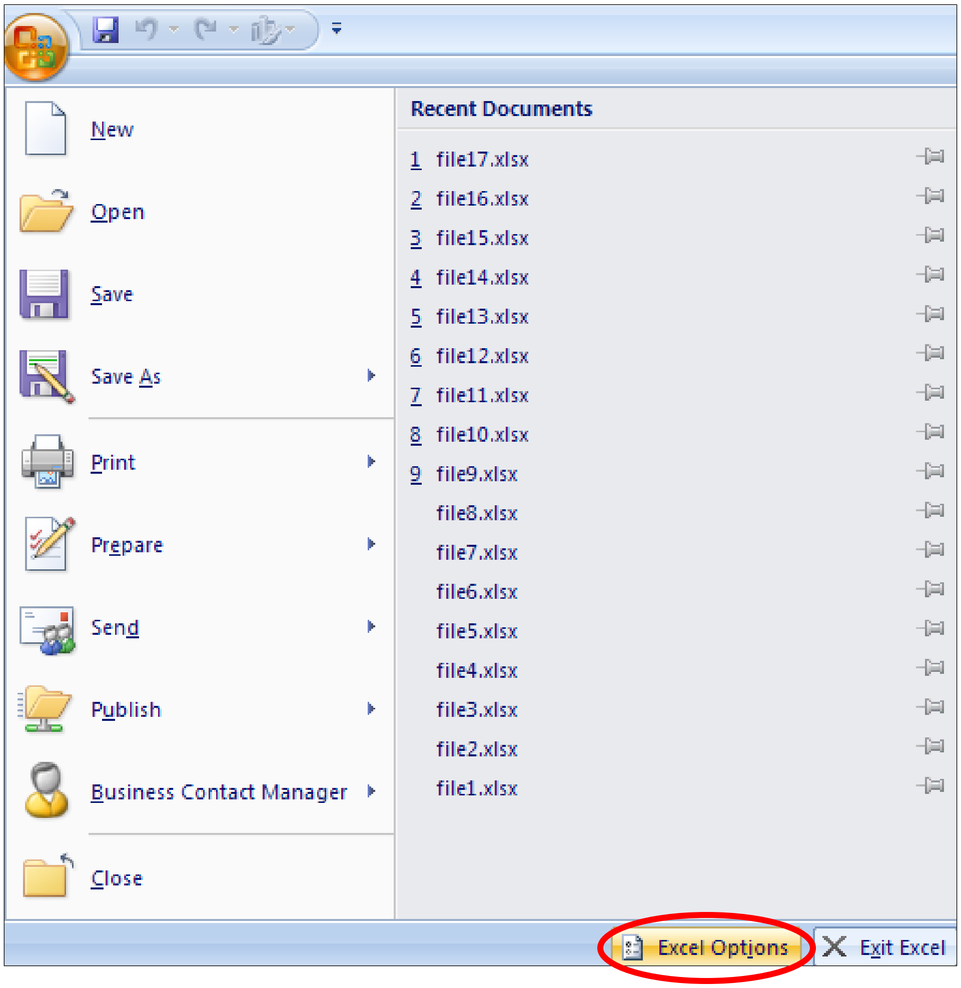
U.S. Environmental Protection Agency. (2006). *Data Quality Assessment: Statistical Methods for Practitioners EPA QA/G-9S*. Available online at: http://www.epa.gov/quality/qs-docs/g9s-final.pdf.

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press. Available online at: http://dspacelocal.library.cornell.edu/DSpace/IFUP_Catalog.pdf.

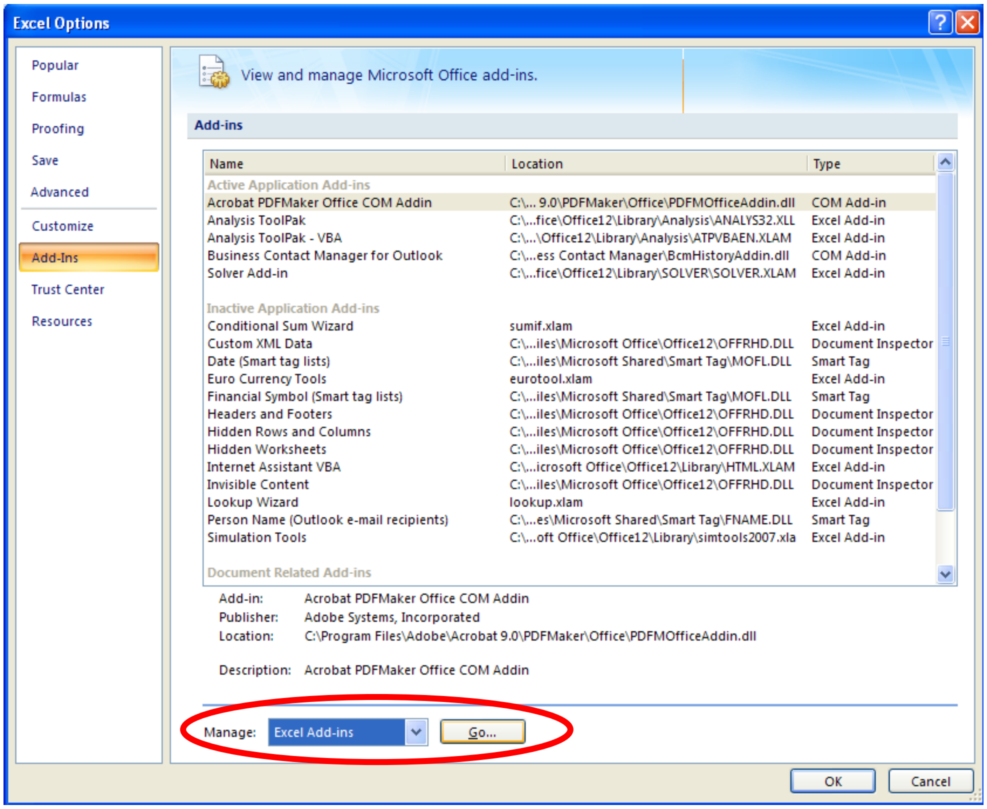
Appendix A. Loading the Analysis ToolPak in Microsoft Excel 2007

Microsoft Excel 2007 includes several built-in data analysis tools. These tools are not accessible to users until the Analysis ToolPak is installed. Below are step-by-step instructions for installing the Analysis ToolPak.

Step 1. Click the Office Button at the top-left of the screen. Click *Excel Options*.



Step 2. In the *Excel Options* box, click the *Add-Ins* tab on the left. In the *Manage* drop-down list, select *Excel Add-ins*. Click *Go…*



Add-ins

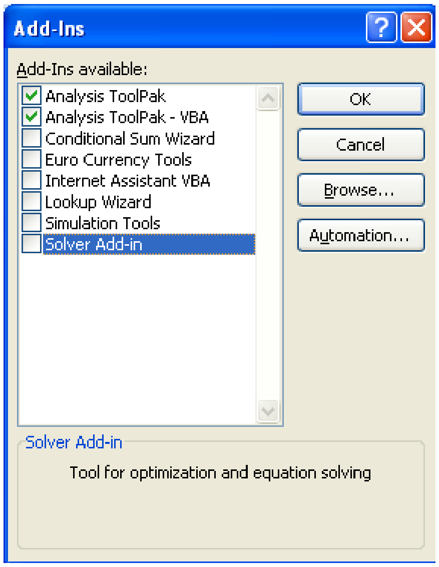
Name	Location	Type
Active Application Add-ins		
Acrobat PDFMaker Office COM Addin	C:\... 9.0\PDFMaker\Office\PDFMOfficeAddin.dll	COM Add-in
Analysis ToolPak	C:\...fice\Office12\Library\Analysis\ANALYS32.XLL	Excel Add-in
Analysis ToolPak - VBA	C:\...\Office12\Library\Analysis\ATPVBAEN.XLAM	Excel Add-in
Business Contact Manager for Outlook	C:\...ess Contact Manager\BcmHistoryAddin.dll	COM Add-in
Solver Add-in	C:\...fice\Office12\Library\SOLVER\SOLVER.XLAM	Excel Add-in
Inactive Application Add-ins		
Conditional Sum Wizard	sumif.xlam	Excel Add-in
Custom XML Data	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Date (Smart tag lists)	C:\...iles\Microsoft Shared\Smart Tag\MOFL.DLL	Smart Tag
Euro Currency Tools	eurotool.xlam	Excel Add-in
Financial Symbol (Smart tag lists)	C:\...iles\Microsoft Shared\Smart Tag\MOFL.DLL	Smart Tag
Headers and Footers	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Hidden Rows and Columns	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Hidden Worksheets	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Internet Assistant VBA	C:\...icrosoft Office\Office12\Library\HTML.XLAM	Excel Add-in
Invisible Content	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Lookup Wizard	lookup.xlam	Excel Add-in
Person Name (Outlook e-mail recipients)	C:\...es\Microsoft Shared\Smart Tag\FNAME.DLL	Smart Tag
Simulation Tools	C:\...oft Office\Office12\Library\simtools2007.xla	Excel Add-in
Document Related Add-ins		

Add-in: Acrobat PDFMaker Office COM Addin
Publisher: Adobe Systems, Incorporated
Location: C:\Program Files\Adobe\Acrobat 9.0\PDFMaker\Office\PDFMOfficeAddin.dll

Description: Acrobat PDFMaker Office COM Addin

Manage: Excel Add-ins

Step 3. In the *Add-Ins* box, check the boxes next to *Analysis ToolPak* and *Analysis ToolPak – VBA*. Click *OK*.



Add-Ins available:

- ☑ Analysis ToolPak
- ☑ Analysis ToolPak - VBA
- ☐ Conditional Sum Wizard
- ☐ Euro Currency Tools
- ☐ Internet Assistant VBA
- ☐ Lookup Wizard
- ☐ Simulation Tools
- ☐ Solver Add-in

Solver Add-in

Tool for optimization and equation solving