Exploring Environmental Data

Dr. Robin Matthews, Professor Emeritus,
Environmental Sciences Department
Western Washington University

May 15, 2023

Introduction to Environmental Data Exploration

- Scientists spend a large amount of time planning research projects and collecting data
- Equally important, however, is learning how to display data clearly, without bias, in an appropriate format
- Unfortunately, it is all too easy to obscure important elements from research, intentionally or unintentionally
- Here are a few illustrations of how the selection and formatting in figures can hide or highlight results

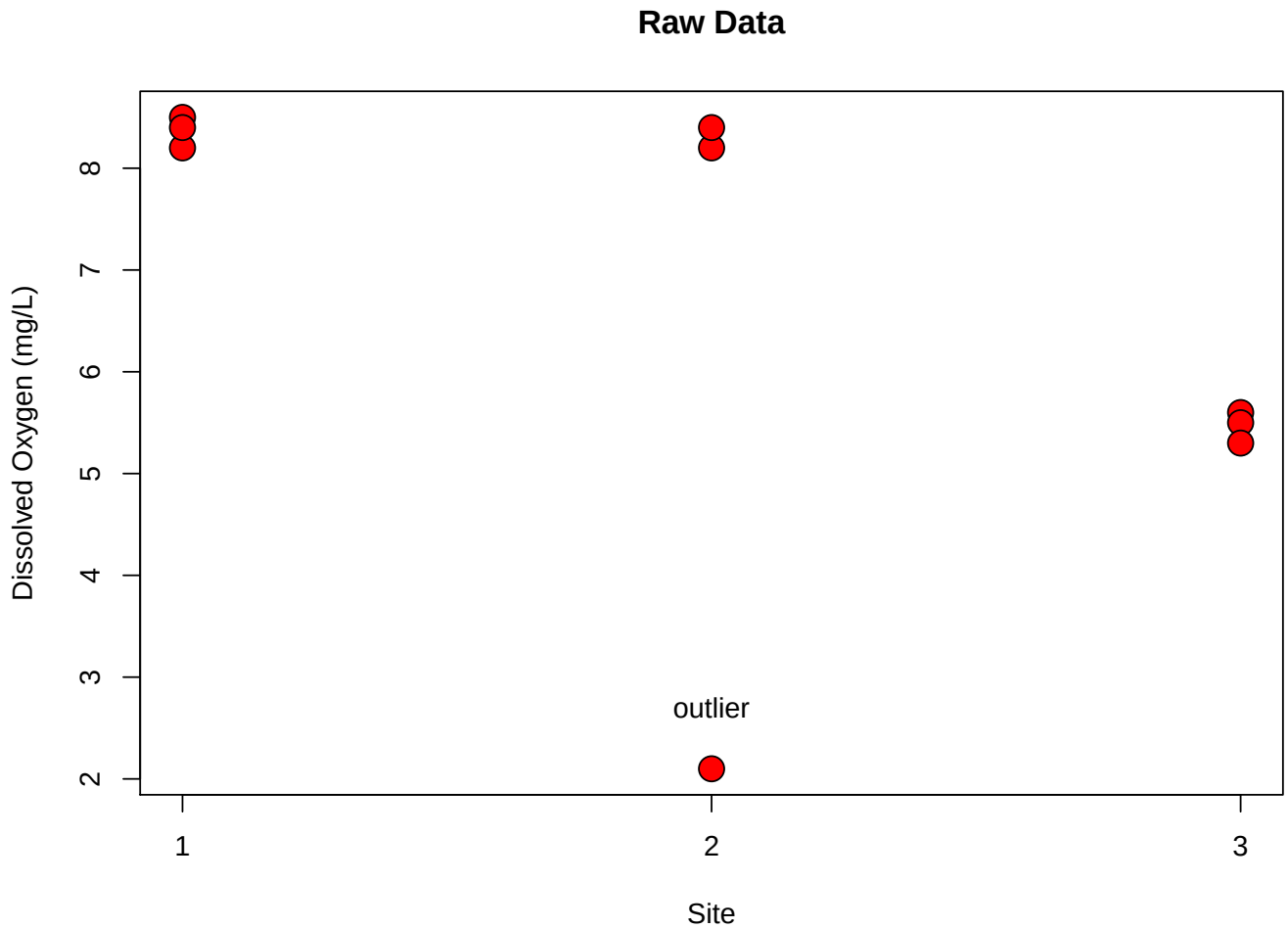
First - Some Common Data Set Problems

- Variables are chosen based on whether the scientist thinks the measurement is appropriate
  - Data collection is also influenced by physical constraints (e.g., money, people, equipment, time); biological constraints (e.g., population cycles, distribution of test organisms); and ethical constraints (e.g., endangered species, experimentation on sentient organisms)
- Don't assume that all measured parameters will show an effect
- Don't assume that all important parameters were measured
- Variation should be expected; causation should not!

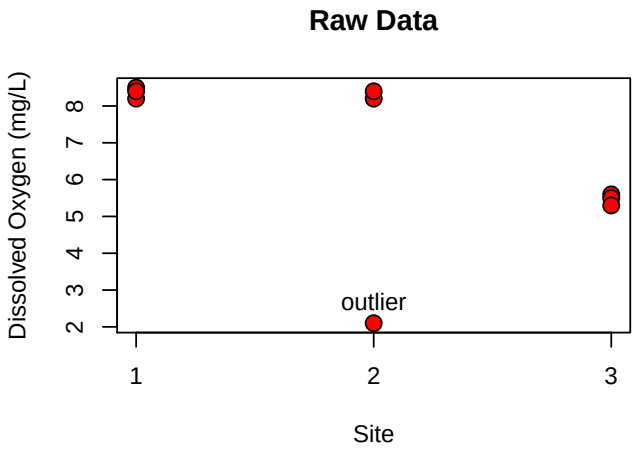
Review of Basic Statistical Approaches

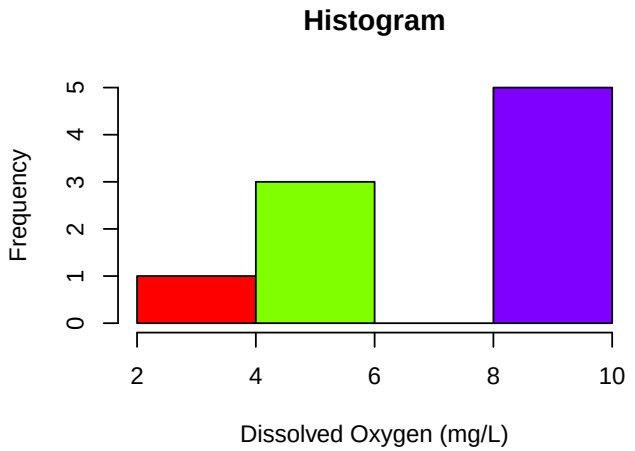
Exploratory vs. Confirmatory Statistics

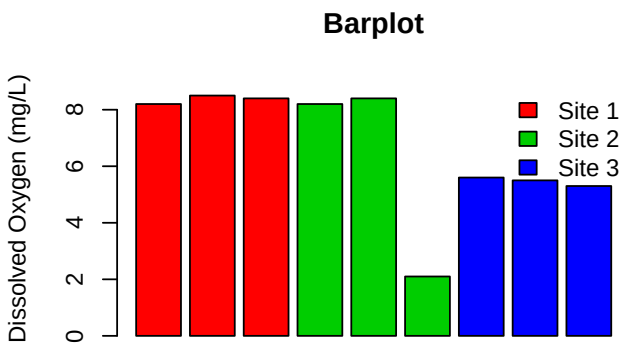
- Fundamental statistical questions are usually:
  - What is happening? (exploratory)
    - Examples: descriptive statistics, plots, distributions, clustering, PCA
  - Is it real? (confirmatory)
    - Examples: t-test, regression/correlation, ANOVA, LDA
- Regardless of which approach you choose, you should start your data analyses using simple descriptive statistics and plotting

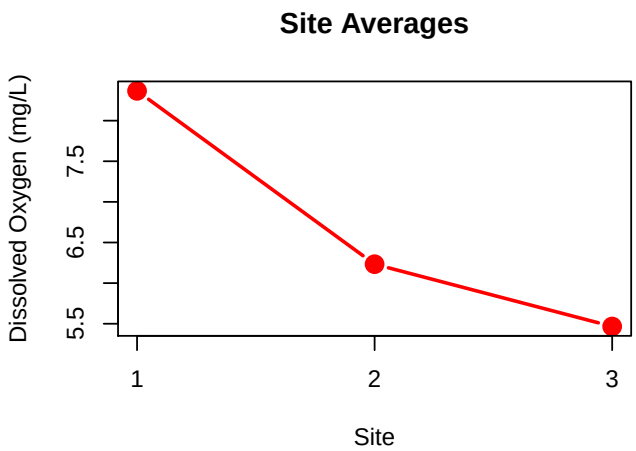


Here is an example where three measurements of dissolved oxygen were collected at three different sites. This figure shows the actual measurements. Can you describe the pattern?

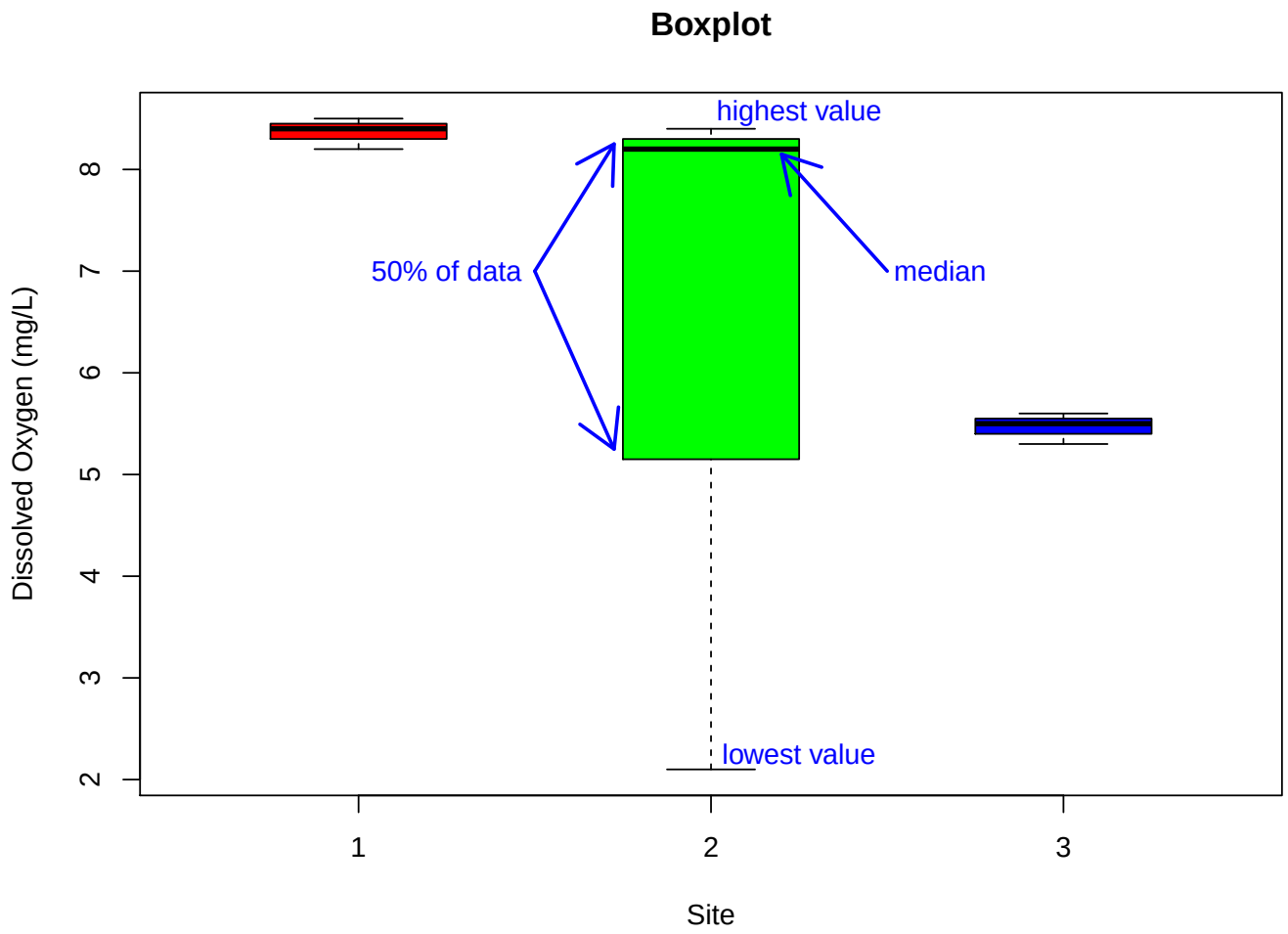








Here are the original data with three summary plots, all of which are poor illustrations of how dissolved oxygen changes at the three sites



This *boxplot* is a good scientific summary figure because it shows the influence of the outlier at Site 2 as well as the similarity between Site 1 and 2; however, it may not be the best way to summarize data to non-scientists

Memory Used for Processing Visual Information

We use three basic types of memory to process scientific information:

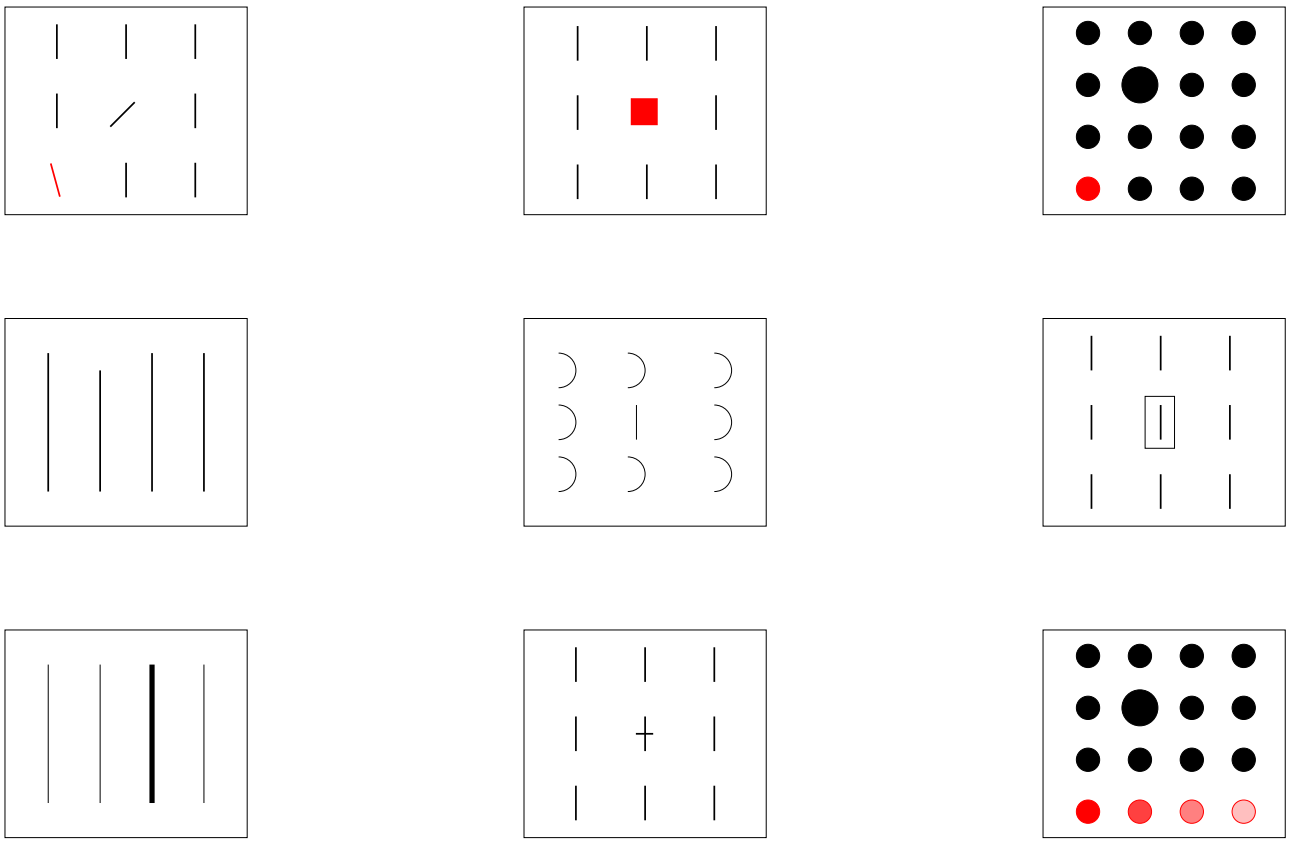
- *Iconic memory* for detecting visual information
  - *Pre-attentive processing*
- *Short-term memory* for temporary (limited) storage
  - *Attentive or perceptual processing*
  - Limited to $\sim$ 3–9 items
- *Long-term memory* for retaining information
  - Long-term memory can be created consciously or unconsciously
  - Information is stored more permanently, with cross-links that allow access back into short-term memory
  - Required for recognizing images, interpreting words and numbers, understanding context

Pre-Attentive Processing of Visual Information

- **Pre-attentive processing** (iconic memory) provides quick, subconscious processing of graphical information and is influenced by variations in:
  - form
  - color
  - spatial position
  - motion
- Graphics that make use of these features tend to make a strong impression on us, even when we don't know why
- In creating visual displays of scientific information, the careful use of color, shape, position, and other design effects can **emphasize** or de-emphasize information in the figure

Pre-Attentive Processing

Pre-attentive processing involves high-speed, subconscious processing of variations in form, color, spatial position, and motion



Figures modified from Show Me The Numbers by Stephen Few, Analytics Press, 2004

Example using Color and Form

How many zeros are there?

6	4	4	2	1	5	7	2	2	2	2	8
9	8	9	3	6	5	5	5	7	8	7	6
1	3	5	9	5	6	0	6	7	6	6	6
7	4	2	5	7	7	1	5	5	5	4	2
5	2	1	1	4	2	6	6	4	9	6	3
5	7	2	0	6	1	6	8	0	6	0	2
9	8	7	4	4	5	4	4	9	1	5	1
2	1	3	7	8	6	2	0	2	9	4	9
3	4	9	6	2	1	7	9	4	8	2	8
2	5	5	2	2	4	5	5	8	7	1	5

How many zeros are there?

6	4	4	2	1	5	7	2	2	2	2	8
9	8	9	3	6	5	5	5	7	8	7	6
1	3	5	9	5	6	**0**	6	7	6	6	6
7	4	2	5	7	7	1	5	5	5	4	2
5	2	1	1	4	2	6	6	4	9	6	3
5	7	2	**0**	6	1	6	8	**0**	6	**0**	2
9	8	7	4	4	5	4	4	9	1	5	1
2	1	3	7	8	6	2	**0**	2	9	4	9
3	4	9	6	2	1	7	9	4	8	2	8
2	5	5	2	2	4	5	5	8	7	1	5

Here is an example of how pre-attentive color processing helps clarify tabular information

Perceptual Processing of Visual Information

- **Perceptual processing** (conscious interpretation of visual information) is based on perceptions of difference rather than the ability to recognize absolute values
- As a result, it is quite easy to fool our visual perception of data using simple optical tricks and illusions
- Some of the best examples of this are found in advertising
  - One trick for staging houses for sale involves tricking our perception of room size by adding mirrors, light colored paint, and removing most of the furniture to make rooms look larger
  - Another advertising trick is to place warnings or qualifying information at the bottom of the screen in small, light colored lettering, against a backdrop of screen motion. (Just *try* to stay focused on the text at the bottom of your TV screen!)
- Perceptual processing in essential for understanding data in tables and written discussion, but remember that it can slow down our ability to interpret information

Designing Effective Figures

Example #1 - Drayton Harbor Coliforms

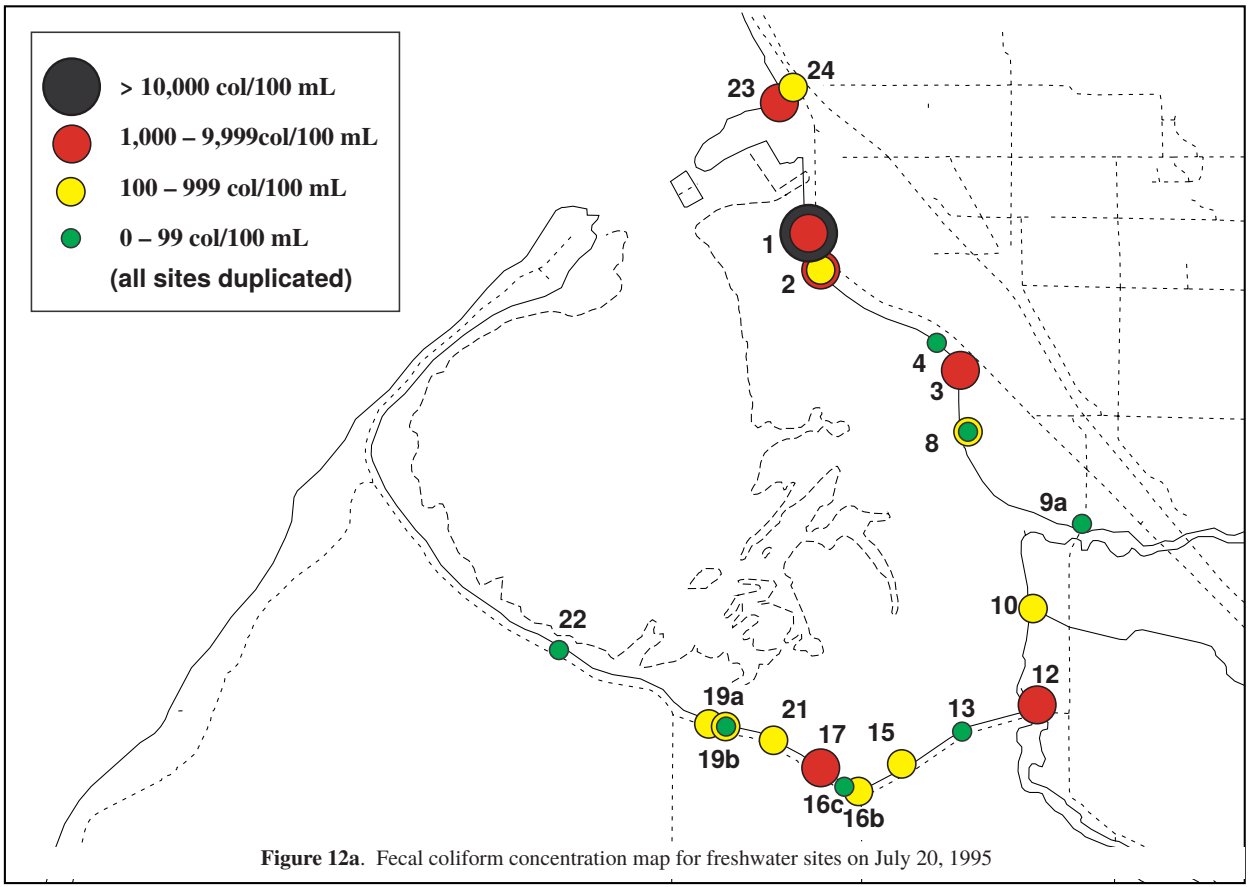


Figure 12a. Fecal coliform concentration map for freshwater sites on July 20, 1995

This figure has a few bad features (e.g., key is distracting), but uses pre-attentive processing techniques effectively

Designing Effective Figures

Example #1, continued

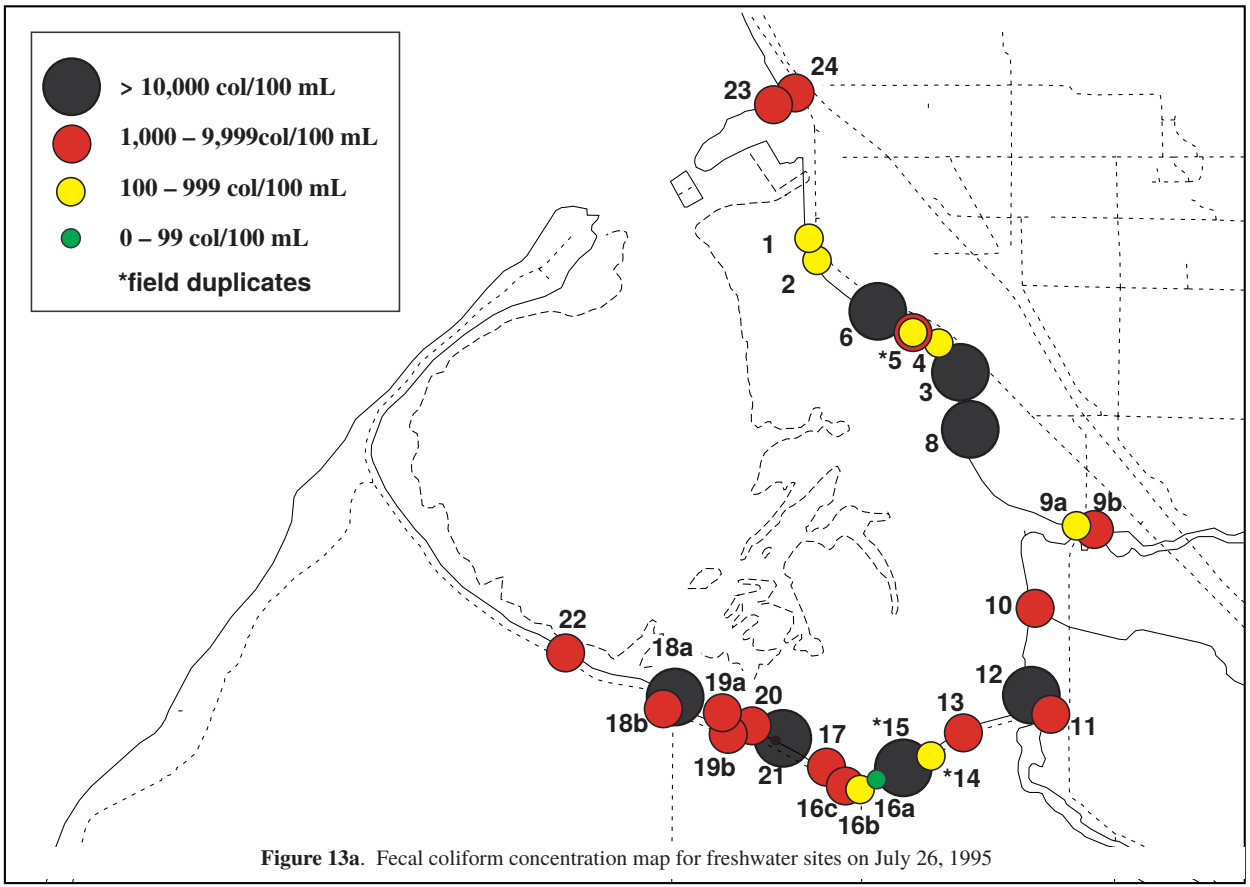
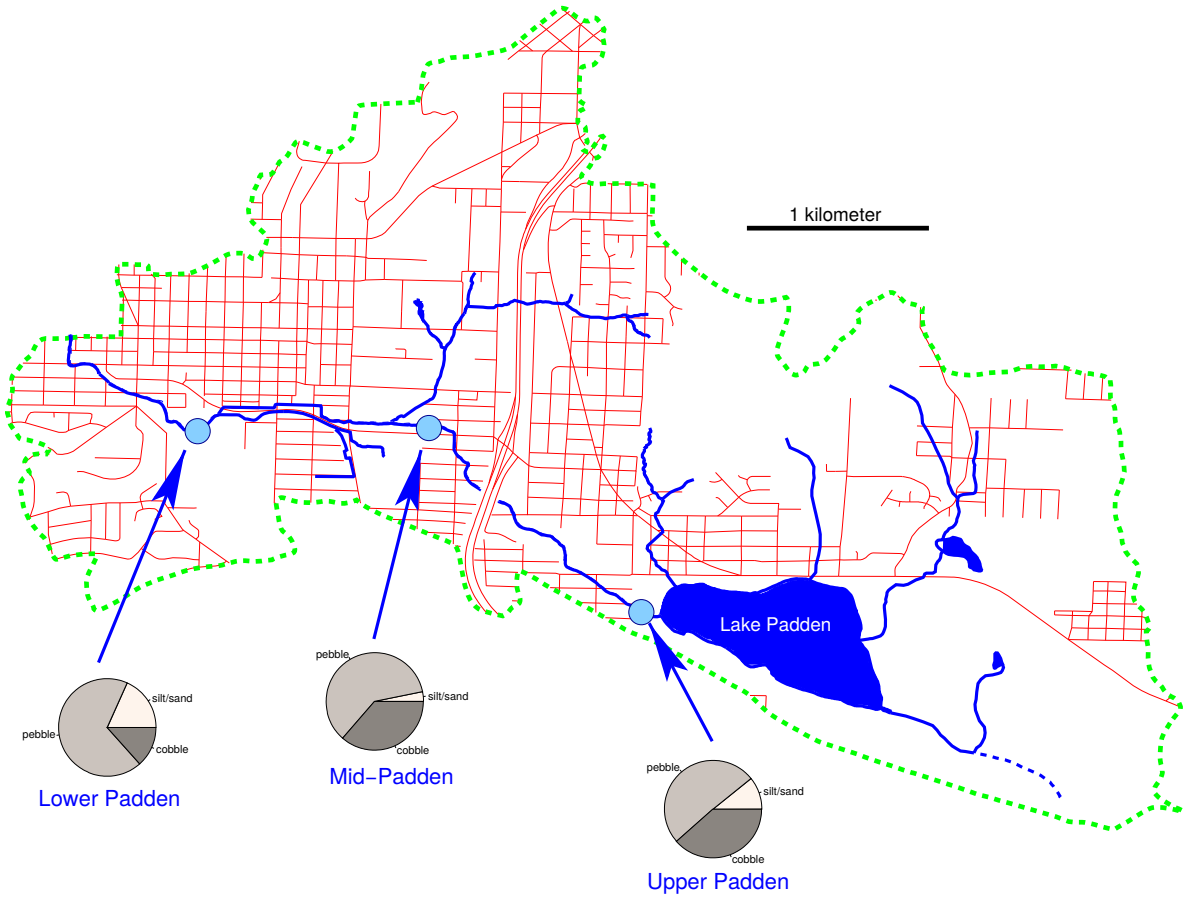


Figure 13a. Fecal coliform concentration map for freshwater sites on July 26, 1995

Note how the large red and black circles emphasize the much higher coliform counts on July 26 compared to July 20

Designing Effective Figures

Example #2, Padden Creek Sediments

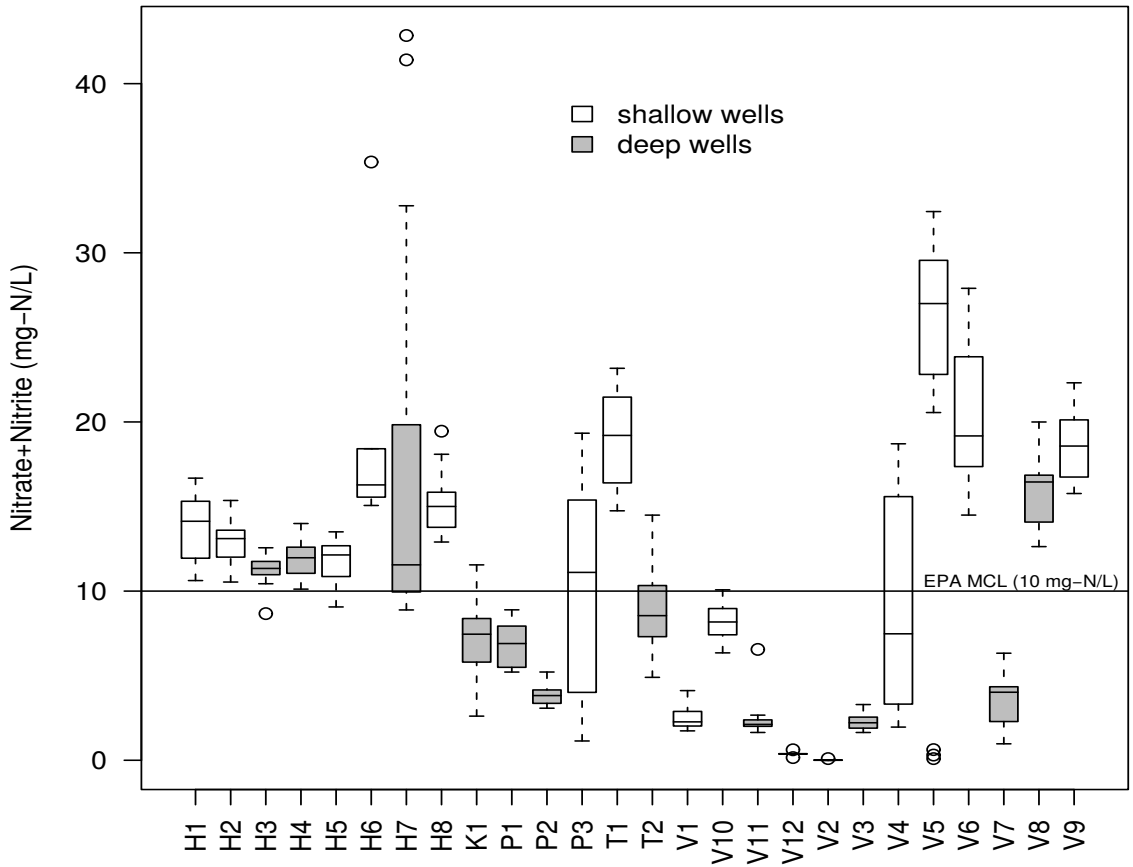


Padden Creek sediment size distributions, summer 1998 (adapted from Hachmoeller, 1988)

Here is example that uses boxplots to display data. What are other options for presenting the information?

Designing Effective Figures

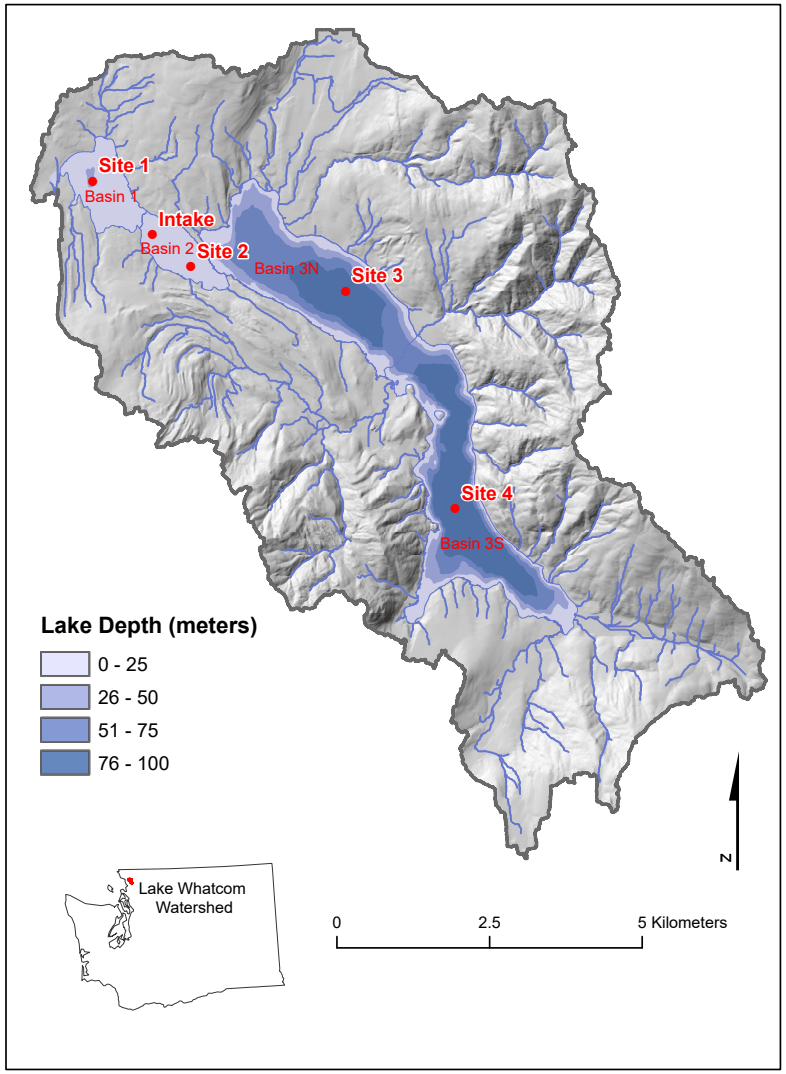
Example #3, Nitrate Levels in Abbotsford/Sumas Wells



This figure contains a large amount of summary information. What can you determine from the graphics?

Designing Effective Figures

Example #4 - Are Cyanobacteria Increasing in Lake Whatcom?



Designing Effective Figures

Example #4, Scientific background:

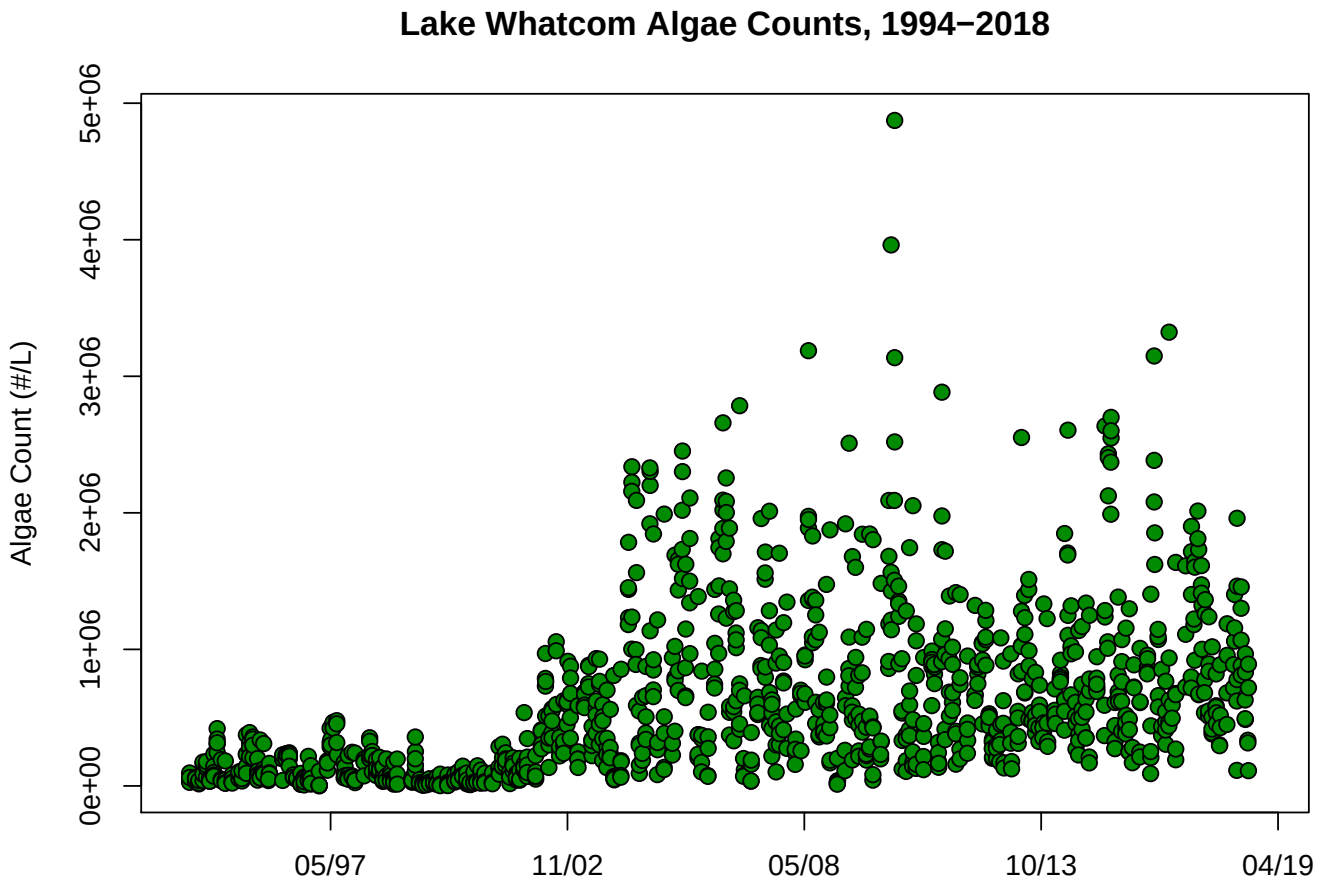
- Cyanobacteria (aka bluegreen algae) are water quality indicators that often increase in abundance in polluted lakes
- Researchers associated with Western Washington University have been counting algae in Lake Whatcom for more than 20 years
- Algae samples were collected monthly (except Jan and Mar) at 5 meters below the surface at four sites in the lake as part of a long-term monitoring project
  - Prior to 1994, most of the counts were collected by graduate students, with minimal coordination between different individuals. After 1994, the counts followed a more consistent procedure
- Intensive sampling (more depths, more sites, more frequent sampling) has been done by various researchers, but the data are sporadic and there is little, if any, similarity between methods

Designing Effective Figures

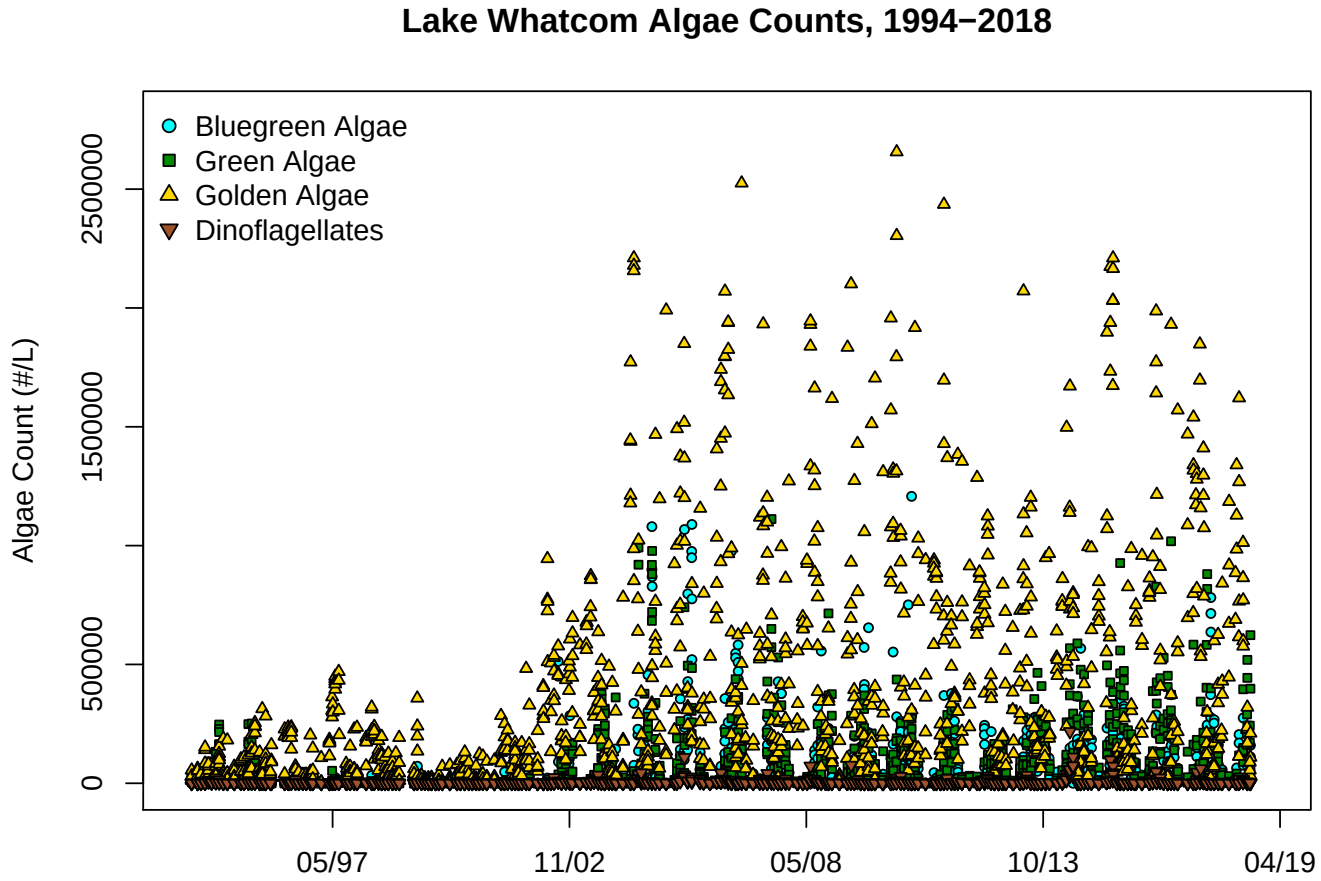
Example #4 - Preliminary Data Analysis Decisions

Assessing long term trends requires that we isolate "time" as the factor influencing algae counts, and reduce or eliminate other factors that might cause changes in the counts

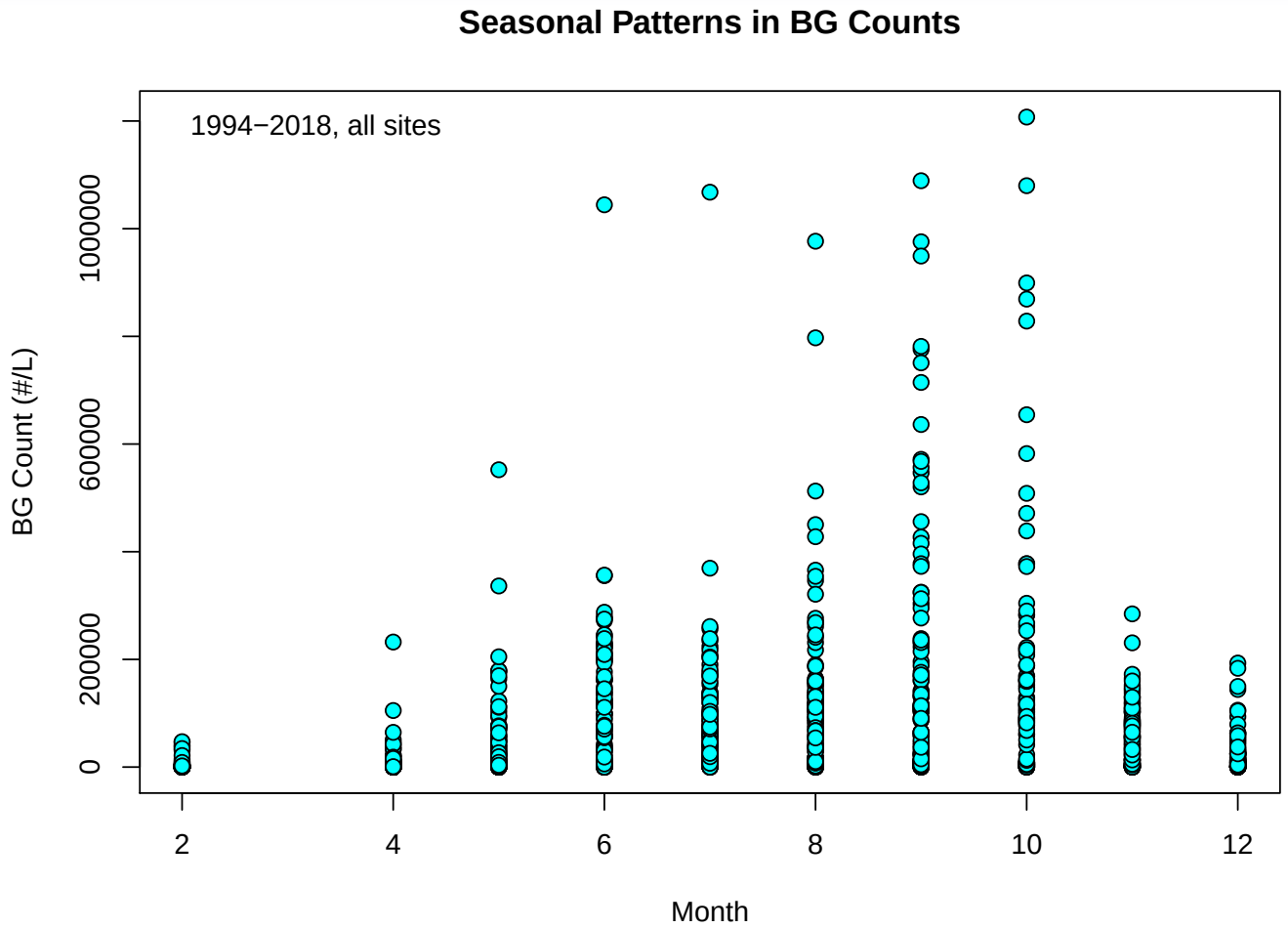
- Decision #1: Use data beginning in 1994 to minimize influence of different counting techniques
- Decision #2: Use data from monthly sampling to minimize influence of different sampling intensities
- Decision #3: Use data from Sites 1–4 collected 5 meters below the surface to minimize differences due to depth and location in the lake



Here is a plot of the raw algae counts from 1994–2018, without distinguishing site or type of algae. (Remember, our goal is to say whether the bluegreen algae are increasing, not whether all algae are increasing)

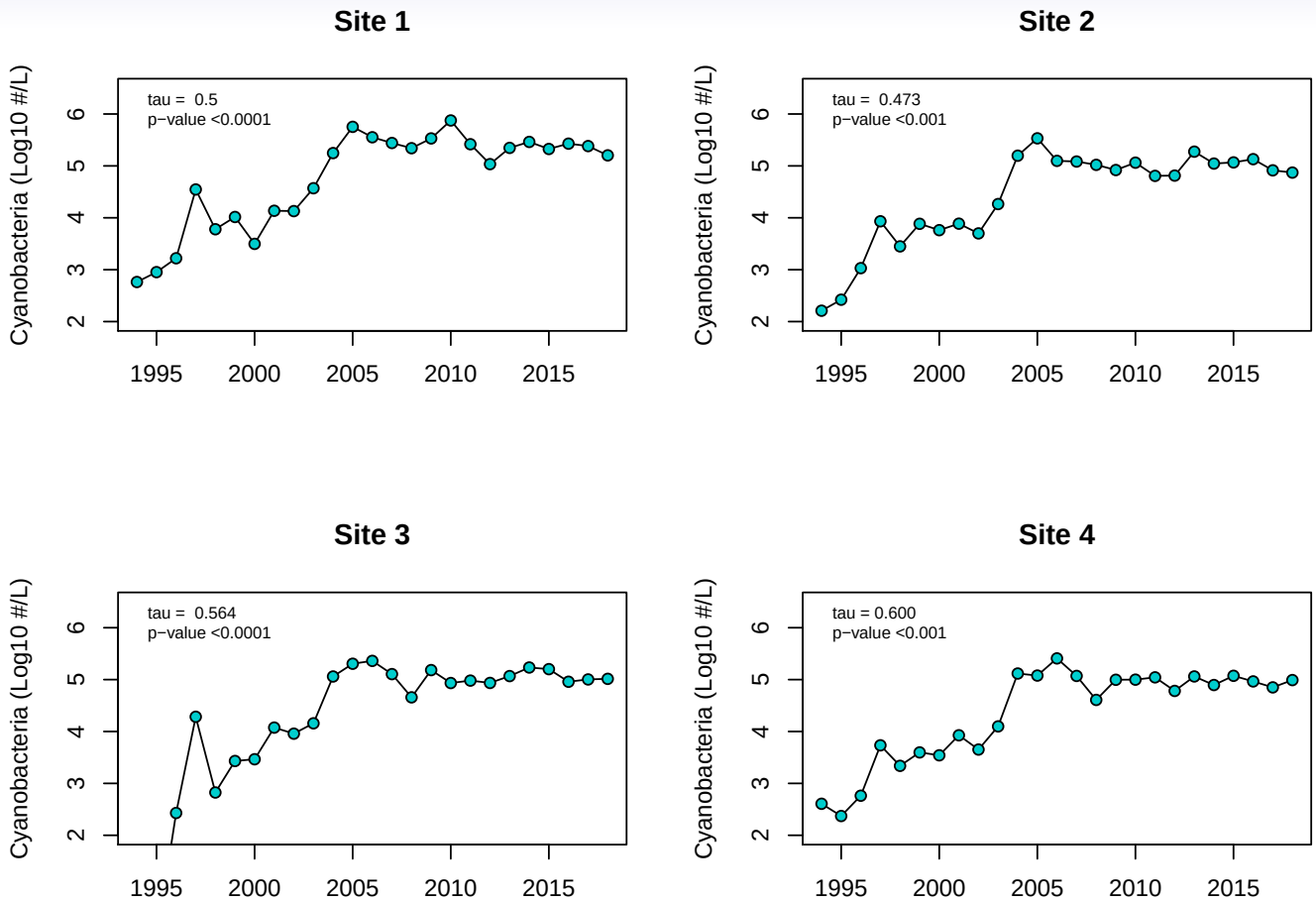


If we assign plotting codes to separate the bluegreen algae from the other major algal types, we begin to see a potential problem. Bluegreens are not the dominant type of algae in the lake . . . golden algae (diatoms) are much more abundant. Also, algal counts are very "noisy" (the counts have a large range)



Here is another problem: the previous figures included counts from all months and depths. Bluegreen algae are only abundant in near the lake surface during summer and early fall

One of the easiest ways to eliminate a trend is to average in data that you know won't show any patterns



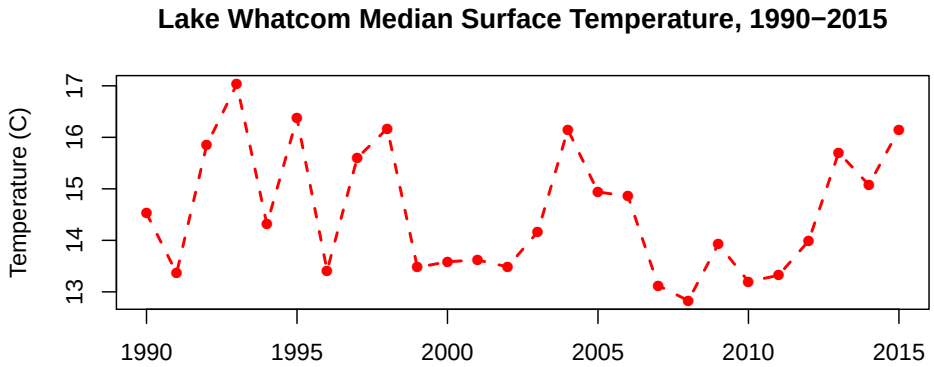
Here are the figures used for the annual reports, showing near-surface medians (June-October, $\leq$5 meters), with correlation results confirming that cyanobacteria are increasing significantly with time. Three items should be apparent: the cyanobacteria counts increased between 1994 and 2005; the correlation with year is strongest at Sites 3–4; the counts have been stable for about half the time period, so the statistical analysis *should* be divided into pre- and post-2005

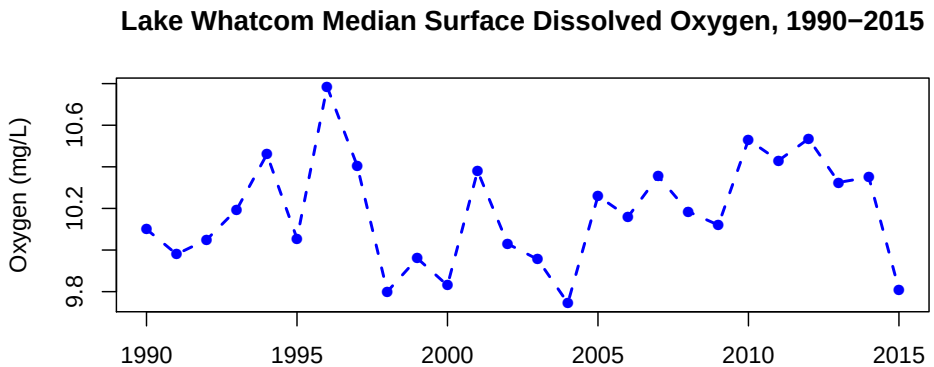
Designing Effective Figures

Example #5 - Are Temperature and Oxygen Levels Changing in Lake Whatcom?

- Preliminary data from the early 1990s suggested that oxygen concentrations were declining near the bottom of Lake Whatcom
- Researchers associated with Western Washington University have been collecting temperature and oxygen data in Lake Whatcom since the 1960s
- Beginning in late 1989, the sampling and analysis methods were standardized
    - Decision #1: Use data collected from 1990 to ensure similar sample sizes for all years

Lake Whatcom Temperature and Oxygen Data

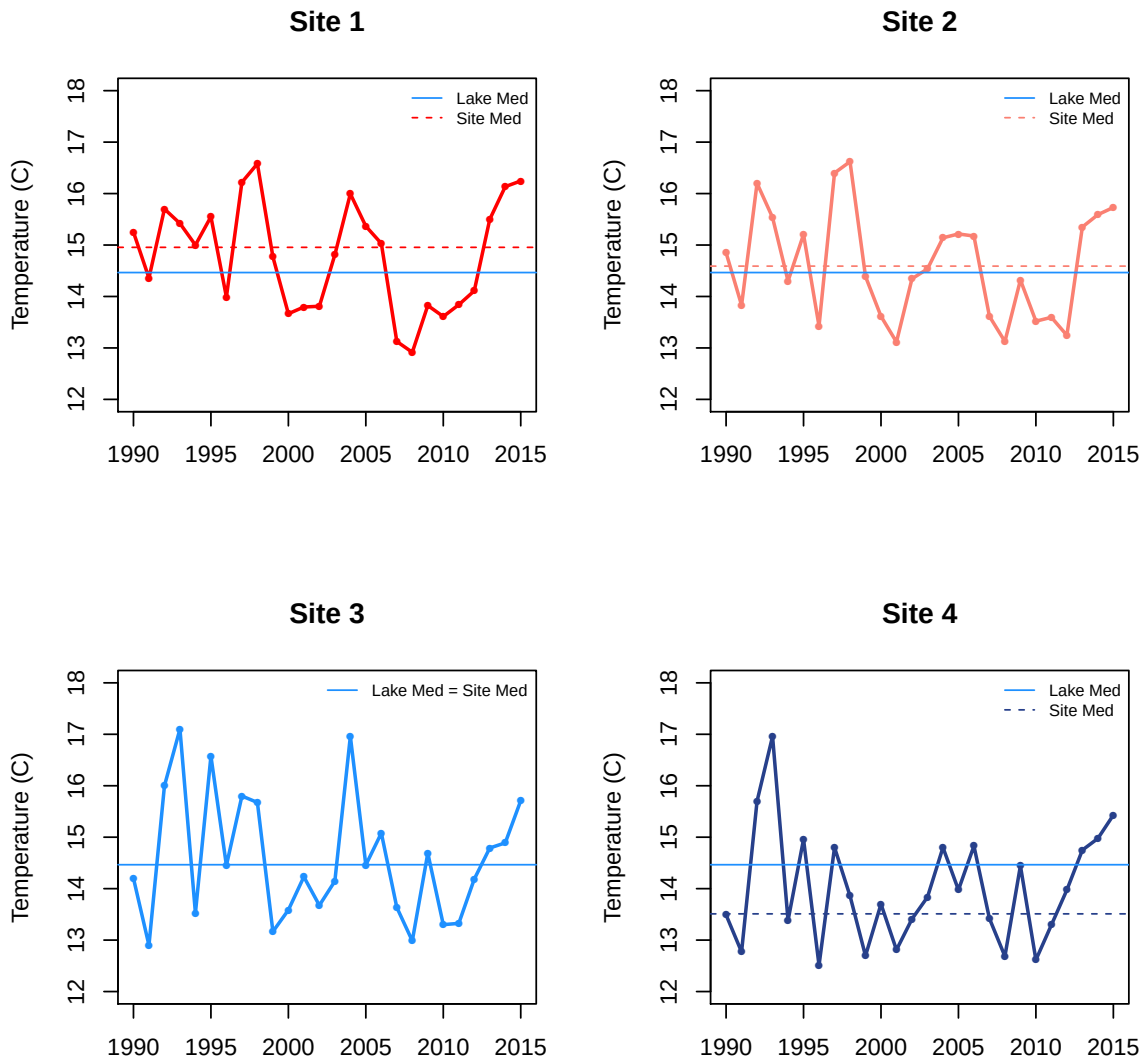




Here is an example using simple x-y scatterplots to look for patterns in Lake Whatcom surface temperature and dissolved oxygen levels. This figure shows median surface temperature and oxygen levels for all sites and months ($\sim$40 samples per year)

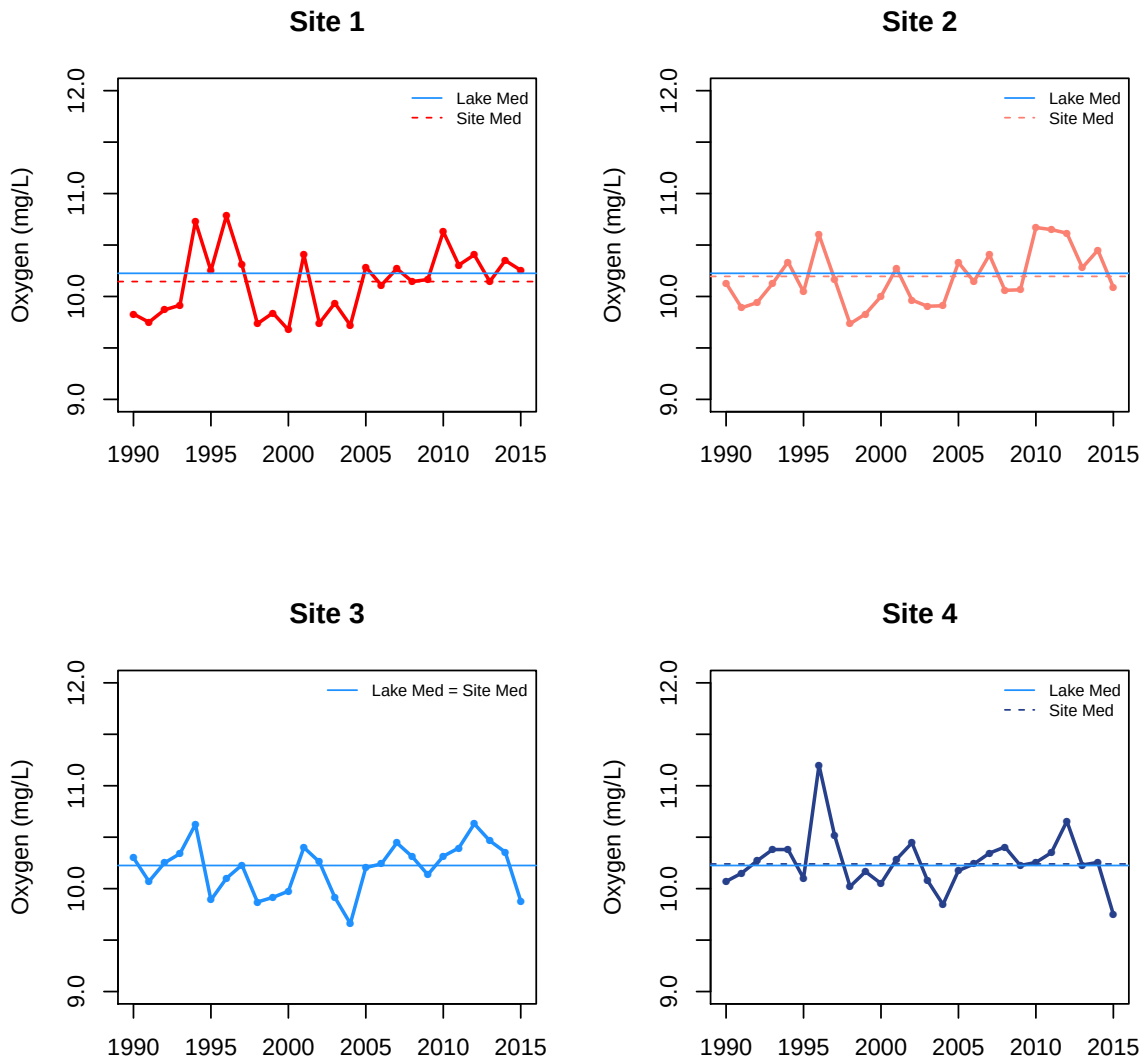
Lake Whatcom Temperature Data

Influence of Spatial Location Revealed Using Trend Lines



Lake Whatcom Oxygen Data

Influence of Spatial Location Revealed Using Trend Lines



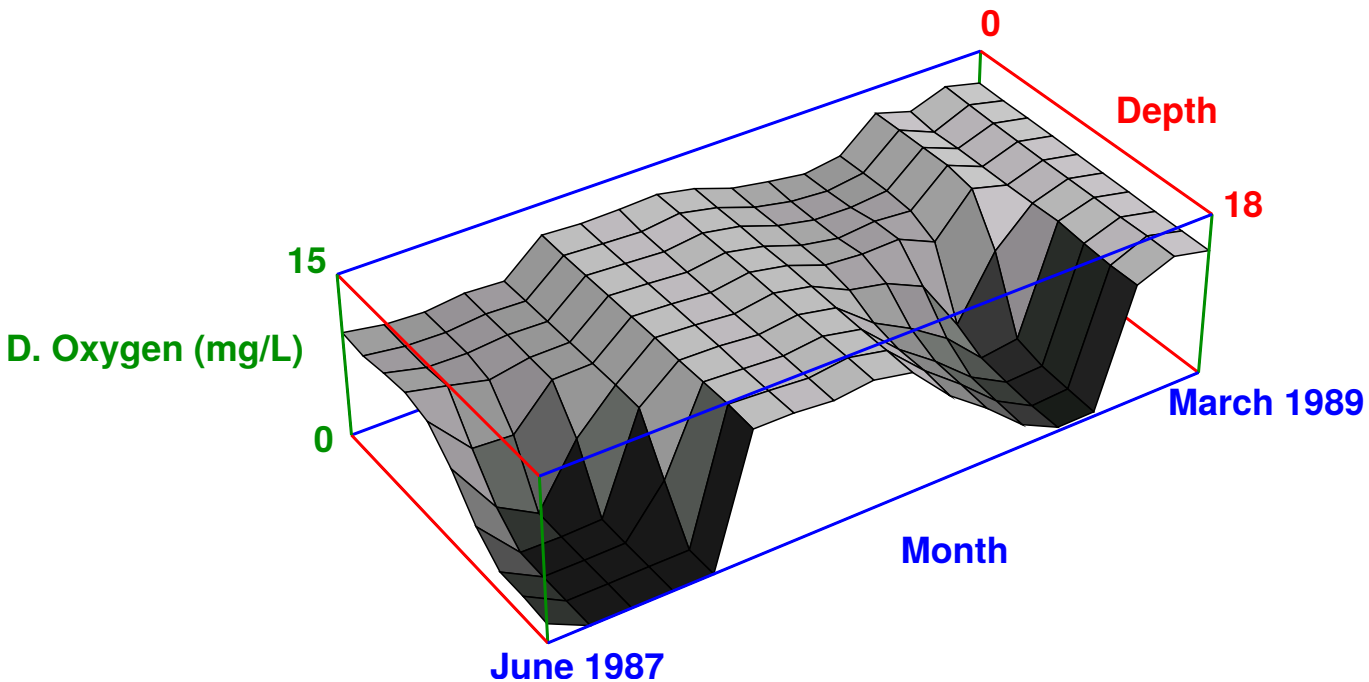
Lake Whatcom Temperature and Oxygen Data

Example #5, continued

- The lake is very large and deep (>100 meters). The majority of the lake (96%) maintains consistently high oxygen concentrations
- The temperature data, dating back to the 1960s, do not show a clear trend
- The shallow sites (Sites 1–2), which are from basins that collectively represent only 4% of the lake volume, both experience severe oxygen depletion during the summer
- Only Site 1 appears to have decreasing oxygen concentrations, and this was only present in summer and late fall in samples from >10 meters deep
- Decision #2: Use oxygen data collected at Site 1 during the period of lake stratification. Oxygen data from 1988-1989 can be included with these selection criteria

Lake Whatcom Oxygen Data

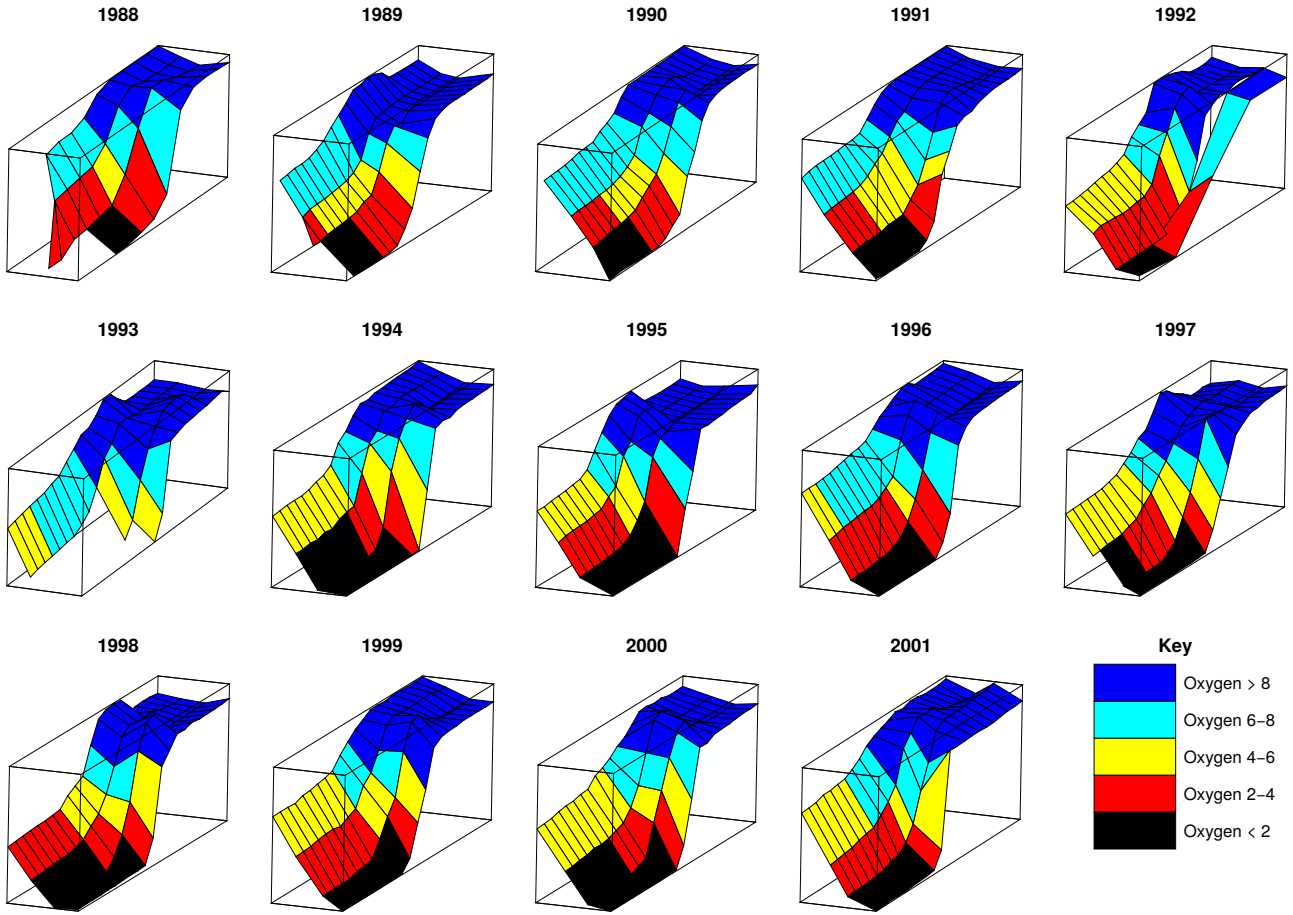
3D Plot of Oxygen, Depth, and Time (1987–1989)



This figure shows Lake Whatcom dissolved oxygen at Site 1 by time and depth for two sequential summers. Although this figure reveals some seasonal and depth-related differences, the patterns are $\Rightarrow$difficult to interpret

Lake Whatcom Oxygen Data

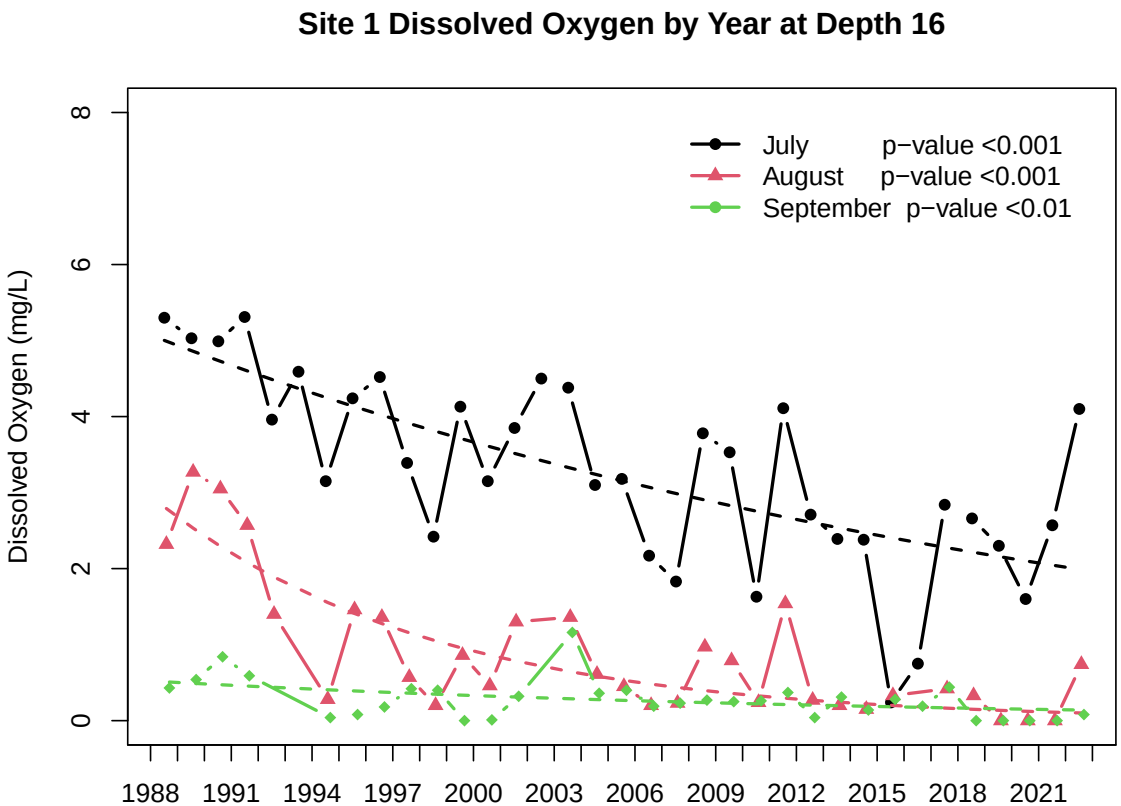
3D Plot of Oxygen, Depth, and Time (1988–2001)



While colorful, this figure is impossible to interpret in anything other than very broad terms

Lake Whatcom Oxygen Data

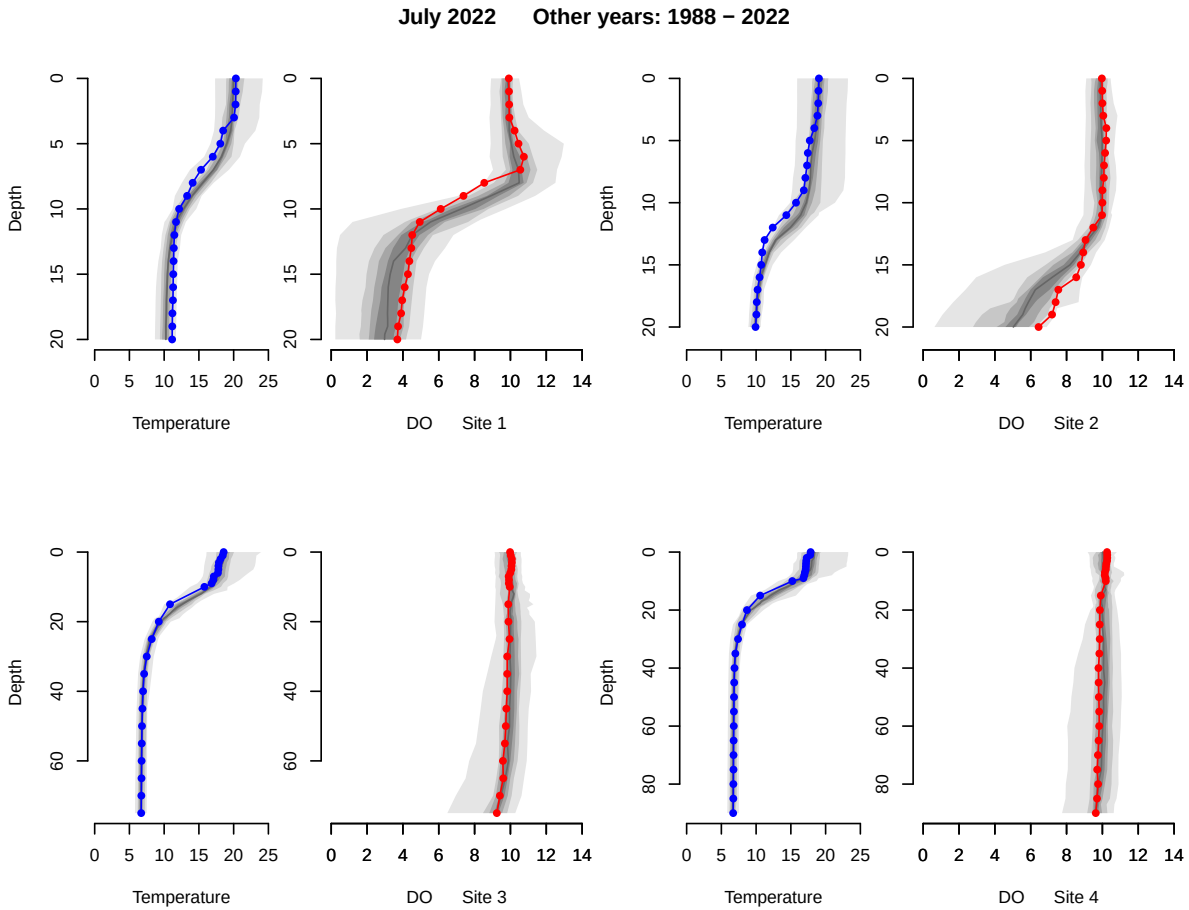
Back to Basics - Simple Scatterplots



The important pattern in the two preceding figures is that there are different rates of oxygen loss during the summer in the deep samples. Rather than try to show all depths and dates, this figure focuses on the changes that are occurring at specific depths and times

Lake Whatcom Oxygen Data

Back to Basics - Simple Scatterplots



This figure shows July 2022 temperature and oxygen profiles superimposed on polygons shaded to show quartiles on either side of the median (historic ranges)

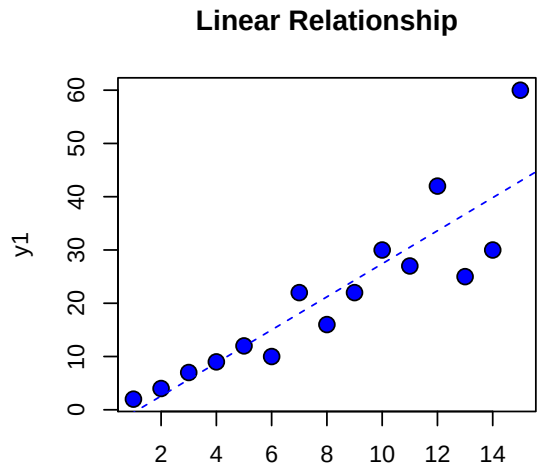
Correlation and Regression

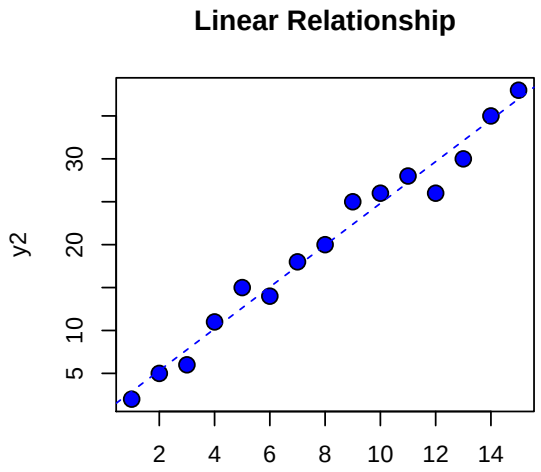
The Foundation of Multivariate Analysis

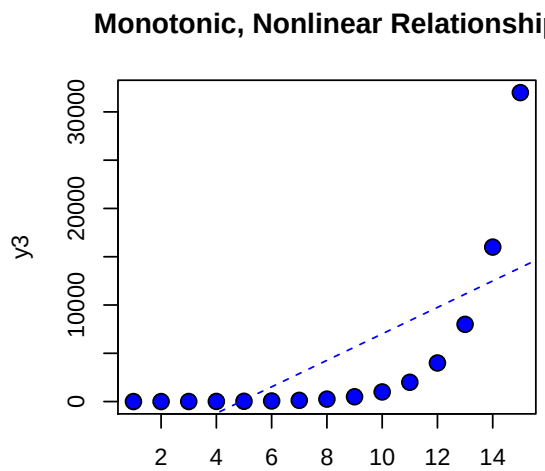
- Both measure *monotonic* relationships (see page 34)
- Regression measures the relationship between an independent variable (x) and one or more dependent variables (y_i)
  - Used to predict (model) unmeasured values of the dependent variable(s)
- Correlation analysis measures the relationship between two variables that are not necessarily functionally dependent
  - Used to explore patterns in measured variables and to identify *indicators* that predict responses in other variables
- Parametric and nonparametric versions are available for both regression and correlation analysis

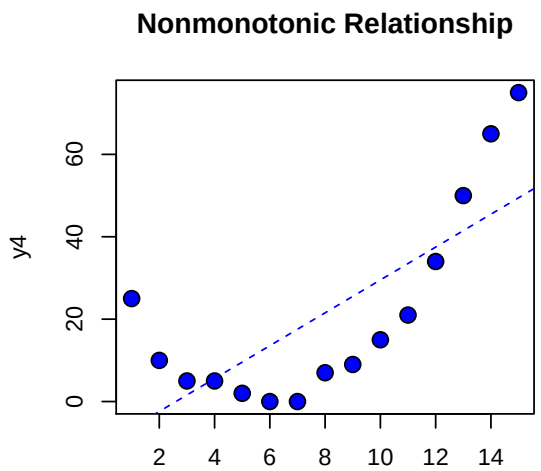
Correlation/Regression Summary

Examples of Monotonic Relationships



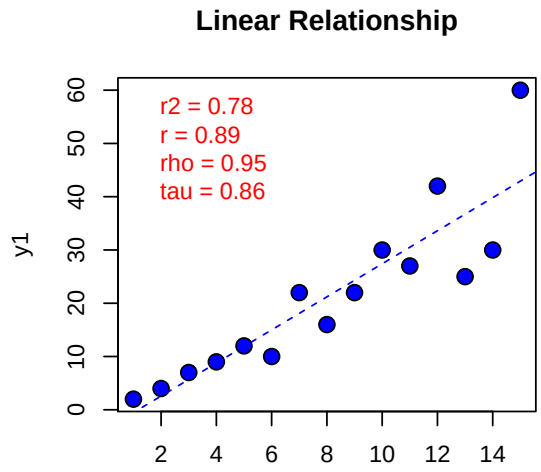


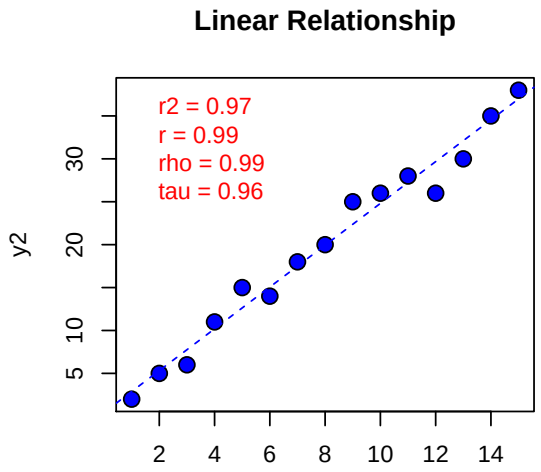


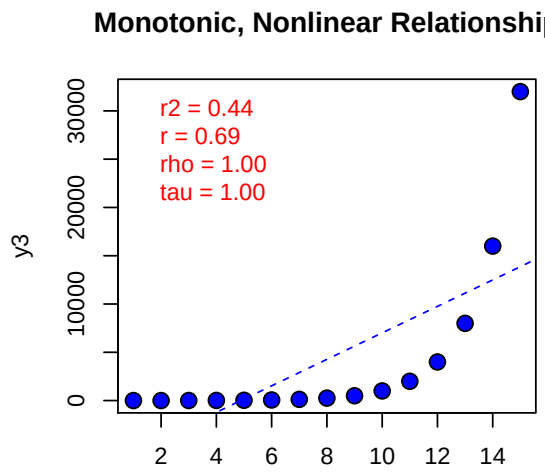


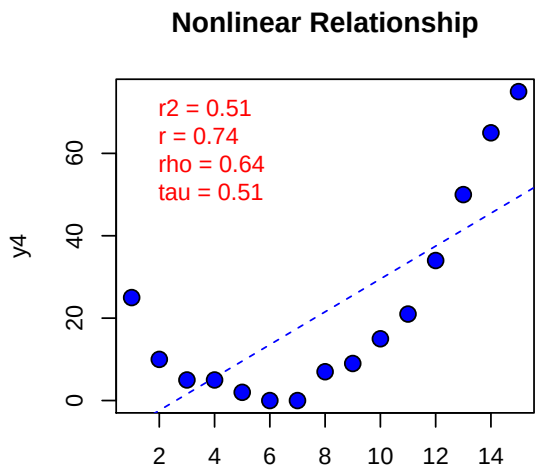
Correlation/Regression Summary

Parametric vs. Nonparametric Correlation Statistics



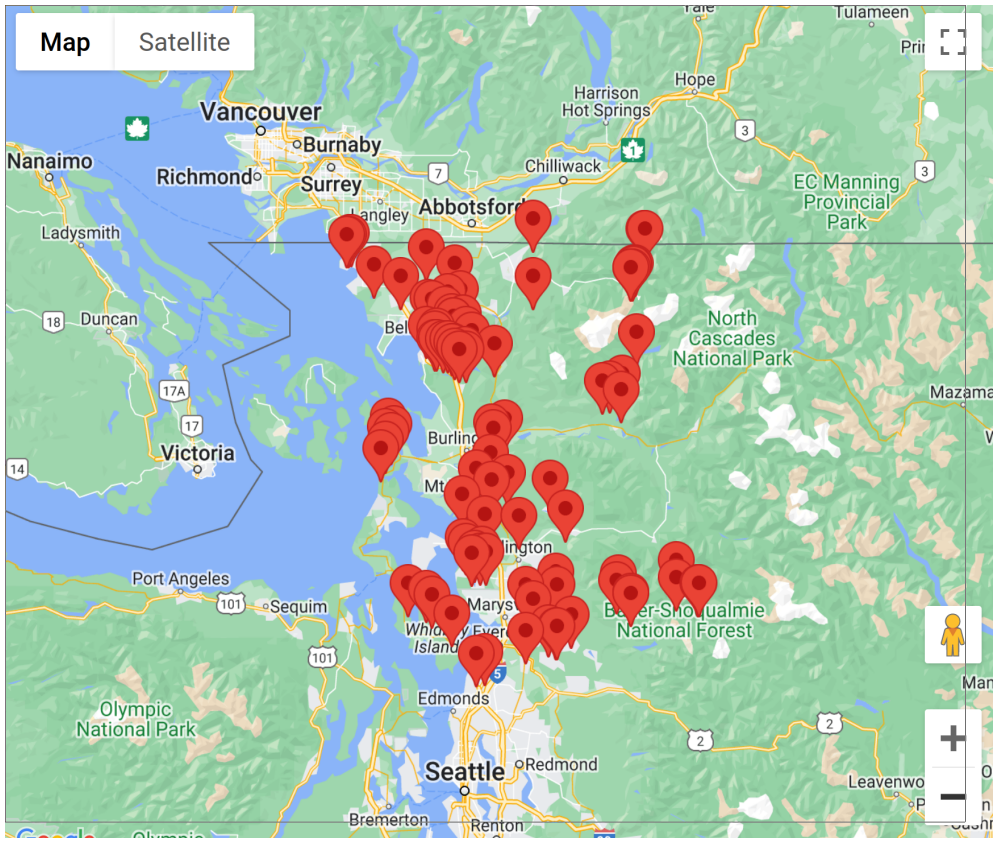






Correlation Analysis

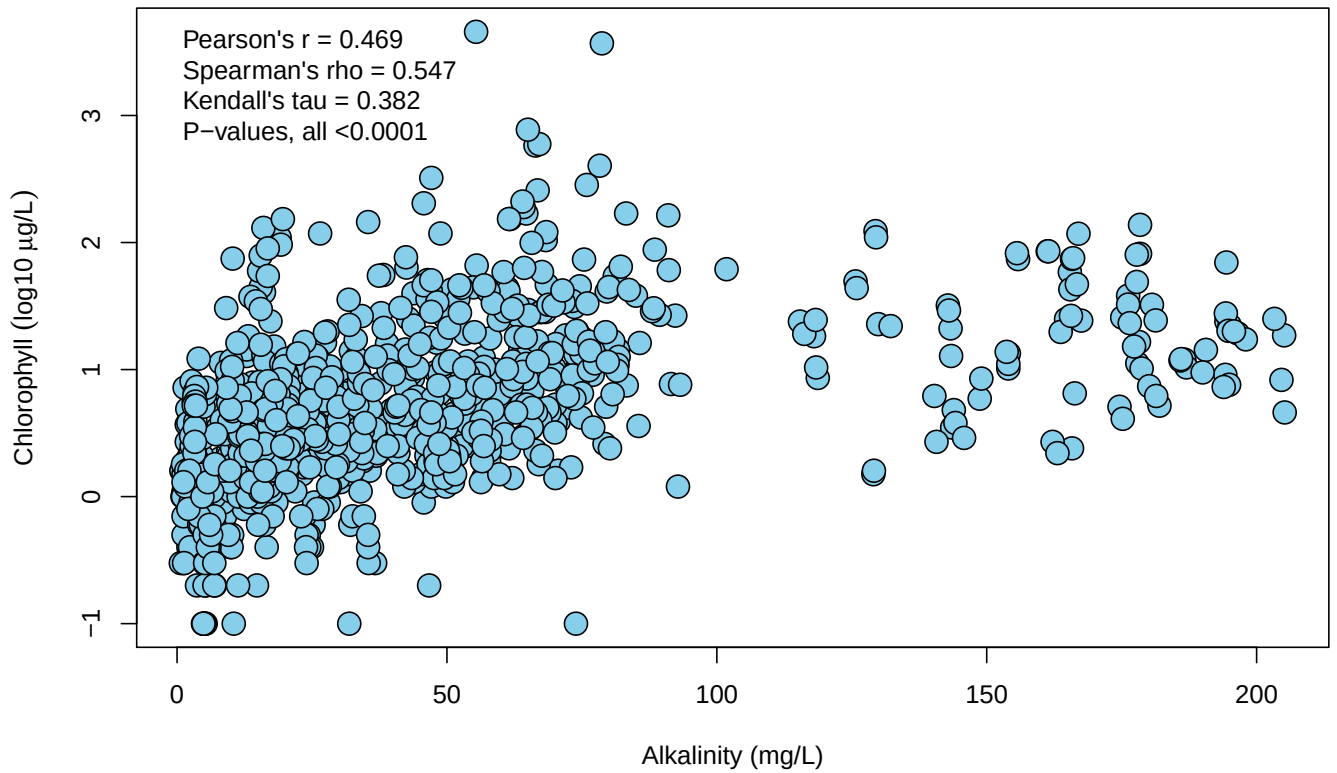
Example #6 - Can Alkalinity Be Used To Predict Chlorophyll?



Western Washington University analyzes water samples from 60–70 lakes each the summer. One of the most time consuming analyses is chlorophyll, which is an important indicator of water quality

Correlation Analysis

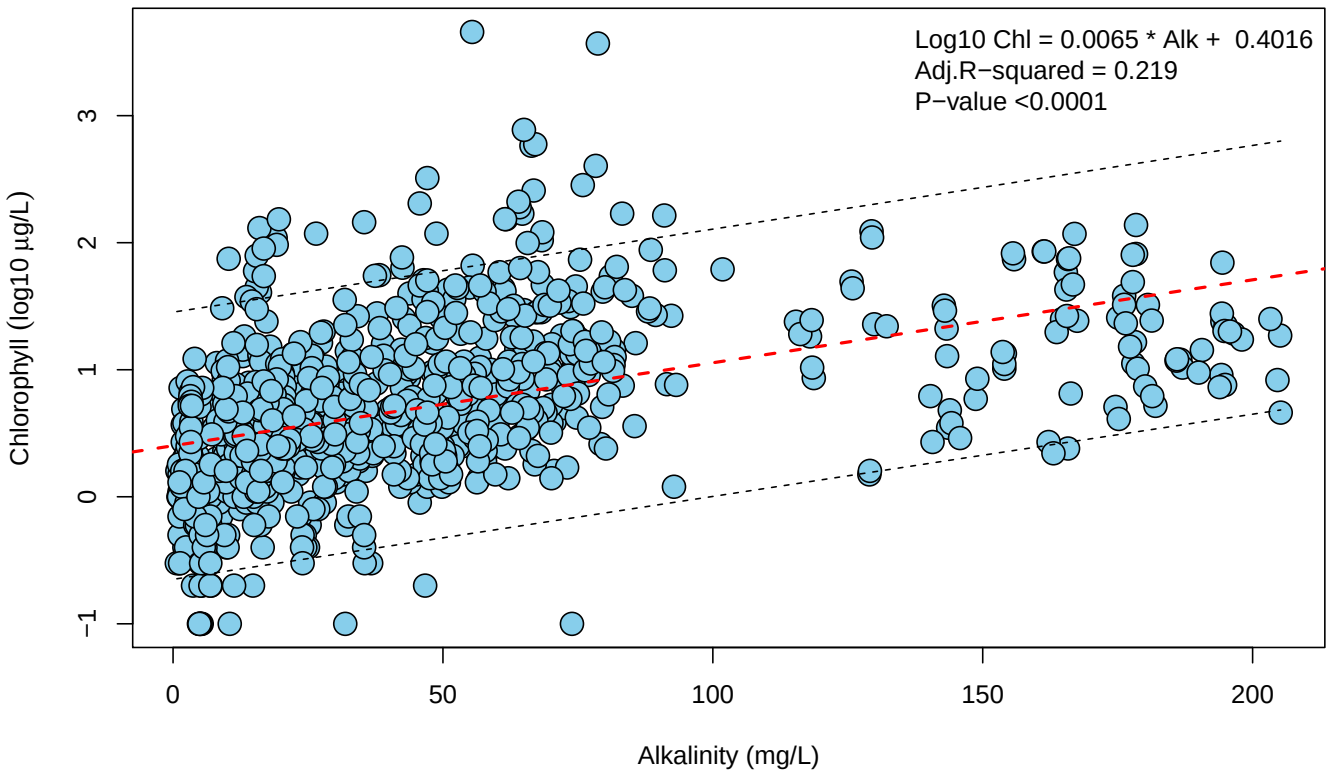
Example #6, continued



Simple correlation analysis shows that there is a monotonic relationship between alkalinity and chlorophyll, which is to be expected from what scientists know about algal photosynthesis. Can we use alkalinity, which is easier to measure, to estimate chlorophyll in lakes?

Regression Analysis

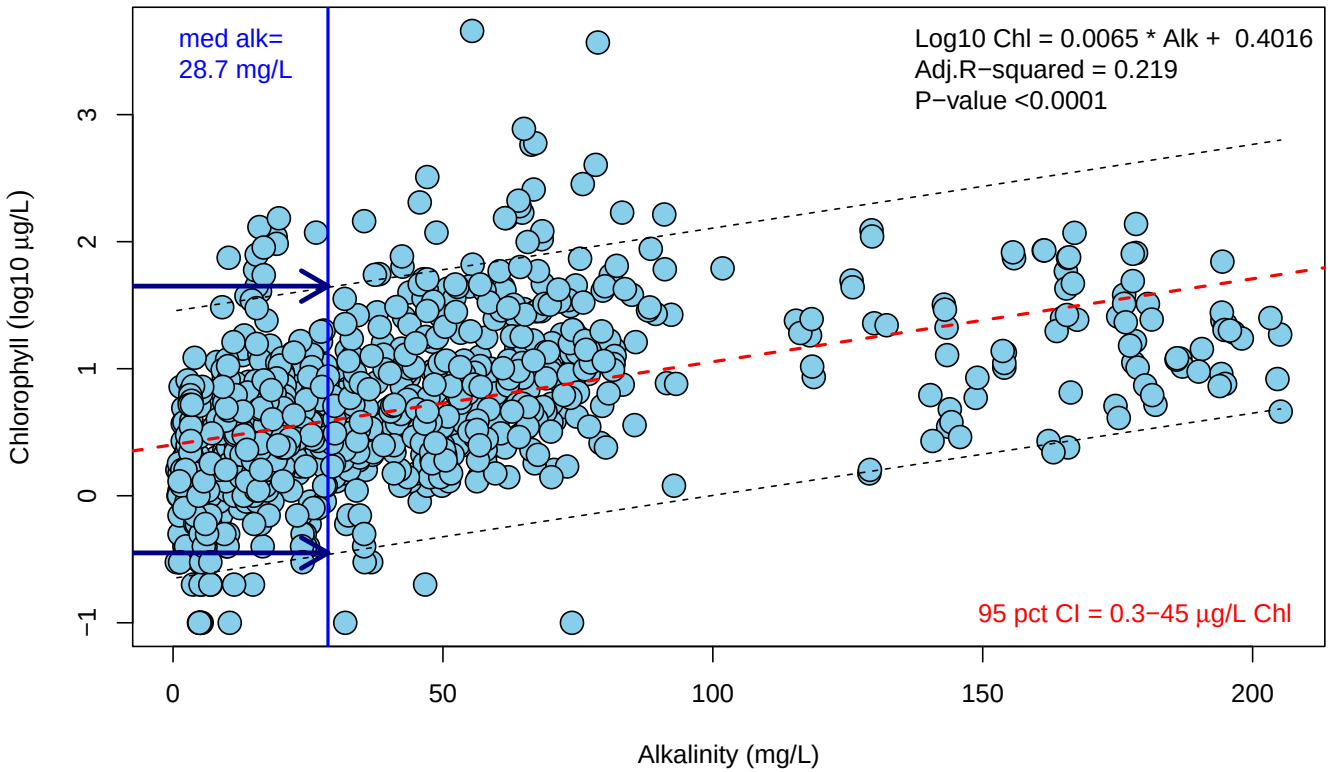
Example #6, continued



Regression analysis confirms that you can create a significant linear model

Regression Analysis

Example #6, continued



But the confidence intervals are huge (log scale). For example, at the median alkalinity concentration (28.7 mg/L), the 95% CI for chlorophyll is 0.3–45 μg/L (back transformed from the log scale); this is too wide to determine lake water quality

Introduction to Multivariate Analysis

Initial Examination of Multivariate Data

- Begin by plotting and using simple exploratory tools like correlation analysis to look for patterns

  The primary purpose of multivariate analysis is data simplification . . . don't use complicated multivariate tests to describe simple univariate or bivariate patterns

- Check for normality and homoscedasticity . . . nearly all multivariate methods are parametric

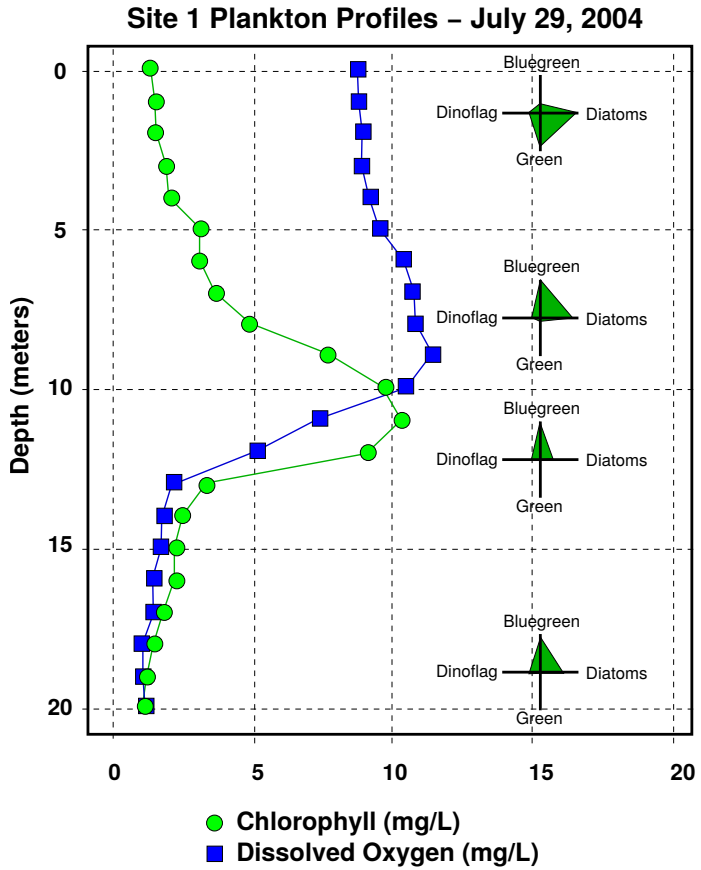
  Heteroscedastic variances are a problem for most multivariate tests

- Identify redundant, nonlinear, and random variables . . . exclude those variables if appropriate

  Including redundant, nonlinear, and random variables in multivariate analysis can obscure patterns in the remaining variables

Simple Multivariate Plotting

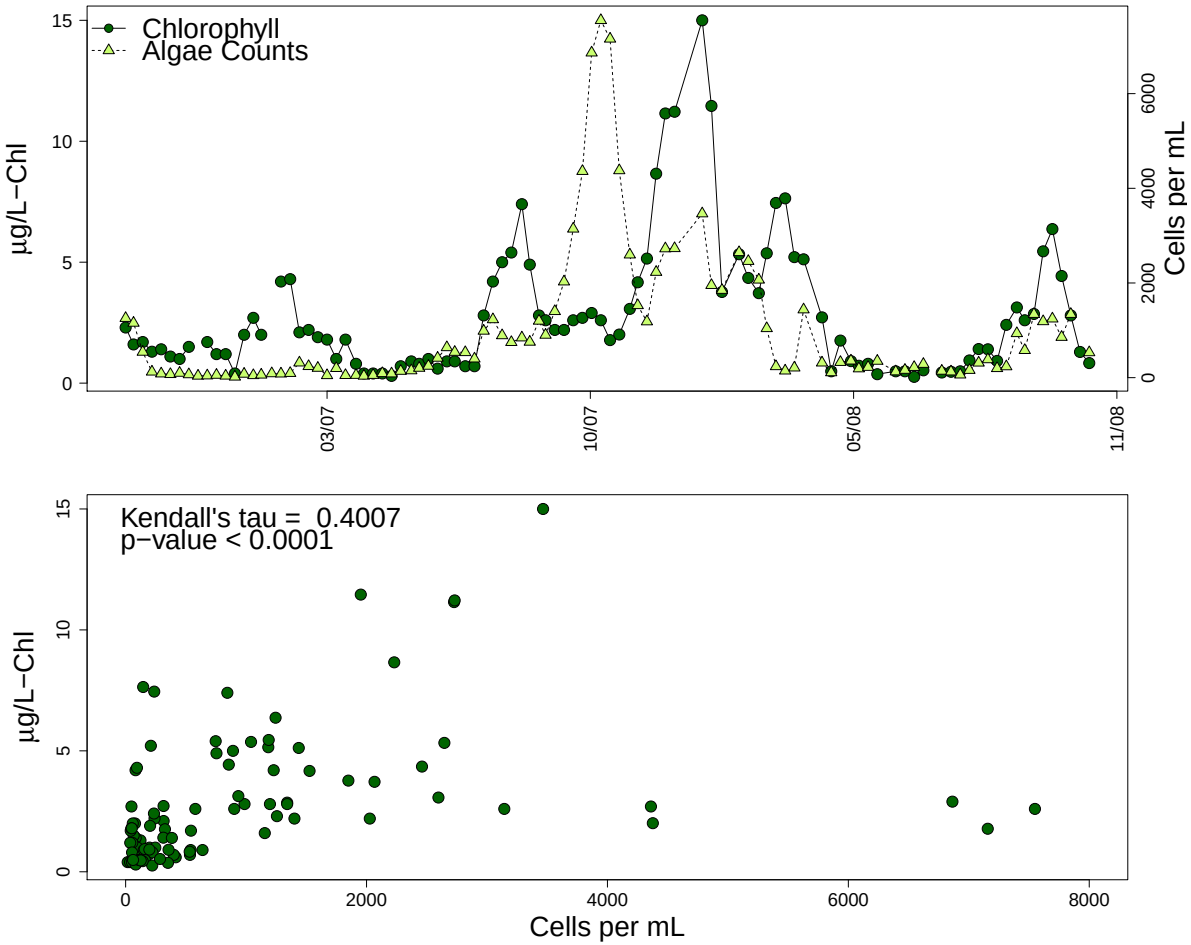
Example #7 - Lake Whatcom Oxygen, Chlorophyll, and Algae Patterns



This example shows a scatterplot of oxygen and chlorophyll concentrations, plotted by depth, with kite diagrams showing the relative abundance of different types of algae

Simple Multivariate Plotting

Example #8 - July Reservoir Algae and Chlorophyll



The upper figure is a dual axis plot showing total algae cell counts (cells per mL) and chlorophyll concentrations (μg/L-Chl) by date. The lower plot shows the correlation between cells counts and chlorophyll (using rank-based Kendall's τ)

Multivariate Analysis

Clustering and Ordination

Two common multivariate patterns include similarity among groups of samples (clustering) and increasing dissimilarity along a gradient (ordination)

Clustering involves finding similarity among groups of samples:

A	B	C	D	E	F	E	C	A	D	F	B
2	3	1	2	1	3	1	1	2	2	3	3
2	3	1	2	1	3	1	1	2	2	3	3
3	1	2	3	2	1	2	2	3	3	1	1
3	1	2	3	2	1	2	2	3	3	1	1
1	2	3	1	3	2	3	3	1	1	2	2
1	2	3	1	3	2	3	3	1	1	2	2

Ordination looks for increasing dissimilarity along a gradient:

A	B	C	D	E	F	E	C	A	D	F	B
3	6	2	4	1	5	1	2	3	4	5	6
2	5	1	3	6	4	6	1	2	3	4	5
1	4	6	2	5	3	5	6	1	2	3	4
6	3	5	1	4	2	4	5	6	1	2	3
5	2	4	6	3	1	3	4	5	6	1	2
4	1	3	5	2	6	2	3	4	5	6	1

Multivariate Analysis

Hierarchical vs. Divisive Clustering

- Most commonly used technique is agglomerative, hierarchical clustering
  - Each sample (row) containing a set of measurements is called a *point*
  - Similarity (distance) is calculated between all points (samples)
  - The closest (most similar) two points are combined into a single joined point
  - The distances between all remaining points and the new joined point are recalculated
  - The next closest two points are joined, and so on
- Another common technique is divisive clustering, where initial groups are defined, then all clusters are iteratively regrouped until the distances are minimized

Cluster Analysis

Can You Test Significance of Cluster Results?

- Clustering is a "blind" exploratory statistic that looks for groups in multivariate data. The output does not provide p-values to confirm the significance of the clusters

- You can add an Association Analysis to test whether the cluster groups are significantly associated with specific groups in the data (like `Species` in the iris data)

Cluster Analysis

Example #9 - Fisher's (Anderson's) Iris Data

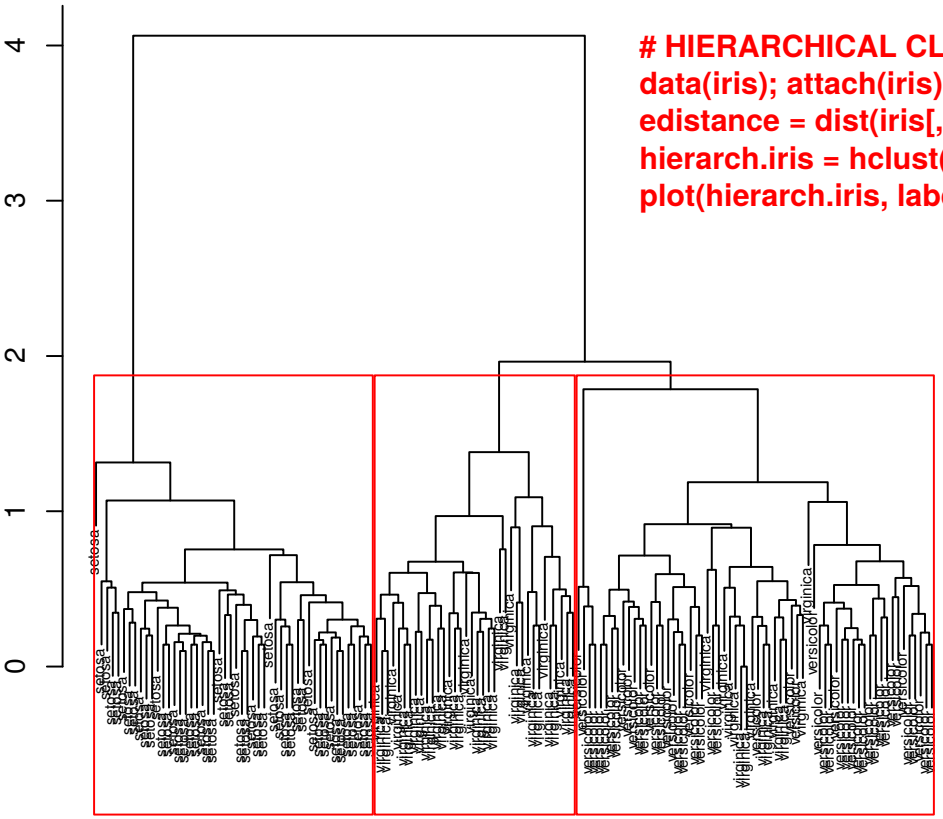


Photographs by C. Hensler and D. Kramb; downloaded with permission from http://www.signa.org

- Fisher's Iris data are widely used to illustrate multivariate patterns and evaluate new statistical techniques
- The data consist of sepal and petal width and length measurements from 150 iris flowers representing three species of iris
- The iris data are included with the `R` "base library" and can therefore be loaded using the statement `data(iris)`

Hierarchical Clustering and Association Analysis

Testing Association Between Hierarchical Clusters and Iris Species




```
# HIERARCHICAL CLUSTERING – ALL IRIS DATA
data(iris); attach(iris)
edistance = dist(iris[, c(1:4)], method="euclidean")
hierarch.iris = hclust(edistance, method="average")
plot(hierarch.iris, labels=iris[, 5], cex=0.5, ann=F)
```

```
### ADD RECTANGLES:
rect.hclust(hierarch.iris, 3)
```

Hierarchical Clustering and Association Analysis

Testing Association Between Hierarchical Clusters and Iris Species

```
irisgroup = cutree(hierarch.iris, 3) # form 3 groups from cluster results
table(irisgroup, iris$Species)       # show cluster groups vs species
```

```
irisgroup setosa versicolor virginica
        1     50          0         0
        2      0         50        14
        3      0          0        36
```

```
chisq.test(irisgroup, iris$Species)
```

```
        Pearson's Chi-squared test

data:  irispart$Species and irisgroup
X-squared = 234.375, df = 4, p-value < 2.2e-16
```

This combination of `table` and `chisq.test` is often called *Association Analysis*. We are showing the association between the hierarchical clusters and Iris Species using a contingency table (`table`) and testing the significance of the association using `chisq.test`

Hierarchical clustering correctly grouped all *setosa* and *versicolor* samples, but misclassified 14 of the *virginica* samples (39%). The total misclassification rate was 9.3% (14 out of 150)

Kmeans Clustering and Association Analysis

Testing Association Between Kmeans Clusters and Iris Species

```
kmeans.iris = kmeans(iris[, c(1:4)], 3)
table(iris$Species, kmeans.iris$cluster)
```

```
            1  2  3
 setosa     50  0  0
 versicolor  0 48  2
 virginica   0 14 36
```

```
chisq.test(iris$Species, kmeans.iris$cluster)
```

```
        Pearson's Chi-squared test

 data:  iris$Species and kmeans.iris$cluster
 X-squared = 223.5993, df = 4, p-value < 2.2e-16
```

Kmeans clustering correctly clustered *setosa*, but placed 2 of the *versicolor* and 14 of the *virginica* into the wrong groups (10.7% total misclassification).

Because kmeans clustering has an element of uncertainty, repeating the clustering process can give you different results. For example, one of my cluster "runs" misclassified 17 *setosa* and 4 *versicolor* (14% misclassification)

Kmeans Clustering and Association Analysis

Plotting the Association Analysis Results

One of the features of kmeans clustering is that it provides information about which variables contribute to cluster separation. To access that information, type the cluster object (kmeans.iris) and look at Cluster means:

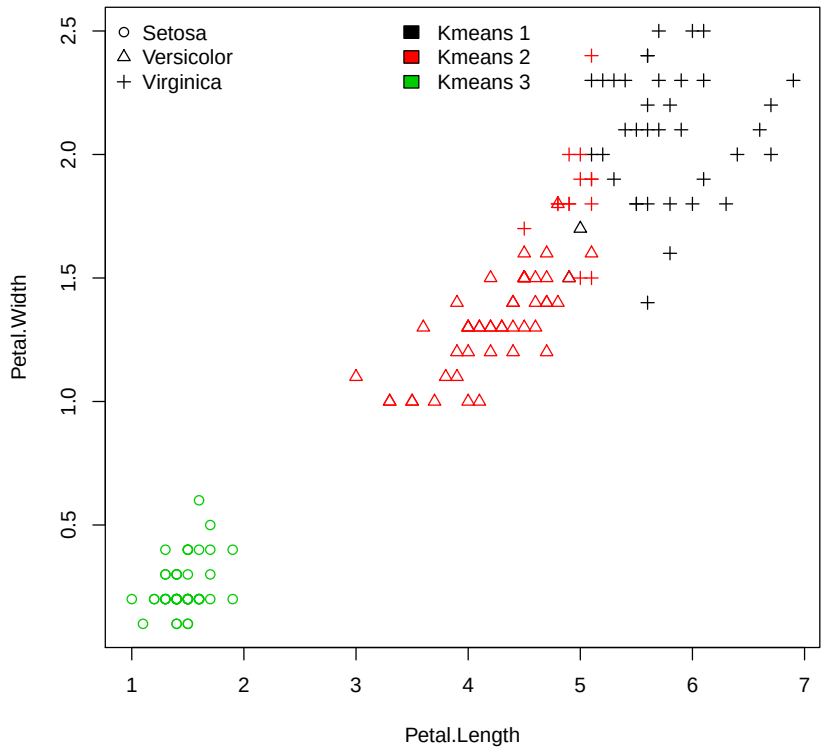
```
kmeans.iris  # review the kmeans cluster features
```

```
K-means clustering with 3 clusters of sizes 50, 62, 38
Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.006000    3.428000     1.462000    0.246000
2     5.901613    2.748387     4.393548    1.433871
3     6.850000    3.073684     5.742105    2.071053
```

The Cluster means are the centers for each variable by group. So, for example, the cluster center for petal width in group #1 is 0.246. Group #1 contains all of the *Iris setosa* samples, which have short, narrow petals. By examining the cluster centers you should be able to see that petal lengths and widths are good plotting choices for displaying the results!

Kmeans Clustering and Association Analysis

Plotting the Association Analysis Results




```
plot(Petal.Length, Petal.Width, col=kmeans.iris$cluster, pch=unclass(iris$Species))
legend(x="topleft", c("Setosa", "Versicolor", "Virginica"), pch=c(1,2,3), bty="n")
legend(x="top", c("Kmeans 1", "Kmeans 2", "Kmeans 3"), fill=c(1,2,3), bty="n")
```

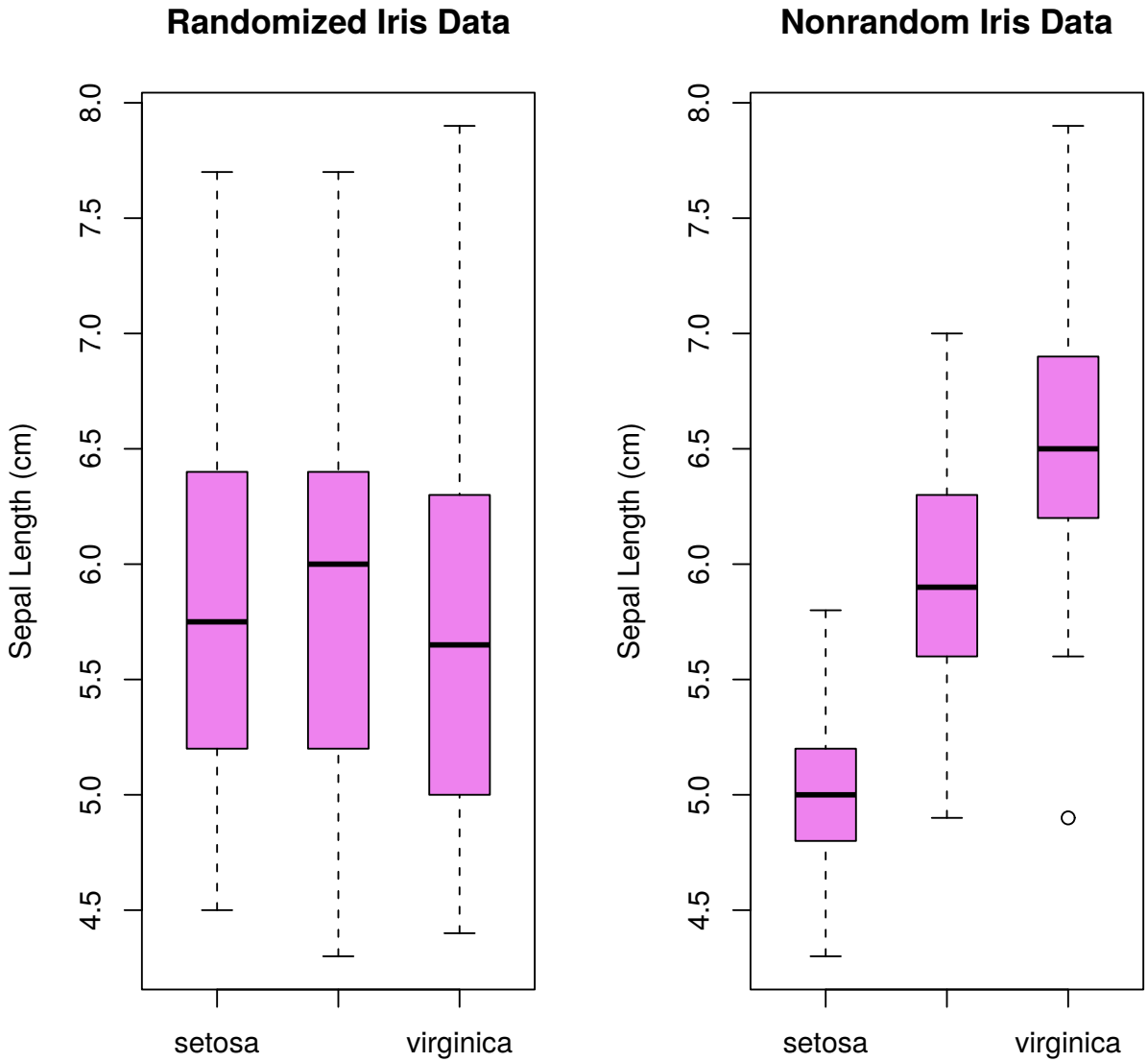
Randomization Testing

Time for a Reality Check!

- Randomization testing is an an excellent reality check when working with multivariate data
- Rather than relying on assumptions about distributions or tables of probabilities, you test whether the results for your nonrandom data are likely to occur by chance
- To create a randomized sample, each variable was sampled in random order, then written back into the same column. As a result, sepal lengths were still in column 1, but the values no longer lined up with the correct iris species

Randomization Testing

Boxplots Showing the Effect of Randomization on Iris Data

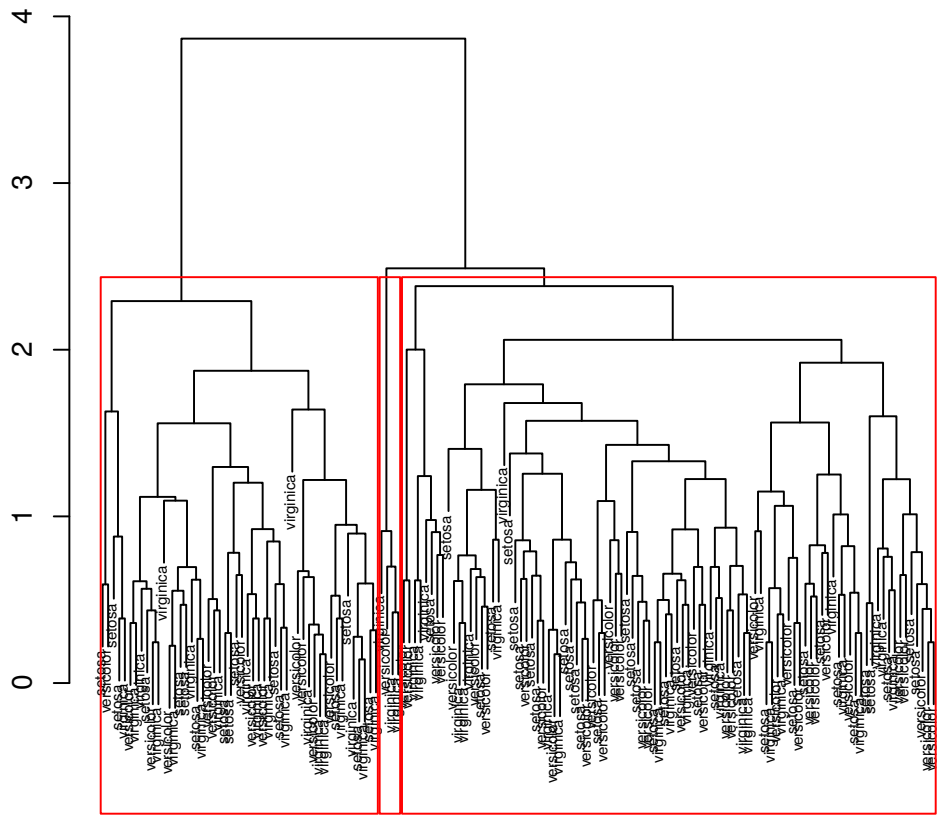


Randomization Testing

Hierarchical Clustering of Randomized Iris Data

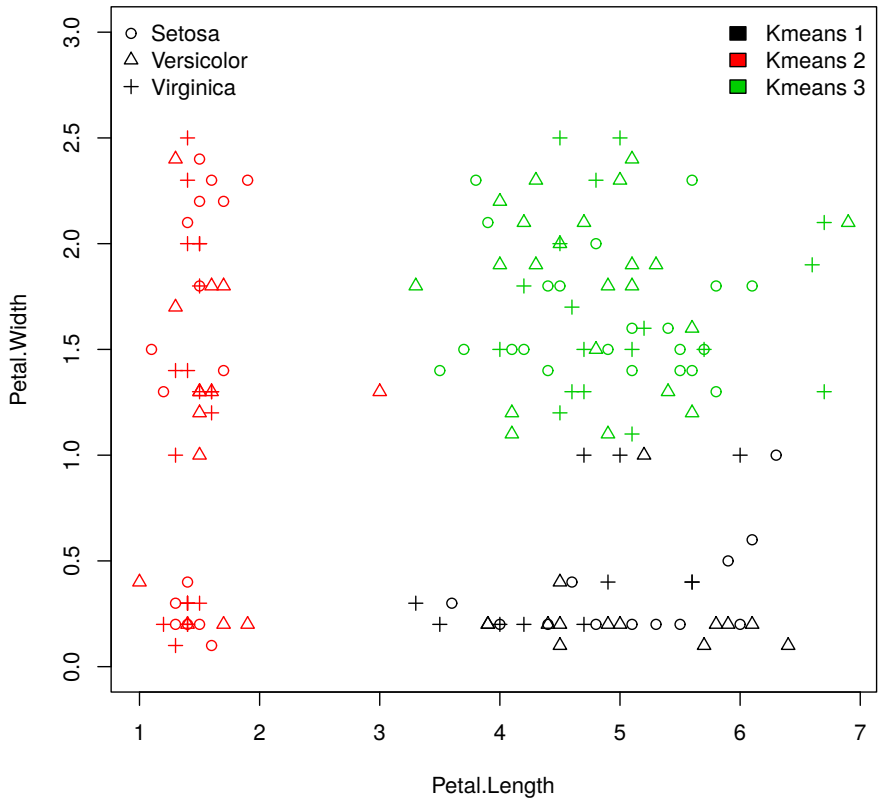
```
table(ririsgroup, riris$Species)

ririsgroup setosa versicolor virginica
         1     34         36        26
         2     16         13        21
         3      0          1         3
```

```
chisq.test(riris$Species, ririsgroup)

Pearson's Chi-squared test

data:  riris$Species and ririsgroup
X-squared = 7.21, df = 4, p-value = 0.1252
```



Randomization Testing

Kmeans Clustering of Randomized Iris Data




```
###EDITED ASSOCIATION ANALYSIS OUTPUT:
             1  2  3
  setosa     12 16 22
  versicolor 14 14 22
  virginica  11 21 18

X-squared = 2.4239, df = 4, p-value = 0.6583
```

Randomization Testing

Kmeans Clustering and Association Analysis of Randomized Iris Data

Trial	Randomized Data		Nonrandomized Data	
	X-squared	p-value	X-squared	p-value
1	2.42	0.6583	246	$<$ 2.2e-16
2	0.36	0.9855	246	$<$ 2.2e-16
3	3.11	0.5387	255.84	$<$ 2.2e-16
4	11.45	0.0219	219.36	$<$ 2.2e-16
5	4.94	0.2935	219.36	$<$ 2.2e-16

You can increase your confidence in the clustering results by comparing repeating the process using different randomized data sets. Here is a comparison using 5 randomized data sets; the `Welch Two Sample t.test` was used to determine whether the X-squared statistics are significantly different

```
NR = c(246,246,256,219,219); R = c(2.42,0.36,3.11,11.45,4.94)
t.test(R, NR, var.equal=F)
```

```
t = -28.527, df = 4.489, p-value = 2.578e-06
```

Multivariate Analysis

Principal Components Analysis

- There are two basic PCA methods: `princomp` and `prcomp`
  - `princomp` ordinates using an eigenvalue matrix
  - `prcomp` is based on a singular value decomposition of the data matrix, which is generally preferred over `princomp`
  - `princomp` and `prcomp` will often produce identical results (number of principal components = number of variables)
  - But if there are a large number of variables, `prcomp` truncates after "almost all" of the variance is contained in the ordination (number of principal components $\leq$ number of variables)
- Both default to a covariance matrix (matches S-Plus), but the best option is a scaled, centered correlation matrix
- In both methods, omit variables that are constant (e.g., all zeros)

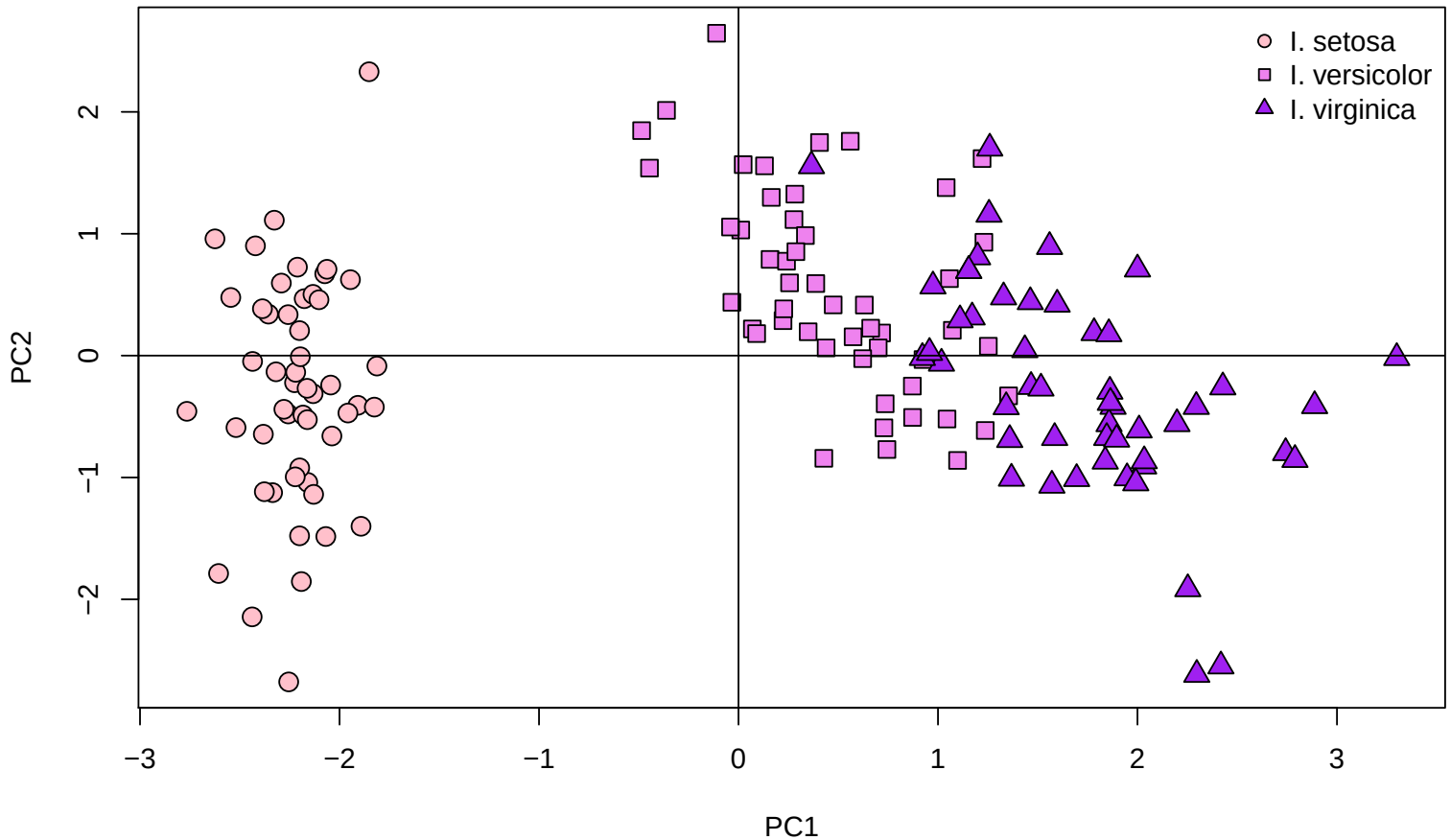
Principal Components Analysis

Example #10 - Fisher's (Anderson's) Iris Data

```
##### PRINCOMP VERSION WITH SCALED/CENTERED CORRELATION MATRIX
data(iris); attach(iris)
iris.princomp <- princomp(iris[, c(1:4)], cor=T) #Basic PCA command
summary(iris.princomp)

Importance of components:
                          Comp.1    Comp.2     Comp.3      Comp.4
Standard deviation     1.7083611 0.9560494 0.38308860 0.143926497
Proportion of Variance 0.7296245 0.2285076 0.03668922 0.005178709
Cumulative Proportion  0.7296245 0.9581321 0.99482129 1.000000000


##### PRCOMP VERSION WITH SCALED/CENTERED CORRELATION MATRIX
iris.prcomp <- prcomp(iris[, c(1:4)], scale=T, center=T)

summary(iris.prcomp)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

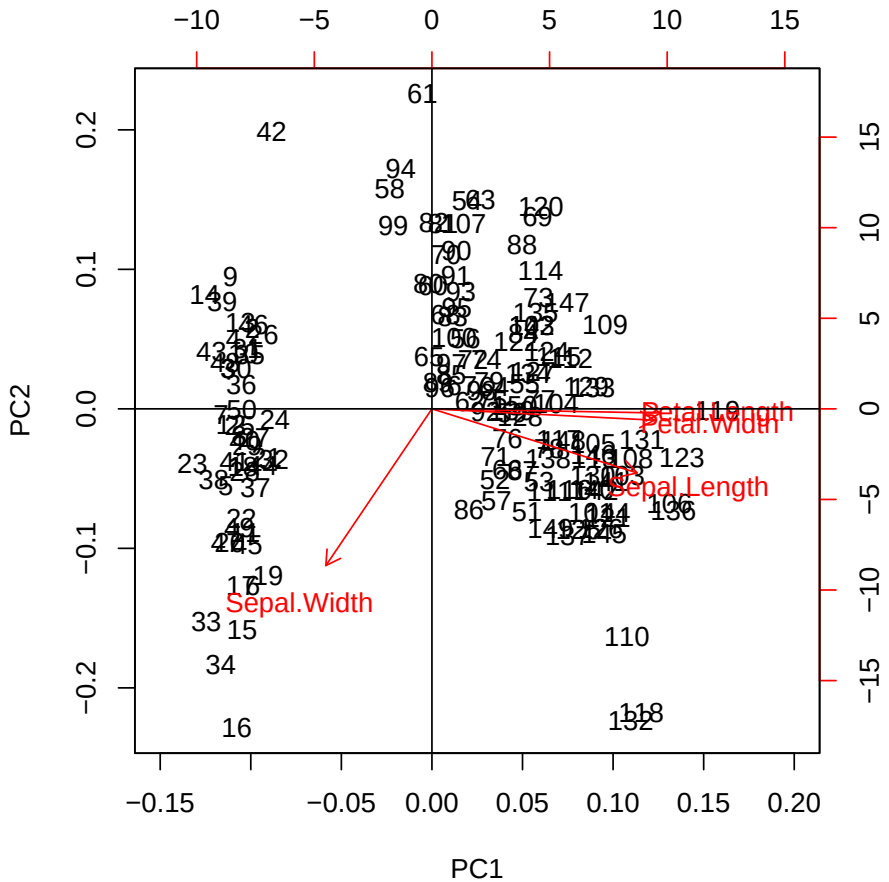
Principal Components Analysis

Principal Components Ordination of Iris Samples



Principal Components Analysis

Biplot Showing Sample and Variable Ordination



One PCA goal is variable reduction. Note how petal length/width and sepal length ordinate together on PC1. We could separate the *Iris setosa* samples using only one variable

Principal Components Analysis

Variable Reduction Features

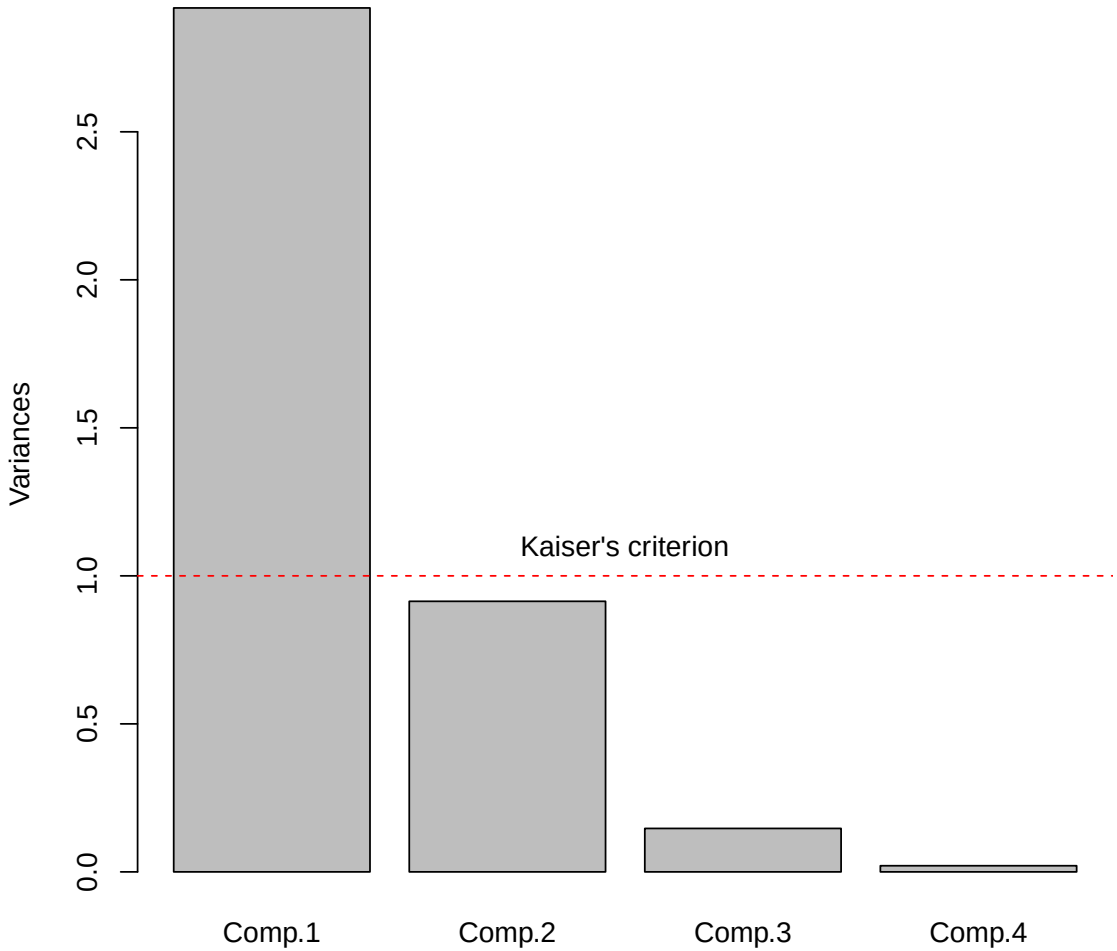
- Another PCA goal is to find the smallest number of components needed to ordinate the samples
- The simplest approach is to look at the variance plot and see where it falls off sharply
- Another method is to use Kaiser's criterion (default in SPSS) - select components with variance ≥ 1[†]

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.7083611	0.9560494	0.38308860	0.143926497
Variance (sd^2)	2.92	0.91	0.15	0.02

[†] Kaiser, H. F. 1960. *The application of electronic computers to factor analysis.* Ed. and Psychol. Measurement, 20: 141-151

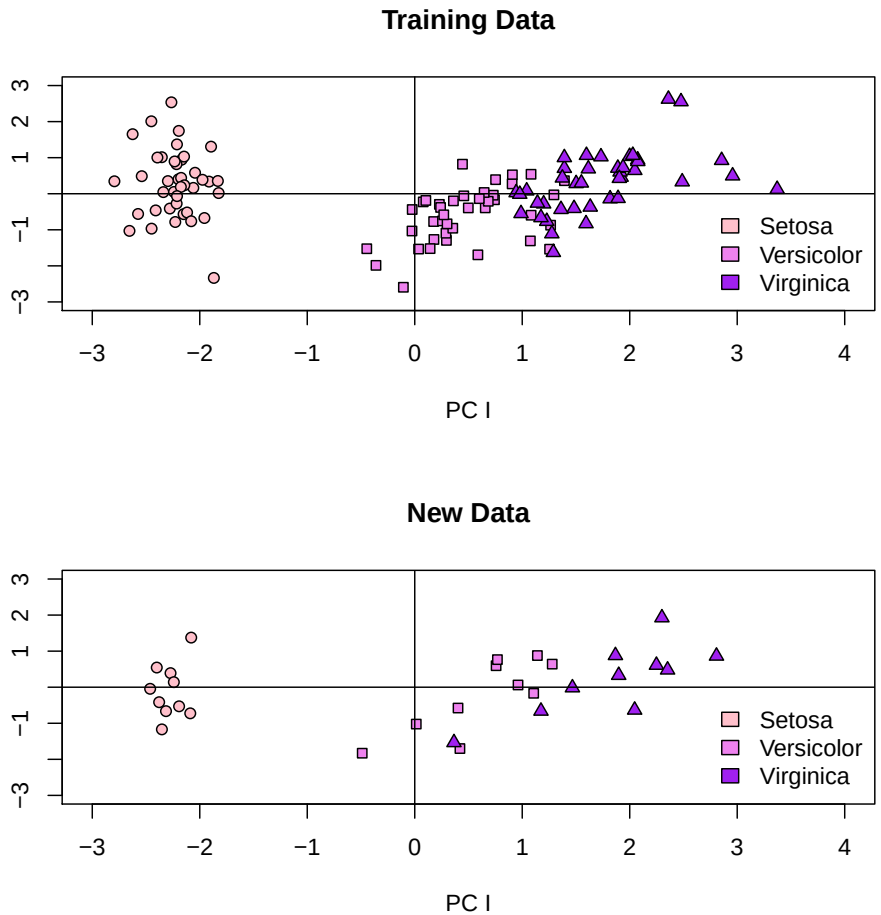
Principal Components Analysis

Variance Plot - Note Importance of PC1



Principal Components Analysis

Using a Trained PCA to Ordinate New Data



A third goal is to use a training set to ordinate new data. This example builds the model using all but the first 10 rows for each iris species, then uses the PCA results to ordinate the "new" data

Multivariate Analysis

Clustering on Principal Components

- Many multivariate methods default to using all variables in a data set
  - In an earlier slide I mentioned that a common data set problem is that not all the measured parameters will show an effect
  - In data sets with only a few variables, you can make reasonable decisions about which variables to omit, but this is not feasible for large data sets and can lead to questions about your selection criteria
- Alternatively, we can use the variable and component reduction features of PCA to improve clustering results without making (potentially) arbitrary selection choices

Clustering on Principal Components

Example #11, Effects of Triclosan on Sediment Biota

- This example is based on data published by Chariton, et al. (2014), following the approach described by Ben-Hur and Guyon (2003).
- The data are from a sediment toxicity test to determine the effects of triclosan, a commonly used antibiotic/antifungal compound, on sediment biota.
- The biota were identified using molecular markers that identified presence or absence of >850 unique sediment organisms, listed by *operational taxonomic units* (OTUs) rather than genus and species. The sediment samples contained zero, low, or high concentrations of the toxicant (triclosan),
- Using OTUs represents an alternative to traditional taxonomic assessment because the sediment organisms don't have to be sorted, identified, and enumerated to measure biological diversity (tedious process!)

Clustering on Principal Components

Preliminary Data Decisions

- The original data file contained presence/absence results for 858 OTUs for three treatments (control, low, high) with six replicates per treatment. As a result, the entire file had only 18 rows.
- Nine of the OTUs had identical values for all 18 samples (variance = zero). These measurements were excluded from the analysis, leaving a data set containing 18 samples and 849 variables
- The data were analyzed using PCA based on a singular value decomposition of scaled, centered data matrix (`prcomp`).
- Using this type of PCA has several advantages over eigenvector-based PCA, including the ability truncate components if the remaining variance approaches zero. In this case, the PCA truncated at 18 components (residual variance <3.3e-15) rather than forcing the use of all 849 components

Clustering on Principal Components

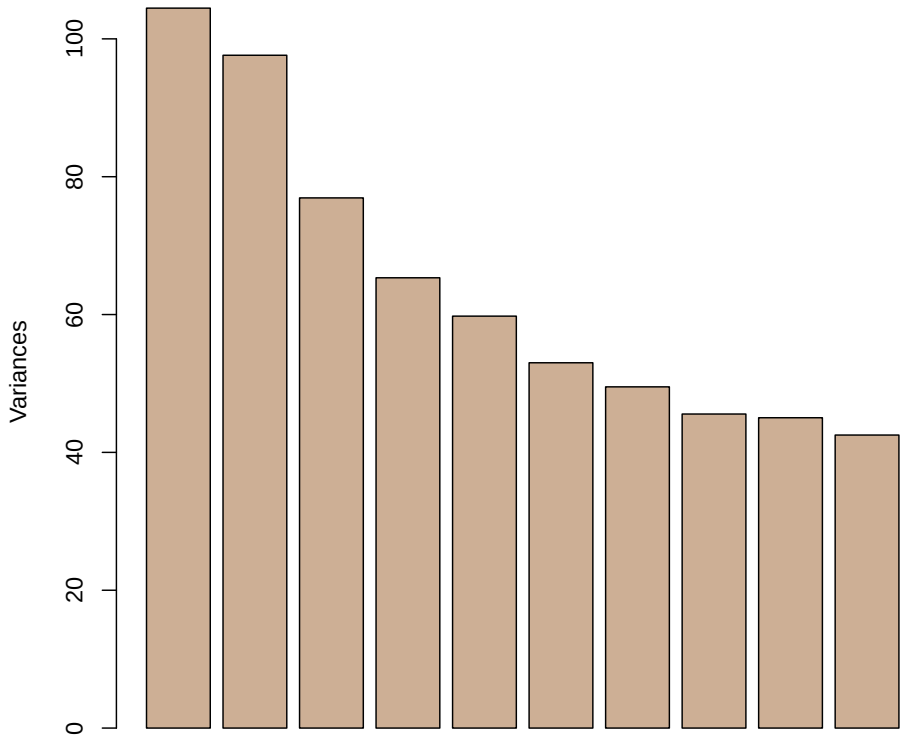
Step 1: Creating New Variables from Component Scores

The prcomp program was used on the entire data set (columns 4:852; excluding group, treatment, replicate). The summary function shows that virtually all variability can be explained using the first 18 principal components (out of 849)

```
Importance of components:
                          PC1   PC2     PC3     PC4    PC5     PC6     PC7
Standard deviation     10.221 9.880 8.77070 8.08297 7.7312 7.28034 7.03630
Proportion of Variance  0.123 0.115 0.09061 0.07695 0.0704 0.06243 0.05832
Cumulative Proportion   0.123 0.238 0.32863 0.40559 0.4760 0.53842 0.59673

                           PC8     PC9    PC10   PC11    PC12    PC13    PC14
Standard deviation     6.75020 6.71040 6.52041 6.3837 6.27286 5.86208 5.52219
Proportion of Variance 0.05367 0.05304 0.05008 0.0480 0.04635 0.04048 0.03592
Cumulative Proportion  0.65040 0.70344 0.75352 0.8015 0.84786 0.88834 0.92426

                          PC15    PC16    PC17      PC18
Standard deviation     5.05834 4.56602 4.22722 3.321e-15
Proportion of Variance 0.03014 0.02456 0.02105 0.000e+00
Cumulative Proportion  0.95440 0.97895 1.00000 1.000e+00
```

Clustering on Principal Components

Variance Plot for First 10 Components



The variance plot is more gradual than most of our previous examples, but note that all 18 components exceed Kaiser's criterion

Clustering on Principal Components

Comparison of Original and New Data Sets

Original Sediment Microcosm Data (18 rows, 851 columns)

Treatment	Replicate	OTU 1	...	OTU 849
control	1	0 or 1		0 or 1
control	2	0 or 1		0 or 1
(etc.)	(etc.)	(etc.)		(etc.)
high	5	0 or 1		0 or 1
high	6	0 or 1		0 or 1

Sediment Microcosm PCA Data (18 rows, 20 columns)

Treatment	Replicate	PC 1	...	PC 18
control	1	-4.97		$< \pm 0.01$
control	2	-15.53		$< \pm 0.01$
(etc.)	(etc.)	(etc.)		(etc.)
high	5	13.81		$< \pm 0.01$
high	6	13.10		$< \pm 0.01$

Clustering on Principal Components

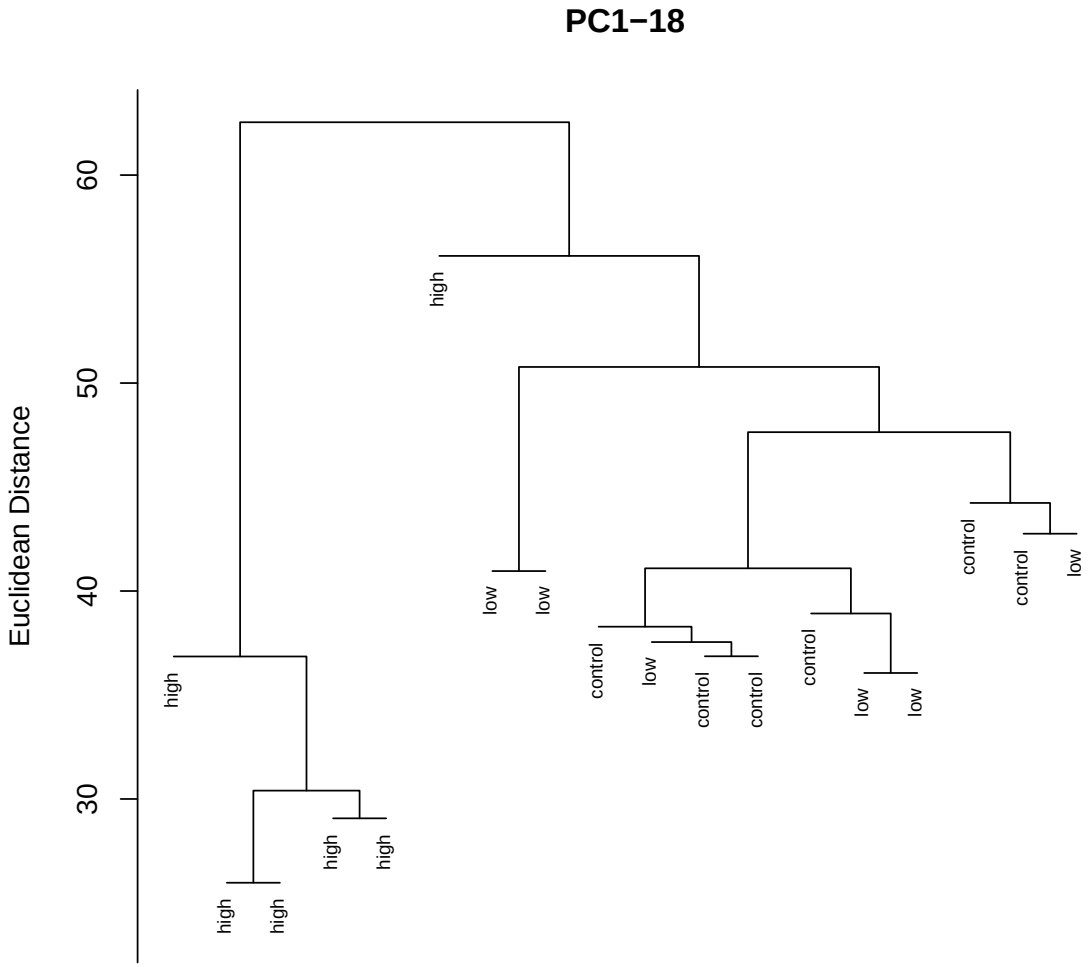
Step 2: Clustering on the Component Scores

- This next step requires a good understanding of what is accomplished by ordinating the original data using principal components.
- A scaled, centered PCA creates a multivariate correlation matrix, with the "best" correlations contained in the first component. Each successive component contains a smaller fraction of "good" correlation.
- What we want to do is cluster using a small subset of the component scores rather than all the components scores. This allows us to focus on just the good multivariate correlations.
- In this example, we had 849 variables (OTUs), but `prcomp` stopped the ordination at 18 components, which contain nearly 100% of the variance.
- The next figure shows the dendrogram for clustering the samples using all 18 principal components. I used euclidean distance and Ward's minimum variance clustering method. (`ward.D` used to duplicate results in Chariton et al. 2014)

  Note: We don't want to use all components, but it is a good place to start.

Clustering on Principal Components

Dendrogram Results using 18 Components



Clustering on Principal Components

Step 3: Identifying Stable Clusters

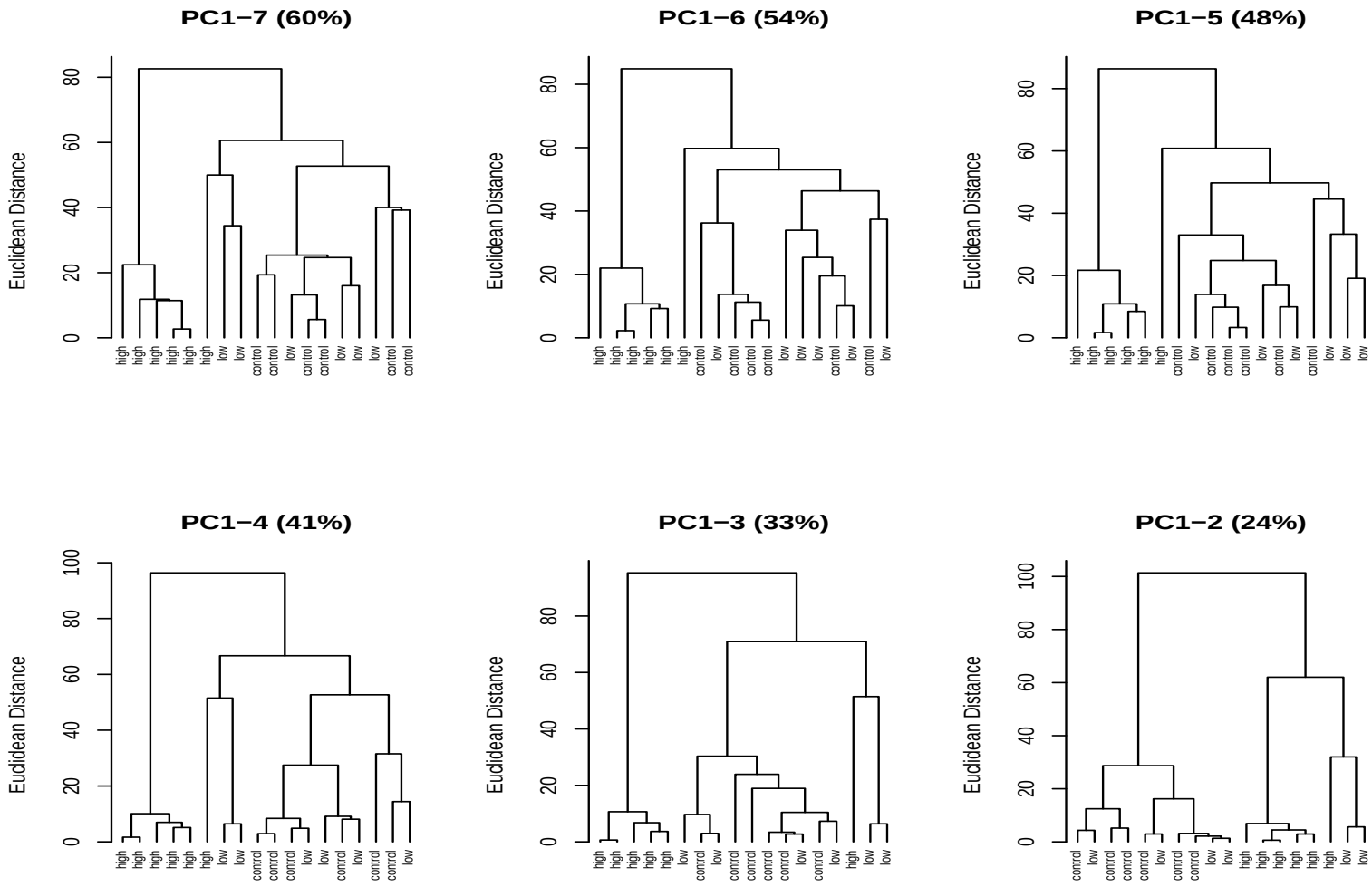
- We want to cluster using fewer than 18 component scores. But how many should we select?
- This is actually a rather difficult question, but the short answer is to use the fewest principal components that produce "stable" clusters (see Ben-Hur and Guyon, 2003).
- Preliminary evaluation of the 18-component dendrogram shows that there are only two *treatment* responses (high or control+low), and one outlier ("high").
- Using Association Analysis, we can test whether the separation into two groups is statistically significant (FYI: "eward" is the name of my hierarchical clustering object)

```
HCgroups = cutree(eward, 2)
table(HCgroups, treatment)
```

```
         treatment
HCgroups control high low
       1       6    1   6
       2       0    5   0
```

```
chisq.test(HCgroups, treatment) #edited output
X-squared = 13.8462, df = 2, p-value = 0.0009848
```

Dendrogram Results using PC1-PC7



Cycling through all clustering options (PC1–PC18, PC1–PC17, PC1–PC16, etc), each have 1 misclassification until you use only PC1–PC2, which has 2 misclassifications. *You only need PC1–PC3 to produce stable clusters*

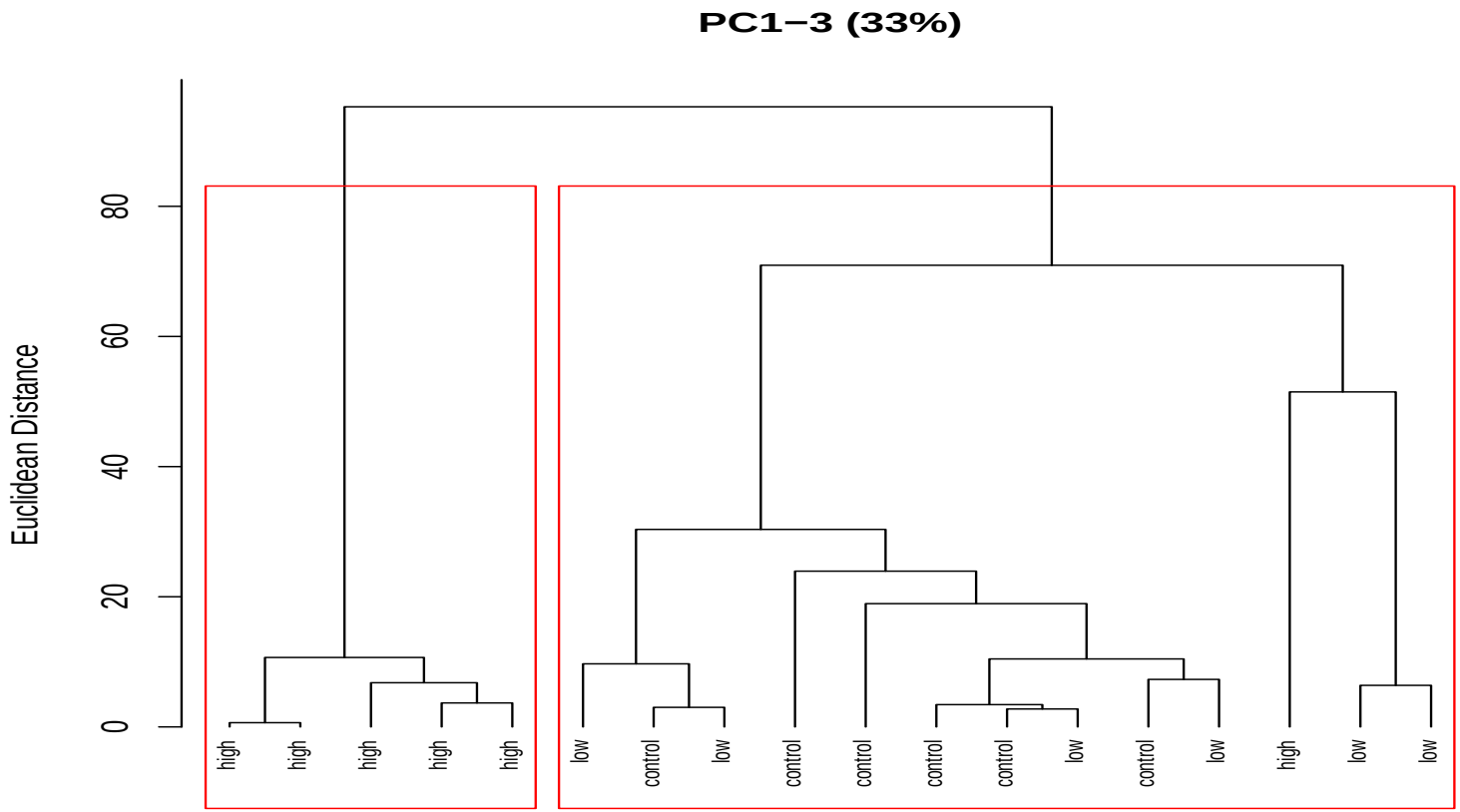
Clustering on Principal Components

Step 4: Refining Cluster Membership and Checking Significance

- Once we decide on the number of principal components to use (PC1–PC3), it is important to revisit the question of cluster membership
- You can see from the next figure that there are two obvious clusters that match treatment groups. One cluster contains five "high" treatment samples and the other cluster contains all of the control and low samples, plus one high outlier
- But in the next figure, note that you could also describe the data using three clusters. Two of the clusters show treatment effects (high or control/low); the third cluster contains three outlier samples.
- In both cases, the associations between clusters and treatment groups are statistically significant (p-value $\leq$0.001).
- Choosing which way to display the results depends on your overall goals, but it is usually desirable to discuss outliers separately from the treatment effect.

Clustering on Principal Components

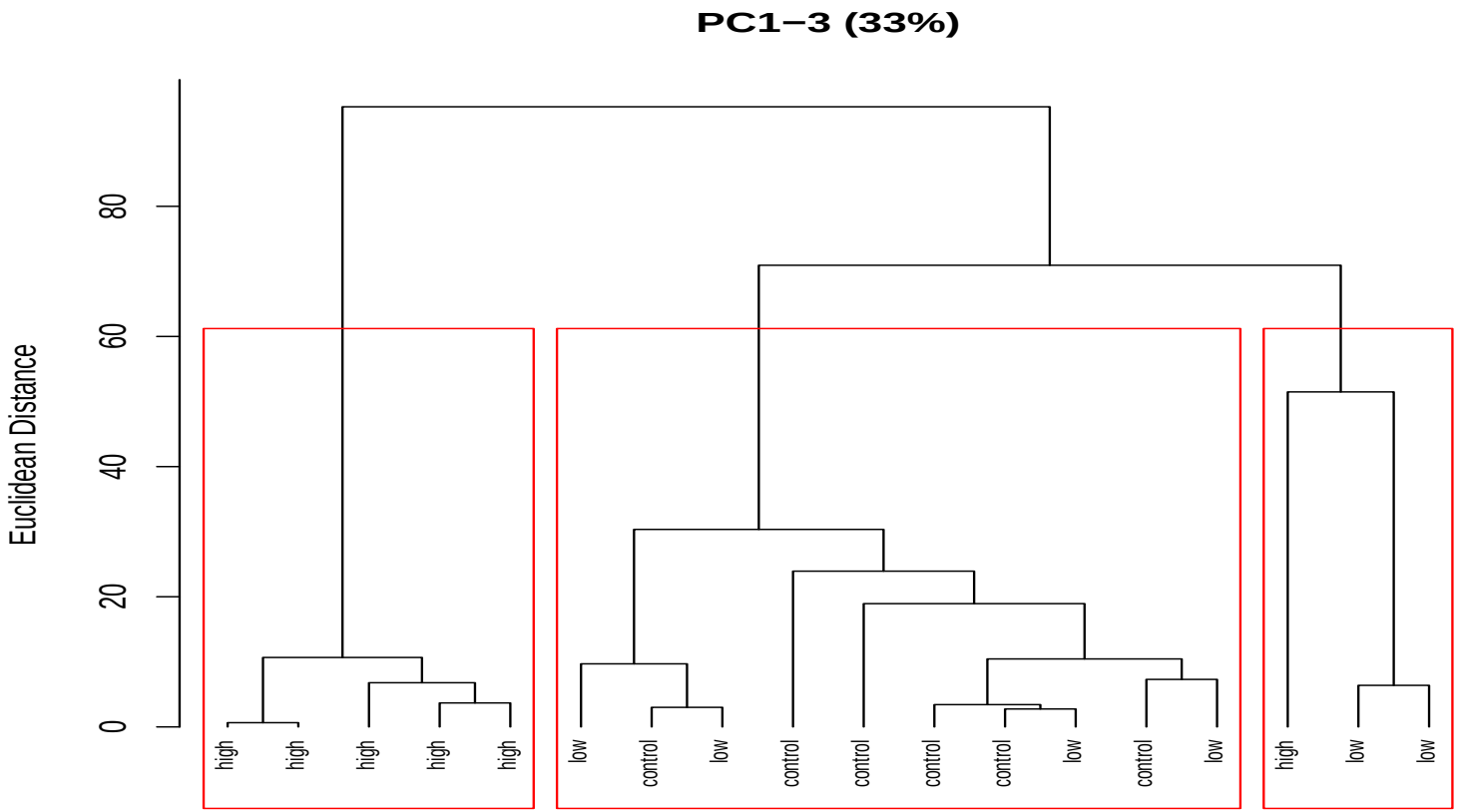
Two-Cluster Dendrogram




```
HCgroups = cutree(eward, 2); chisq.test(HCgroups, treatment)
X-squared = 13.8462, df = 2, p-value = 0.0009848
```

Clustering on Principal Components

Three-Cluster Dendrogram




```
HCgroups = cutree(eward, 3); chisq.test(HCgroups, treatment)
X-squared = 17.6, df = 4, p-value = 0.001477
```

Clustering on Principal Components

Step 5: Interpreting the Results

- Using this multivariate method, we determined that there were only two treatment responses, not the original three. In addition, there was a group of outliers from *different* treatment groups.
- To finish the evaluation, we need to know how the original variables influenced the first three principal components, and how that in turn can be used to describe the cluster groups.
- An approach that works with some data sets is to show summary statistics (e.g., minimum, median, maximum) for each cluster group. But that isn't feasible for presence/absence data.
- Another approach is to look at the variable scores for the components used to cluster the data. Because there were so many OTUs, we need to focus on the "best" variables, but defining "best" is somewhat arbitrary.

Clustering on Principal Components

Best Negative Variable Scores in PC1-PC3

OTU	PC1	OTU	PC2	OTU	PC3
M.5324	-0.085	M.521	-0.087	M.33548	-0.082
M.17875	-0.083	F.671	-0.087	M.29634	-0.076
M.18385	-0.083	uk.euk.3152	-0.087	uk.euk.9604	-0.074
M.34715	-0.083	S.3978	-0.087	M.10729	-0.074
M.25065	-0.082	S.7008	-0.087	M.4300	-0.073
M.28527	-0.082	M.7687	-0.087	R.848	-0.071
uk.euk.6428	-0.082	S.9894	-0.087	uk.euk.422	-0.071
M.1091	-0.081	S.10341	-0.087	S.947	-0.071
M.15718	-0.080	S.13318	-0.087	uk.euk.1428	-0.071
M.25537	-0.079	S.15201	-0.087	uk.euk.378	-0.071

Best Positive Variable Scores in PC1-PC3

OTU	PC1	OTU	PC2	OTU	PC3
R.30870	0.077	M.5776	0.052	uk.euk.5340	0.088
A.11234	0.073	M.4361	0.048	uk.euk.4435	0.084
A.5381	0.068	S.37132	0.044	uk.euk.8723	0.081
R.9197	0.066	M.15282	0.041	M.35761	0.080
S.3563	0.065	uk.euk.35868	0.040	M.35148	0.076
M.37060	0.060	M.10534	0.040	M.20011	0.075
M.37021	0.057	M.25299	0.039	R.7994	0.075
R.1633	0.056	M.588	0.038	A.25541	0.075
R.28351	0.056	F.35789	0.037	M.29451	0.075
S.20330	0.056	uk.euk.28316	0.037	M.26907	0.075

Clustering on Principal Components

OTUs Translated into Biotic Richness, from Chariton, et al.

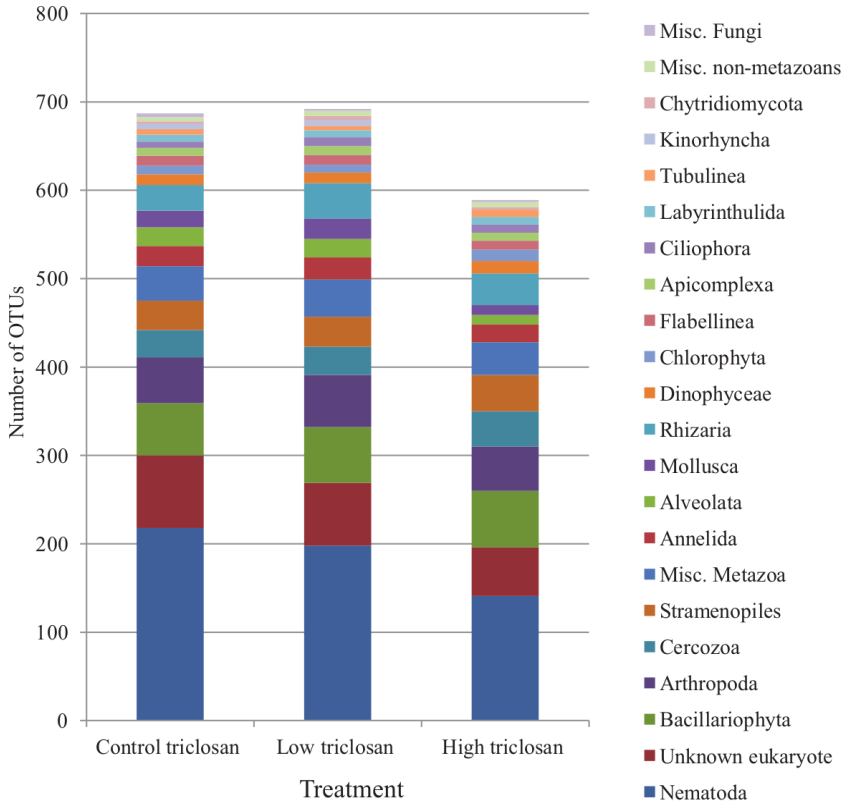


Figure 1. The biotic composition and richness of phyla (and other coarse taxonomic groups) sequenced from the surficial sediments of the 3 modified treatments at the completion of the experiment. To aid visual interpretation, data have been aggregated at the level of phylum and higher. OTU = operational taxonomic unit.

Clustering on Principal Components

Plotting the Samples by PC Scores

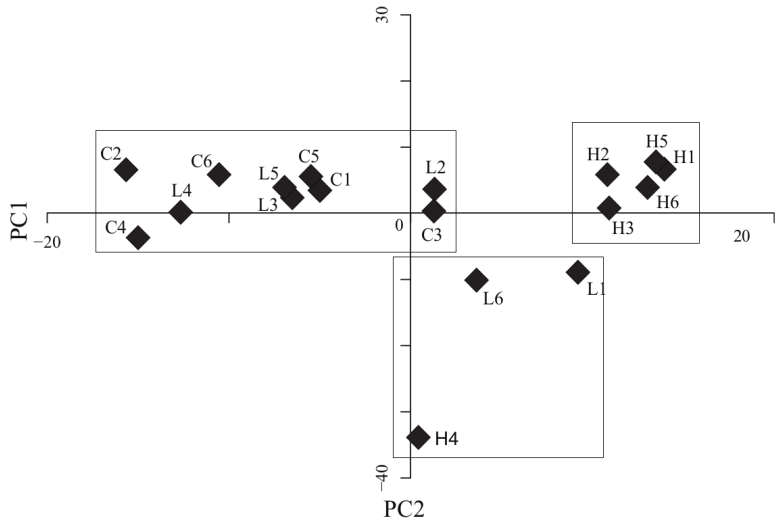


Figure 3. Samples plotted by principal components (PC)1 to 2 and grouped by principal component analysis cluster groups. Dashed rectangles show outlier groups (L1, L6, and H4). C = triclosan control; L = low triclosan treatment; H = high triclosan treatment.

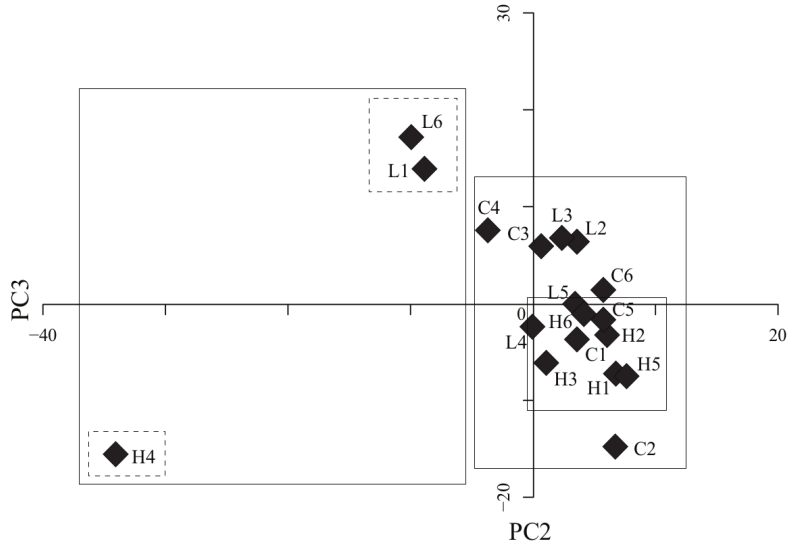


Figure 4. Samples plotted by principal component (PC)2 and 3 and grouped by principal component analysis cluster groups. Dashed rectangles show outlier groups (L1, L6, and H4). C = triclosan control; L = low triclosan treatment; H = high triclosan treatment.