

## N-STEPS Objectives

Provide regions, states, and tribes with support related to nutrient criteria development

Provide access to expert assistance with issues related to nutrient criteria development and implementation

Improve communication nationwide.

## What Is It?

Correlation is a measure of the strength of a relationship between two variables. Correlations do not indicate causality and are not used to make predictions; rather they help identify how strongly and in what direction two variables covary in an environment. In the context of nutrient criteria development, correlation analysis is a powerful tool to explore which variables may be strongly related to nutrient concentrations.

Types:

- Pearson (parametric, assumes linear relationship)
- Spearman (non-parametric, can be non-linear)
- Kendall's Tau (non-parametric, can be non-linear)

Example Question: How strongly is total phosphorus concentration related to the richness of macroinvertebrate taxa?

## How is it Applied to Nutrient Criteria Development?

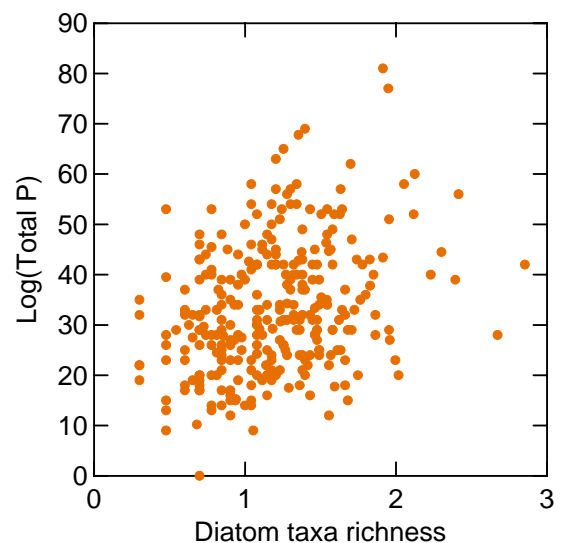
Nutrient criteria development involves three main processes: identifying relationships between biological responses and nutrient stressors, examining these relationships, and establishing nutrient and/or biological thresholds or criteria.

Correlation analysis is a powerful tool to identify the relationships between nutrient variables and biological attributes. The purpose of correlation analysis is to discover the strength of these relationships among a suite of nutrient and biological attributes and to select the most interesting relationships for further analysis. Correlations do not indicate causality and are not used to make predictions; they help identify how strongly and in what direction two variables covary in an environment.

## How Does It Work?

**Pearson Product Moment** - Calculates a correlation coefficient ( $r$ ) that is the ratio of the covariance of two variables (sums of products of both variables) to their individual variances (square of the sum of each variable). In other words, how much of the change in one variable is associated with changes in the other. Pearson correlation assumes the two variables are approximately normal and related in a linear fashion. Transformations can be used to help meet these assumptions.

**Spearman Rank Correlation** - Measures the monotonic relationship (one variable simply increasing or decreasing with another) between two variables. It makes no assumptions about the shape of that relationship. Procedure assumes that the two variables are randomly sampled from continuous populations. Values are ranked separately and the ranks among sites are compared using the same correlation coefficient as for the Pearson Correlation.



- Pearson  $r = 0.30$ ,  $p < 0.001$ ,  $N=294$
- Spearman  $r = 0.29$

## Data Requirements

Independently collected numeric data in the form of paired observations are required. These are preferably continuous data, although discrete numeric variables (e.g., taxa richness) can also work. The greater the range of environmental conditions encompassed the better. One way to assure a large range is to use a gradient design and select sites along as large a gradient as possible.

## What Should You Look For & Report?

Examine the bivariate plots and the distribution or behavior of the values. This will help choose which method to use. You should see some relationship between the two variables when plotted together if they are correlated. A non-significant correlation does not mean two variables are not related – the relationship may be non-linear or non-monotonic. Similarly, significant relationships may not mean much – often times large sample sizes produce significant correlations. Also, keep an eye on outliers, they can wreak havoc with correlations – especially with small datasets.

Report the correlation coefficient, the number of data pairs or degrees of freedom, and the significance (p-value) or type I error rate (alpha).

### Pros

- Effective way to convey simple relationships
- Easy to understand
- Quantitative measure of bivariate association

### Cons

- Hard to detect complex relationships (quadratic) without transformations
- Lack of significance does not mean lack of association
- Large sample sizes can lead to significant but small correlations
- Not predictive

## Alternatives

- Linear and non-linear regression
- Loess
- Multivariate analyses (for more than one variable at a time)

## Citations

EPA Statistical Primer - <http://www.epa.gov/bioindicators/statprimer/index.html>

Ott, R.L. 1993. An introduction to statistical methods and data analysis. 4<sup>th</sup> edition. Duxbury Press, Belmont, CA.

*For more information, contact:*

Steve Potts  
US EPA  
202-566-1121  
Potts.Steve@epa.gov

## What Is It?

### N-STEPS Objectives

Provide regions, states, and tribes with support related to nutrient criteria development

Provide access to expert assistance with issues related to nutrient criteria development and implementation

Improve communication nationwide.

One of the most common statistical modeling tools used, regression is a technique that treats one variable as a function of another. The result of a regression analysis is an equation that can be used to predict a response from the value of a given predictor. It can be used to consider more complex relationships than correlation by using more than two variables or combinations of different order equations (e.g., polynomials). Regression is often used in experimental tests where a range of fixed predictor levels are set and one tests whether there is a significant increase or decrease in the response variable along the gradient of predictor levels.

Example Types (\*treated in this fact sheet):

- Simple linear\*
- Multiple linear regression
- Non-linear
  - Logistic regression\*
  - Exponential regression
  - Polynomial regression

Example Question: Can I use total phosphorus concentration to determine the chlorophyll content in a lake?

## How is it Applied to Nutrient Criteria Development?

Nutrient criteria development involves three main processes: identifying relationships between biological responses and nutrient stressors, examining these relationships, and establishing nutrient and/or biological thresholds or criteria.

If a strong relationship between a biological parameter (e.g., algal biomass) and nutrient variables (e.g., total phosphorus) is or is not identified in a correlation analysis, scatterplots and regression analysis can be used to examine the relationship further. Regression analysis includes simple linear regressions, multiple linear regressions, and non-linear regressions. Simple regression analysis is similar to correlation analysis but it assumes that nutrient parameters cause changes to biological attributes. Nonlinear or multiple linear regression analyses can be used to consider more complex relationships between biological attributes and nutrient variables, such as nonlinear relationships and multiple predictors (e.g., both TN and TP are predictors of algal biomass).

## How Does It Work?

Simple linear regression - In least squares regression, the common estimation method, an equation of the form:  $E(y_i) = \beta_0 + \beta_1 x_i$  is estimated by finding values for the parameters ( $\beta_0$ - the intercept and  $\beta_1$ - the slope) that minimize the sum of the squared deviations between the observed responses and the linear equation. The variance of each parameter can be used to evaluate its significance. A significant slope means the slope is different from zero and there is a response to the predictor; a significant intercept means the intercept is generally different from zero.

Some Assumptions:

- Relationship between predictor and response is linear (n.b.: transformations used to make it linear)
- Error term is assumed to be normal (bell shaped distribution) with homogeneous variance -
- Samples are independent

Logistic Regression - Logistic regression is used to model a binary response (e.g., presence/absence of nuisance algae) with some predictor (e.g., nutrient concentration) using an equation of the form:

$E(y) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$ . The probability of response is modeled rather than the actual response value, typically using a computationally intensive iterative process. Logit ( $\log_e$  of odds ratio) gives a linear model.

### Some Assumptions:

- Error term is assumed to be non-normal with nonconstant variance
- Samples are independent

## Data Requirements

Independently collected numeric data in the form of paired observations are required – for both the predictor(s) and the response variable. These are typically numeric data, although discrete numeric and binary variables (presence/absence) can also be used. As with correlation, the greater the range of environmental conditions encompassed the better. One way to assure a large range is to use a gradient design and select sites along as large a gradient as possible.

## What Should You Look For & Report?

**Linear Regression** – Examine the plots and the final regression line. Examine the residuals of the regression for normality (equally spaced around zero), constant variance (no pattern to the residuals), and outliers. Report the regression equation, the significance of the model, the degrees of freedom, and the significance of each of the parameters (t-statistics and p-values for the slope and intercept). It is not uncommon to also report any unusual features of the residuals, if they exist. Finally, report the coefficient of determination ( $R^2$ ) which measures what proportion of the variability in the relationship is explained by the regression. This varies from 0 to 1, where 1 means the regression explains 100% of the variability in the relationship (i.e., all the points fall right on the regression line).

**Logistic Regression** – Examine the plots and final regression line. Use a goodness-of-fit test to determine the appropriateness of the model. Fitted responses should approximate monotonic curves with sigmoidal shapes. Formal and informal tests of this are available (e.g., Hosmer-Lemeshow). Report the parameter estimates and the associated significance, and the goodness-of-fit results. One can use the parameters to determine the probability of a response for a particular value of the predictor.

For both forms, avoid extrapolation – predictions beyond the range of the predictors used to build the models, as confidence outside that range is low.

### Pros

- Efficient analysis that is useful for prediction
- Easy to interpret
- Can use more than one predictor
- Synergistic relationships can be modeled

### Cons

- Assumptions about variables constrain analysis
- Evaluating models becomes more difficult with complexity of model
- Shape of response necessary to choose the best model
- Sensitive to outliers and, like most models, hard to extrapolate

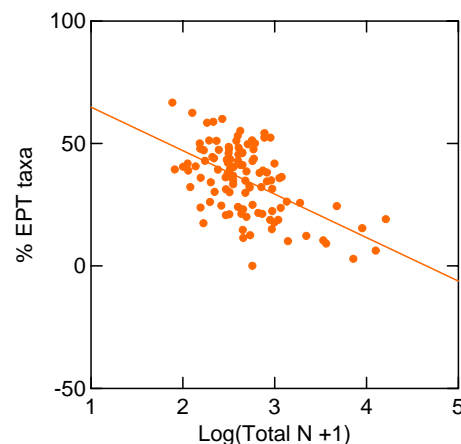
### Alternatives

Generalized Linear Models  
Generalized Additive Models  
Nonparametric regression

### Citations

EPA Statistical Primer -  
<http://www.epa.gov/bioindicators/statprimer/index.html>

Ott, R.L. 1993. An introduction to statistical methods and data analysis. 4<sup>th</sup> edition. Duxbury Press, Belmont, CA.



$$\begin{aligned} \% \text{ EPT} &= 82.6 - 17.8 * \text{Log}(\text{TN}) \\ F &= 44.2, p < 0.001 \\ df &= 106 \\ R^2 &= 0.29 \end{aligned}$$

*For more information contact:*

Steve Potts, USEPA  
202-566-1121  
Potts.Steve@epa.gov

# LOESS (or LOWESS)

## N-STEPS Objectives

Provide regions, states, and tribes with support related to nutrient criteria development

Provide access to expert assistance with issues related to nutrient criteria development and implementation

Improve communication nationwide.

## How does it work?

Loess is fairly straightforward. A specific width of points along the x axis is selected (the bandwidth or tension) adjacent to the point being predicted, and a low degree polynomial equation (often just linear) is fit through that subset of the data. More weight is given to points closest to the value being predicted. This resulting equation is then used to

## What Is It?

Loess stands for locally estimated scatterplot smoothing (lowess stands for locally weighted scatterplot smoothing) and is one of many non-parametric regression techniques, but arguably the most flexible. A smoothing function is a function that attempts to capture general patterns in stressor-response relationships while reducing the noise and it makes minimal assumptions about the relationships among variables. The result of a loess application is a line through the moving central tendency of the stressor-response relationship. Loess is essentially used to visually assess the relationship between two variables and is especially useful for large datasets, where trends can be hard to visualize.

Example Question: Is there some non-linear trend hidden among the noisy relationship between chlorophyll and total phosphorus?

## How is it Applied to Nutrient Criteria Development?

Nutrient criteria development involves three main processes: identifying relationships between biological responses and nutrient stressors, examining these relationships, and establishing nutrient and/or biological thresholds or criteria.

By combined with scatterplots, locally weighted scatterplot smoothing (LOESS) is used to examine biological attribute changes along a nutrient gradient. It is designed to address nonlinear relationships where linear methods do not perform well. Loess fits a regression line through the moving central tendency of a biological attribute along the nutrient gradient. As a result, the trend of biological attribute changes along a nutrient gradient can be observed in a scatterplot with a large dataset. Loess can be used to examine the threshold change of biological community along a nutrient gradient, if a threshold exists.

predict the value for the selected point. The data are then shifted one point to the right and the process continues, with a new prediction for the second point, and so on. The resulting points are then connected together with a line. The user can control how wide a band of points are used – the smaller the bandwidth, the fewer points that are used and the less smooth the final line. Users can also adjust the type of line-fitting that is used – weighted least squares is the most common. Users can also adjust what types of weights are used.

Some Assumptions:

- Very few
- Need a lot of data – the more the better

## Data Requirements

Independently collected numeric data in the form of paired observations are best. These are typically continuous numeric data, although discrete numeric data can be used. As with correlation and regression, the greater the range of environmental conditions encompassed the better.

## What should you look for and report?

One should look for a line that represents the smoothest trend in the data while minimizing random noise. Users should report the bandwidth that was used. One could also report the weights that were applied, if known, and what type of fit was used (linear or some higher order polynomial). There are typically few evaluation parameters to report.

### Pros

- Simple and flexible
- No assumptions about the relationships between variables
- Valuable for visualizing complex relationships
- Users can estimate new values to the fit and validate models if needed

### Cons

- Requires densely sampled datasets
- No ready formula is produced, so it is hard to transport the results
- Computationally intensive – but not a problem for most computers
- Sensitive to outliers

## Alternatives

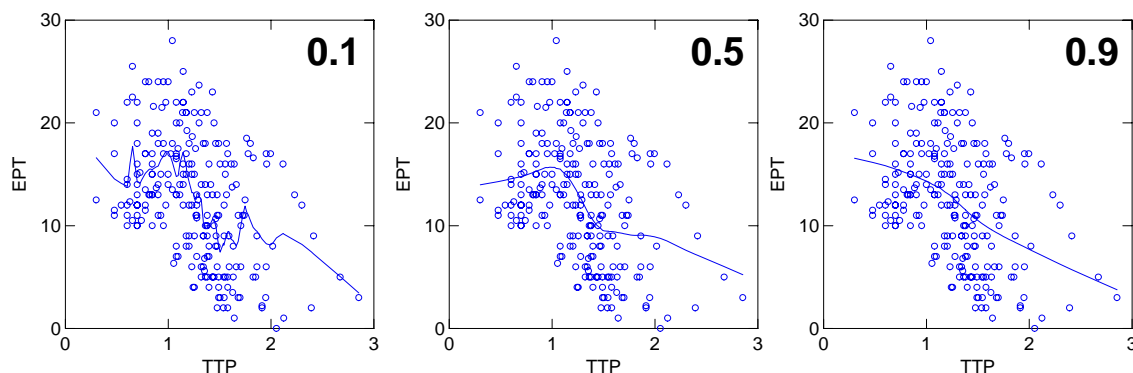
Linear and Non-linear Regression  
Generalized Linear Models  
Generalized Additive Models  
Other non-parametric regression and smoothing techniques

## Citations

Cleveland, W.S. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association 74:829-836

Cleveland, W.S. and Devlin, S.J. 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association 83:596-610

NIST/SEMATECH e-Handbook of Statistical Methods, 4.1.4.4. Loess (aka Lowess)  
<http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm>



Loess plots with different bandwidths

*For more information, contact:*

Steve Potts  
US EPA  
202-566-1121  
Potts.Steve@epa.gov

## N-STEPS Objectives

Provide regions, states, and tribes with support related to nutrient criteria development

Provide access to expert assistance with issues related to nutrient criteria development and implementation

Improve communication nationwide.

## What Is It?

Change-point analysis is a method for identifying thresholds in relationships between two variables. More specifically, it is an analytical method that attempts to find a point along a distribution of values where the characteristics of the values before and after the point are different. In the case of a stressor-response relationship, change-point analysis can be used to identify that point along the x axis where characteristics along the y axis change – implying a shift in the average of variance or a change in slope.

Example Question: Is there a threshold in the response of the number of EPT taxa to gradients in total phosphorus?

## How is it Applied to Nutrient Criteria Development?

Nutrient criteria development involves three main processes: identifying relationships between biological responses and nutrient stressors, examining these relationships, and establishing nutrient and/or biological thresholds or criteria.

Change-point analysis is a statistical method for identifying thresholds and it is essential for nutrient criteria development. More specifically, it is an analytical method that attempts to find a point along a distribution of values where the characteristics of the values before and after the point are different. In the case of algal biomass' response to total phosphorus (TP) concentrations, change-point analysis can be used to identify that TP concentration where average algal biomass shifts significantly before and after that TP concentration.

## How Does It Work?

There are a few methods change-point analysis. The basic method uses deviance reduction. This approach finds the point along a distribution of points where the sum of deviances on either side of the point is lowest compared to the overall dataset deviance. The percent of error reduction associated with splitting the data is then calculated. This is an iterative process, moving along the data systematically and evaluating that point which minimizes the deviance reduction. This is essentially equivalent to a regression tree with one split and that method can be used with one predictor, limiting the process to one split in the data. However, users can select a variety of approaches to use including: non-parametric deviance reduction, least squares, cumulative summing, Bayesian estimation, etc. Bootstrap methods can be applied to estimate error around the identified threshold for non-bayesian methods.

## Data Requirements

Independently collected data in the form of paired observations. These are typically continuous numeric data, although discrete numeric and categorical data can also be used. As with correlation and regression, the greater the range of environmental conditions encompassed the better.

## What Should You Look For & Report?

One is looking for thresholds that result in substantial reduction in the percent error. So most techniques include some measure of error reduction and this should be reported. Some software

applications use a  $\chi^2$  test to evaluate the significance of the error reduction. Confidence in the threshold should also be reported. Bayesian approaches will produce a range for the thresholds, whereas other non-parametric approaches can produce an estimate of confidence in the threshold using bootstrap re-sampling.

### Pros

- Identifies clear thresholds
- Non-parametric techniques are distribution free and have limited assumptions
- Confidence intervals can be used to set criteria

### Cons

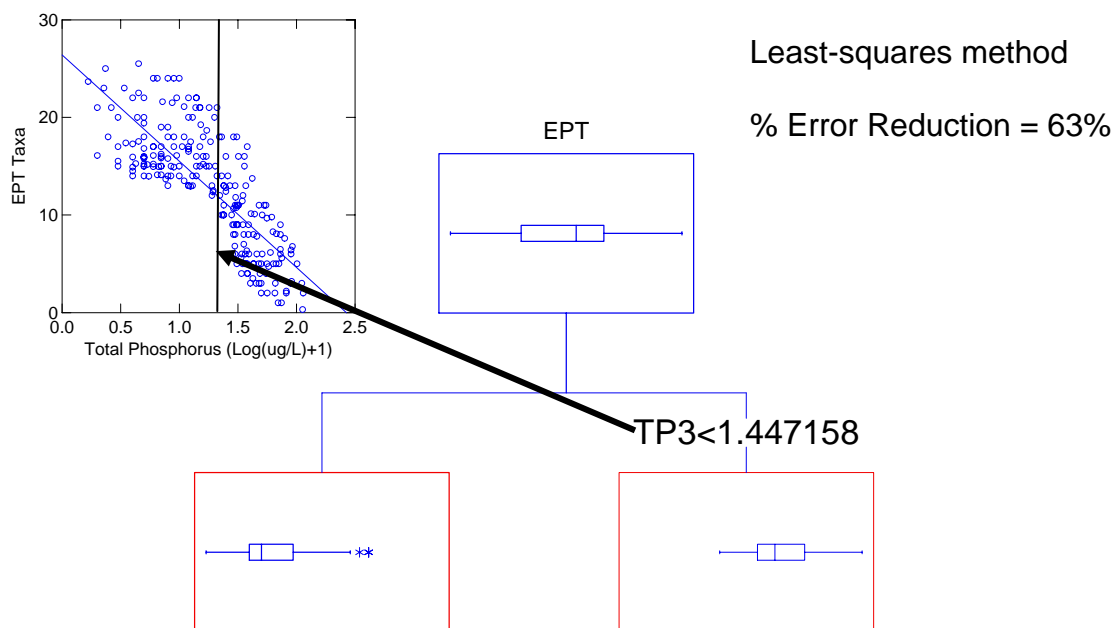
- Data hungry and computationally intensive
- No ready formula is produced, so it is hard to transport the results
- Sensitive to variability in data

## Alternatives

Few - thresholds analysis is not easy or well developed  
Conditional probability analysis  
Visual methods - loess and straight scatterplots

## Citations

Qian et al. 2003. Two statistical methods for the detection of environmental thresholds. Ecological Modelling 166:87-97



For more information contact:  
Steve Potts, USEPA  
202-566-1121  
Potts.Steve@epa.gov