

Irwin, R. 2011. Exploratory Data Analysis (EDA) for Beginners. Presented at the George Wright Society National Meeting New Orleans, March, 2011, Revised to this updated version. Final Working Draft Version, Internet Distributed Only at <https://irma.nps.gov/>

December, 30, 2011, National Park Service Water Resources Division, Fort Collins, Colorado.

Table of Contents:

Acknowledgments.....	4
Introduction:.....	5
EDA Basics:.....	6
Don't Confuse EDA with Bad Practices.....	7
Level One EDA.....	8
Level One Basics:	8
Level One Summary Statistics:.....	8
Don't make the list too short:.....	8
Don't Depend On Summary Statistics Alone.....	9
Don't Rely Only on Standard Deviations When Assumptions Are Violated.....	11
Don't Report the Mean Plus or Minus the SD or SEM as an Interval.....	11
Don't Fall for Common Misunderstandings of SDs, SEMs, and Variances	12
Don't Report The CV Without The Sample Size Or When What You Mean Is The RSD..	13
How Are RSDs used in QA/QC?.....	13
What are Some Rules of Thumb Related to CVs and RSDs?.....	14
Are There Nonparametric Alternatives To The RSD and CV?	14
Don't Read Too Much into Correlation Coefficients	15
Include Single Data Point Uncertainty	15
Level One EDA Plots:	16
Very Common Simple EDA Plots:	16
Less Common Simple EDA Plots:.....	17
There is Always a Model	18
Software Options for Simple EDA Plots:	18
Smoothing Basics:	20
THE ALPHABET SOUP OF SMOOTHERS:	20

Cautions Related To Smoothing	21
Don't Just Force It To A Straight Line.	21
Don't Ignore Assumptions or Discard Raw Data.	22
Don't Throw Out the Baby with the Bath Oil.....	22
Should One Never Test for Trends in Seasonal Medians?	24
More Detail on The common Smoothing Techniques:.....	26
Moving Median Smooths.....	26
Simple Moving Average (SMA) Smooths.....	26
Exponential Moving Average (EMA) Smooths	27
LOWESS and LOESS Smoothers:	27
LOWESS and LOESS Are Not Exactly the Same.....	27
LOWESS and LOESS Software	29
More Complex Smoothers	29
TYPICAL QUESTIONS TO BE ANSWERED WITH LEVEL ONE EDA TOOLS:	30
Is the Distribution Normal?	30
What is a Probability Density Function (PDF)?	31
What are Cumulative Distribution Function (CDF) Line Plots?	32
If not Normal, What is the distribution?	32
What is the Relationship Pattern between Two Variables?	33
Hints from Visual Inspection:	33
Hints from Simple Curve or Line Fitting Software:	35
Cautions on Curve Fitting in MS Excel:.....	36
Hints from a Good Smoother:.....	37
Hints from Dedicated Statistical Software:.....	37
Is it Linear?	38
Is There A Non-Linear Trend Or Other Non-Linear Pattern?	38
Is it Stationary?	39
Are There Cycles?.....	40
Are There Outliers?.....	40
Don't Just Throw Out all Outliers!	41
How Does Software Identify Outliers?.....	42
What is Data Verification?.....	42
What is Data Validation?	43
Are There Patterns in Missing Data?	43

Are There Patterns in Nondetects?	44
Level Two EDA	45
Level Two Basics:.....	45
Level Two EDA Software:	46
TYPICAL QUESTIONS TO BE ANSWERED WITH LEVEL TWO EDA TOOLS:	47
How Often and When Should I Collect One Variable?	47
How Often and When Should I Collect Two Variables I Want to Compare?	48
Do More Complex Lines of Evidence Also Suggest Normality?.....	50
Is My Data Is Independent?	51
What is Autocorrelation?	51
What EDA Plots Help Us Understand Autocorrelation Patterns?	53
What Kinds Of Datasets Tend To Have Fewer Problems With Autocorrelation?	53
What EDA is Done for a Complete Census?	54
How Do I Plot an ACF correlogram	54
Which kind of ACF Correlogram is my Software Drawing?	55
What is the rough approximation ACF correlogram?.....	55
How do I interpret the results of a rough approximation correlogram?	56
How Does a Correlogram Look When There is a Trend?	57
How Does a Correlogram Look When There Are Strong Cyclical Patterns?	57
At What Lag Does Autocorrelation Disappear?	58
Hints from a Rough Approximation ACF Correlogram	58
Hints from a Large Lag Standard Error ACF Correlogram	58
Hints from More Complex Lines of Evidence:.....	59
How Do I Plot The Large Lag ACF Correlogram?	59
What does the PACF tell me?	60
How Do ARMA and ARIMA Models Fit Into The Grand Scheme Of Things?	62
What is Differencing?	63
What is Pre-whitening?	63
Is the autocorrelation first order autocorrelation only?.....	65
Why Do I Need to Know if First-order Autocorrelation is the Only Kind Present?	65
What is Effective Sample Size and How is it Used?	65
Potential Complications with Effective Sample Size	66
Relatively Simple Equation Options for Effective Sample Size	67
What Equation Is Used in a PACF Correlogram?	68

What Insights Do My ACF and PACF Correlogram Give Me About Model Fitting?.....	68
Are There Also Tests for Autocorrelation?	70
What Autocorrelation EDA Tools Work For Data Not Regularly Spaced?	70
What EDA Helps Identify a Longer Term Trend In Spite of Shorter Term Cycles?	71
What Are Some More Advanced EDA Tools for Cyclical Patterns?	72
Is There A Simple Way to Handle Spatial and Temporal Autocorrelation Together?	72
Should I Calculate Confidence Intervals as Part of EDA?	73
Should I Calculate Prediction Intervals or Tolerance Intervals as Part of EDA?	74
What are Measurement or Manufacturing Process QC Control Charts?	75
Level Three EDA	77
TYPICAL QUESTIONS TO BE ANSWERED WITH LEVEL THREE EDA TOOLS:.....	77
How Do I Use Decomposition Plots in EDA?	77
Decomposition Software	79
Can Relatively Simple Trend Tests be Part of EDA?	80
What Should Be Kept in Mind About all Trend Tests?	80
What Are Seasonal Kendall (SK) Tests?	80
How Do I Confirm A Monotonic Linear Trend?	81
What Simple Trend Test Confirms a Longer Term Trend In Spite of Shorter Term Cycles? .	82
How Do I Confirm A Non-Linear Trend Or Pattern?	83
How do I Use Control Charts for More Complex Purposes?	83
Looking for Differences from Background with Control Charts	84
Looking for Trends with Control Charts:	84
How Do I Handle Nondetects in Advanced EDA and Trend Testing?	85
Software for Handling Nondetects	86
How Do I Handle Spatial and Temporal Autocorrelation Together and Still Use all the Data?	
.....	87
Is There Multivariate EDA?	88
How Does EDA Compare to FDA?	89
How Do I Follow Up EDA with Trend Analyses When I Have Rotating Panels and/or a Probabilistic Sampling Design?	89
Some Examples of Putting It All Together	89
References Cited:	91

Acknowledgments

Helpful comments on selected portions of text within his document and/or or permission to quote certain statements or websites were received from Dennis Helsel of Practical Stats (Consulting Firm) in Denver Colorado, Graham McBride and Ian Jowett of NIWA in New Zealand, David Meko of the University of Arizona, Jim Loftis of Colorado State University, Patrick Wessa of the Lessius Business School in Antwerp Belgium, Ken Gerow of the University of Wyoming Statistics Department, Jeffrey Morisette, David Lorenz, and Robert Hirsch of USGS, Kirk Cameron of MacStat Consulting, Anita Singh of Lockheed Martin, Stan Brown of Tompkins Cortland Community College, Craig Polhemus of Statgraphics Software, David Garson of North Carolina State University, Matthew Hall of the State of Maryland, and James Kirchner of the University of California at Berkeley. Within the National Park Service, helpful comments on selected portions of text were received from Tom Philippi, Gary Smillie, Jeff Runde, David Thoma, Pete Penoyer, Rick Inglis, Eric Starkey, Jeff Shearer, Bill Monahan, Rob Daley, Dilip Shinde, Tahzay Jones, and Christopher Sergeant. I am thankful for all the many suggestions. However, any imperfections or less than optimal discussions or other shortcomings herein should be attributed only to Roy Irwin.

Introduction: This summary is a longer and revised version of the handout given out at the George Wright Society National Meeting in New Orleans in 2011. It might be thought of as a basic introduction to EDA for environmental monitoring specialists. It is being called a final working draft because I ran out of time for final editorial polishing before retiring December, 2011. Those who notice needed final edits are asked to send suggestions to me at rji44@yahoo.com. However, it is being distributed in this version so that those who have requested it can use it as a working draft without delay.

The target audience for this summary is statistical beginners, often biologists or environmental monitoring specialists whose statistical expertise could use some updating. The emphasis herein is on long term aquatic/water quality monitoring such as is being done in the National Park Service Vital Signs Monitoring Program. A frequent focus is on trends. However, hopefully the EDA basics discussed herein will also be helpful for those doing shorter term projects and projects more focused on spatial issues.

At the risk of oversimplification, the attempt is to make this summary as plain language as possible. There is often a choice between complex and simpler EDA. The main emphasis herein is trying to cover the relatively simple approaches, with shorter mentions of more complex approaches. Once beginners have been empowered to understand the basics and take the baby steps summarized herein, the hope is that they may then be motivated to learn more as their careers progress. If nothing else, hopefully the following summary may help teach readers some basics which would make them able to communicate their remaining questions to statistical experts with more clarity.

Example EDA plots are not universally included in this document in order to keep the word count and digital file space magnitude down, but example plots are only few clicks away in the free Internet references cited. Since this document is designed to be used primarily as a digital document helpful even in very remote National Parks, when possible we cited information from comprehensive and free Internet summaries. Thus, the references cited section gives hyperlinks to most references, and those available on the Internet are bolded.

EDA Basics: Exploratory Data Analysis (EDA) is an approach to data analysis that seeks to find patterns in data, to reveal structure, to suggest possible relationships among variables, and to find other hidden secrets in the data. EDA tends to give us some of our first hints about processes possibly behind the patterns seen in the data. EDA should be thought of more as an “**attitude taken**” rather than a standardized list of techniques, since different methods may be needed for different datasets. EDA techniques tend to include many plots, displaying results on graphs with no or few assumptions in an attempt to explore the data in an open-minded way by looking at the data from many different angles. During early EDA, one also typically begins to assess [outliers](#) and to look for clues about the validity of initial assumptions about data characteristics such as **independence, normality, linearity, and stationarity** (lack of trend). Understanding these data characteristics is important in choosing logical statistical procedures to be done next. During later EDA steps, one begins to use progressively more complex methods to more conclusively determine data characteristics.

A 2001 National Water Quality Monitoring Council summary on statistics concluded that one argument against having standardized “recommendations” for routine statistical analysis methods is that having them would tend to censor the methods that might be used for exploratory data analysis. In other words, exploratory data analysis employed by researchers needs to remain un-tethered and flexible (Griffith et al. 2001).

Especially in long term monitoring or other long term projects, EDA should not be something one does early on and then never does again. It is something one should do early on then again later when more data and understanding is gained. It is better to continue the EDA attitude to "let the data do the talking" rather than holding too strongly to early preconceptions about potential patterns, explanatory factors, or how the data should be grouped into categories.

In the EPA Mid-Atlantic intensive project 73% of the initial professional judgment-picked reference sites later failed to meet the independently-established multi-lines of evidence (chemical, biological, and habitat) criteria for reference condition. Probabilistic sampling not only disqualified some reference sites, it also revealed new ones not previously recognized as unimpaired (EPA, 2003). Thus, early theories should always be confirmed with hints from evidence from real data expressed as EDA plots, summary statistics, and comparisons with objective criteria.

In long term monitoring for status and trends, one goal of EDA is not to bias data interpretation or to hinder finding potentially important patterns not initially recognized or that might emerge in response to new stressors. In this regard, continuing to look at data from different angles to get new insight into what the dataset is telling us is consistent with information needs related to “scenario planning” for potential climate change. We don’t know what combination of stressors may be important to natural resources in the future. Remember no amount of statistical sophistication can make up for missing the most important variable(s). Keeping an open mind might also help identify information helpful in future adaptive management.

Environmental data is seldom normally distributed. Helsel and Hirsch (2002), and the father of EDA Tukey (1977) and others have emphasized many nonparametric methods that can be used

to avoid the overuse of only parametric statistics and only parametric EDA on data not normally distributed. This helps avoid back transformation bias and “changing the question” pitfalls sometimes associated with trying too hard to make the parametric methods work. Not factoring in the bias resulting from back-transformations may partly explain why some early predictive models failed to predict the relatively fast rate at which some reservoirs have been filling up from sedimentation (Koch and Smillie, 1986).

Although simple exploratory graphs could often suggest that normal distribution assumptions are probably not met, too many investigators skip the simple EDA plotting steps necessary and don't fully check assumptions, and instead just use normal distribution (parametric) statistics (McBride, 2005). It is also perhaps too common for investigators to skip quickly to linear regression or linear trend methods without discovering the pattern was clearly non-linear (which they might have discovered had they plotted the data in simple graphs).

Continuing EDA is a very good practice but takes time and financial resources. Therefore, it should probably be (more often) planned for during budgetary planning. In other words, preparing data for EDA often includes processing and combining selected environmental and geophysical data as potential covariates with the sample data from the system or process being studied. In some cases it takes a good deal of time and expertise to prepare these inputs for exploratory analysis and other statistical analyses, so program and project managers should consider and account for the professional resources required to support not only data processing for routine analysis, but also for a meaningful level of exploratory analysis (Rob Daley, NPS GRYN, Personal Communication, 2011).

Don't Confuse EDA with Bad Practices: Whether EDA is used or not, avoid bad practices such as transforming data in different ways until the answer that supports a pet theory appears or using test after test until one is found that gives a desired p-value or other result. Such practices are associated with terms data dredging or data fishing. These are simply unethical and non-transparent/bad science practices to be avoided and should not be confused with either EDA done correctly or with data mining done correctly. Concern about these bad practices is more often associated with shorter term experimental and hypothesis driven projects, but they should be avoided regardless of how long a project lasts.

Anderson and Burnham (2002) and Anderson et al. (2001), though sometimes (wrongly) cited in arguments to avoid EDA, actually primarily emphasized avoiding bad practices, particularly in limited term experimental work. Any methodology can be misused and that misuse of methods is not the fault of the methods. Anderson and Burnham (2002) stated that they fully support some post-hoc exploratory activities to better understand the data, but that one should transparently admit one is doing so. The potential for finding spurious effects and wrongly handling data interpretation increases when the data are not fully representative of the target population, when assumptions are not met, when sample sizes are too small, when lack of independent data is not documented or accommodated, and when sophisticated methods are used instead of simple analytical approaches and the data are not sufficient to support the premise of the more sophisticated approach (Anderson et al., 2001),

Usually it is better to look for hints of trends or other patterns from multiple lines of evidence while keeping some lines of evidence relatively simple.

The structure of complex outdoor environmental and ecological data is seldom fully known ahead of time. So, although specifying hypotheses in advance is a good idea in science, specifying every step of a statistical analysis in an advance is a luxury we don't have in most environmental and ecological studies. The key is to admit what one did and what didn't and didn't work. This approach is not only valid but often the only approach available (Zuur et al, 2007)

Remember to put science before statistics. Statistical considerations are used to identify a plan that is effective in meeting scientific goals -- not the other way around. Use pilot studies, or past data to help plan future work (Lenth, 2011).

It is possible to combine exploratory analysis and hypothesis-driven analysis into a larger study. One way of doing this is to perform a 2-phase study, in which the first phase is an exploratory analysis, perhaps involving subjectively located plots and employing many variations on analysis. The patterns found in the first phase are then posed as hypotheses for the second phase. The second phase involves the collection of fresh data from objectively located plots, and an entirely planned data analysis (Palmer, 2011).

EDA is discussed below in order of increasing complexity, so that most readers can find something fitting their level of understanding:

Level One EDA

Level One Basics:

Early EDA done for progress reports has often been an early look at the basic characteristics of results by looking at summary statistics and simple plots. Among other things, early plots and summary statistics often provide our first clues about whether or not the distribution being sampled may be a normal distribution and whether or not trends through time and relationships between variables are linear. Strict normality and linearity tend to be unusual for environmental data, but recognizing the true underlying patterns in the data can be important in choosing later statistical analyses. The Kendall-group of trend tests are more tolerant of a lack of strict normality and strict linearity than a linear regression, which is one reason they tend to be so popular for testing water quality trends.

Level One Summary Statistics:

Don't make the list too short:

Too often early summary statistics have included only a few normal-distribution statistics, such as means, standard deviations, variances, [coefficient of variation \(CV\)](#) and a standard error of the mean ([SEM](#)). That hints the investigator has already assumed a normal distribution model would fit the data, perhaps without even checking to see if there are any summary statistics or EDA graphical hints of that pattern.

One reason not to focus only on a few parametric summary statistics is that (parametric) means tend to drive ecological responses less than extremes, and how often or how long values stay at extremes. Standard deviations are not ideal expressions of variability when the distributions are not normal, sample sizes are too small, a sample is not representative, or when autocorrelation is present (as it often is, even though many do not even check for its presence).

Whether or not they are included in reports, for typically skewed data from the environment, one should also routinely calculate and interpret the meaning of medians, percentiles above or below benchmarks, kurtosis, skewness, range, minimum, maximum, sum, count (N), and nonparametric confidence intervals about a mean.

Percentiles calculated often include the 25th and 75th percentiles (for IQR ranges in box and whisker plots), as well as other percentiles to help identify outliers. The 5th and 95th or the 10th and 90th percentiles are often useful in comparison with thresholds or other benchmark standards, background levels, or outliers. MS Excel makes it easy to quickly and simultaneously calculate all such early EDA summary statistics using the Toolpak add-in, though many prefer to calculate final statistics in more uniformly accurate dedicated statistical software programs.

It is often also helpful to estimate the percent of nondetects and the percent of missing data.

To get away from over reliance on normal distribution statistics such as a standard deviation or variance to describe basic variability one might not only report IQR but also consider calculating a nonparametric [F-pseudosigma](#), a kind of nonparametric alternative to a standard deviation.

One might also calculate an average deviation. The average deviation is the simpler way to say average absolute deviation. Another option would be to calculate a median deviation from a sample median. This is referred to in some references as a median absolute deviation (MAD).

Although average deviations and median deviations do not have the same uses in higher inferential statistics that a standard deviation has, they can reported as stand-alone summary statistics that are easier to understand and are less dependent upon a normal distribution model than a standard deviation..

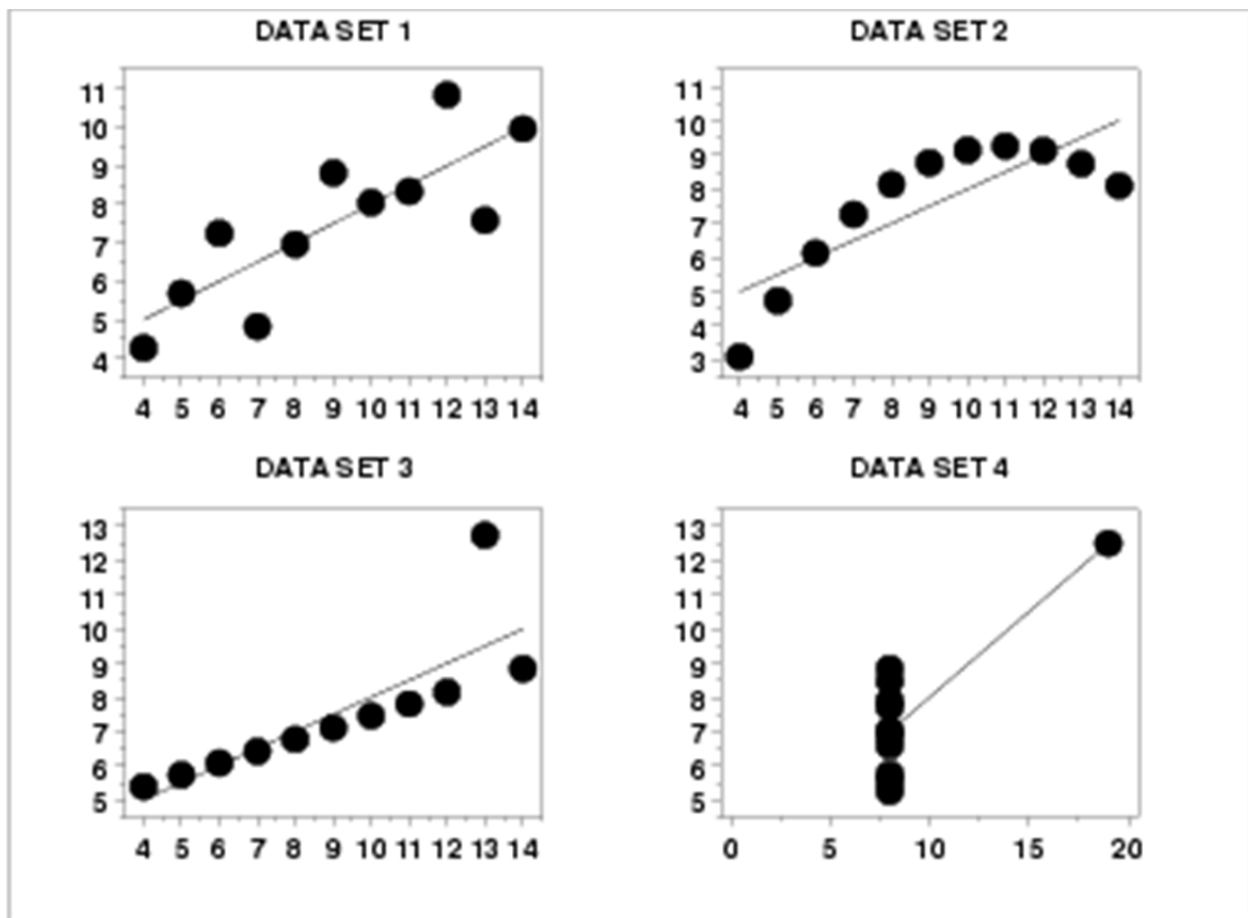
When comparing two variables common to calculate a [correlation coefficient](#)

Don't Do These:

[Don't Depend On Summary Statistics Alone.](#)

Classical parametric estimation techniques have the characteristic of taking all of the data and mapping the data into a few numbers (estimates of means, variances, standard deviations, etc.). This is both a virtue and a vice. The virtue is that these few summary statistics focus on important characteristics (location, variation, etc.) of the population. The vice is that concentrating on these few characteristics can filter out other important characteristics

(skewness, tail length, autocorrelation in time or space, percentile based characteristics, etc.) of the same population. In fact, basic data characteristics and distributions as suggested by plots can look radically different in plots even though summary statistics are identical. Thus, an EDA approach often entails less loss of information due to this "filtering" process of reducing information to summary statistics, especially if those summary statistics are only means and standard deviations and include few if any nonparametric statistics NIST (2006). That same NIST summary provided the following dramatic example from <http://www.itl.nist.gov/div898/handbook/eda/section1/eda16.htm>. The below four patterns have identical means, standard deviations, slope, and [correlation coefficients](#)!



This provides a dramatic example of why one should not rely only on summary statistics, but should also include simple EDA plots as another line of evidence. The means and standard deviations simply don't tell the full story. The more important stories from the plots above include the following:

The upper left plot hints at a (mostly) linear relationship with some scatter. If the data points were closer to the line, it would hint at an even stronger linear relationship.

The plot on the upper right hints at a nonlinear relationship, most likely a quadratic, pattern. It could also be part of a longer cycle.

The plots on the lower left hints at a linear pattern with one [outlier](#).

The plot on the lower right hints at a poor survey design and one outlier.

Don't Rely Only on Standard Deviations When Assumptions Are Violated

As introduced in our discussion of the mean plus or minus the standard deviation, summary statistics that rely on a standard deviation (including CVs, RSDs, and variances) are only optimal when the distribution is normal or something close (at least mostly symmetrical). If the distribution is radically different, the assumption is violated. Also, as will be discussed in more detail just below, a standard deviation is only a good representation of the target population being measured when the full range of conditions is sampled in way that insures that both common values and extremes are measured in proportion to their occurrence in the underlying target population.

The full range of conditions is typically not covered with small sample sizes. Those tempted to calculate a standard deviation when sample size is two to eight might consider the following. The main reason for calculating the commonly used $n-1$ sample standard deviation is to get an estimate of the population standard deviation. “An obvious but often overlooked fact is that there is no point in estimating something about the parent population unless our estimates are reliable” (Phillips, 1999).

Don't Report the Mean Plus or Minus the SD or SEM as an Interval

One of the most common and flagrant misuses of summary statistics in both EDA and in final data reporting, especially in certain biological journals, is to report the mean and/or the standard error of the mean (SEM) as an interval about the mean. This is especially problematic when sample size is not reported with the interval.

Neither of these intervals is accurate at small sample sizes and neither is optimal unless the sample is fully representative of the full range of conditions of the underlying target population being sampled and that population is perfectly normally distributed, both of which are unusual for environmental data.

It is fine to report the standard deviation as a common stand alone summary statistic, just don't report that as the only form of variability estimate, especially if the distribution is not normal, the sample size small, and/or the sample is not fully representative of the target population being sampled. When the standard deviation is reported as an interval about the mean it is just a crude 68% tolerance interval of questionable value with no accompanying statement about our confidence in the size the interval. This interval provides little to help one interpret data, especially when not accompanied by sample size and/or when applied to data that is not from a normal distribution. More detail:

A mean plus or minus the standard deviation is a very crude 68.26% parametric tolerance interval, where the upper and lower tolerance levels should bound roughly 68% of the underlying population being sampled. Unlike better tolerance intervals, this crude tolerance interval has no accompanying statement about our percent confidence in the interval width. How does a poorly-estimated 68% tolerance interval (estimated to represent about 68% of the values in the population being sampled) help on in data interpretation?

Likewise, it is fine to include the standard error of mean (SEM) among more helpful summary statistics, but don't report it as an interval. In a perfect world of high sample sizes and perfectly normal distributions, a mean plus or minus the SEM is a 68.26% parametric confidence interval about the mean. Even if the assumptions were met, who cares about 68% confidence about a mean? SEMs about a mean may be over reported partly because they are a smaller value than a 95 or 99% confidence interval about the mean. Often better nonparametric confidence intervals are even wider. Thus a confidence interval formed by the SEM tends to give the uniformed a false impression of that the mean estimate is more precise than it really is.

Don't Fall for Common Misunderstandings of SDs, SEMs, and Variances

The standard error of the mean is sometimes based on de Moivre's equation: where standard error of the mean = [the standard deviation of the sample divided by the square root of sample size]. This equation has been nominated as "the most dangerous equation," because of the extreme length of time over which ignorance of it has caused confusion, the variety of fields that have gone astray and the seriousness of the consequences that such ignorance has caused. Again and again for the last 1000 years, and including very recently, supposed experts have failed to understand that (Wainer, 2007):

The standard deviation does not increase in proportion to the sample. Instead, it is proportional to the square root of the sample size

Variances tend to decrease as sample size increases.

Smaller sample sizes tend to lead to more variation and more extreme variation, while larger sample sizes tend to lead to less variation and more stable rates of variation.

Examples of problems from not understanding this equation include inexact amounts of gold in currency starting as far back as 1150 AD, which was perhaps understandable at that time since de Moivre's equation was not discovered until more recently. In 1150 it was assumed that variability increased proportionally to the number of coins and not to its square root, as de Moivre's equation from 1730 would indicate. However, there are also some very recent examples, including:

- 1) Spending more than a billion dollars in the since the 1990s to break large schools into smaller ones based on the mistaken notion that smaller schools achieve, higher student scores. Smaller schools are simply more variable and thus have a disproportionately high % of schools performing well. When it was finely realized that smaller schools also have a disproportionately high % of schools performing badly and that the differences were

simply due to sample size differences as would be expected, the Gates Foundation announced in 2005 that it was moving away from encouraging large schools to convert into smaller ones.

- 2) Recent maps based on US counties suggested that rural counties had the lowest rates of certain types of cancer. Later it was noticed that rural counties also had the highest rates, the equation could have been used to predict this based on low sample sizes in the rural counties.
- 3) There was no overlap between 2006 Allstate insurance lists for the 10 biggest cities in the US vs. the 10 safest for automobiles vs. the 10 least safe. The 10 safest and 10 least safe were at the extremes partly due to lower population driving variability rates higher.
- 4) Daniel Kahneman's 2002 Nobel Prize in economics was for his studies that showed that humans don't intuitively "know" that smaller hospitals would have greater variability in the proportion of male to female births.
- 5) Why males tend to be more variable than females. Mental traits conveyed on X chromosome will have larger variability among males than females, for females have two X chromosomes whereas males have one X and one Y.

Don't Report The CV Without The Sample Size Or When What You Mean Is The RSD.

The coefficient of variation (CV) is a common summary statistic calculated as the standard deviation divided by the mean. A CV has no units since the numerator and denominator have the same units (Zar, 1984). So a CV might be something like 0.95. The relative standard deviation (RSD = $CV \times 100$). A RSD is in % units, so a CV of 0.95 when expressed as a RSD would be 95%.

Too often investigators say CV when what is meant is RSD. They are not exact synonyms, contrary to what some think. When dealing with percents, to avoid confusion, it is probably safest to simply define the RSD as is the standard deviation divided by the mean, all times 100, and avoid the mention of CVs.

Never report the CV or RSD without including sample size. They won't be very useful for very small sample sizes or if the sample is not fully representative of the population being sampled. Also, remember that the CV and RSD are only optimal for normal distributions, so for our commonly skewed environmental data, consider not emphasizing the meaning of CVs or RSDs at all.

How Are RSDs used in QA/QC?

RSDs are calculated when sample size is above two to estimate QC precision. If only two values are being compared, precisions estimates are instead expressed as relative percent difference (RPDs). For contrast, QC sensitivity estimates are usually reported in terms of AMS, MDLs or MLs) and are given in original units of measure (Irwin, 2008).

What are Some Rules of Thumb Related to CVs and RSDs?

The RSD is sometimes used as a rule of thumb in deciding how to express centrality in transformed, lognormal data. If the underlying distribution is lognormal (and one has a good estimate of the variance of the transformed variables), the sample mean is the preferred estimator of centrality if the RSD is believed to be less than 1.2. Otherwise, a minimum variance unbiased (MVU) is preferred (Gilbert, 1987).

There is also a connection between the RSD and an 'Effect Size.' The RSD is simply the standard deviation divided by the mean times 100. This starts to look a bit like an effect size expressed as a %, which is $(\text{difference}/\text{SD} \times 100)$. Since the units are both (albeit inversely) normalized to standard deviations, there is a relationship between RSDs and effect sizes. A CV/RSD may be used to estimate a minimum detectable difference in original units. The minimum detectable difference (the MDD) = difference between mean observed vs. true mean of reference site in original units = $\text{ES}/100 \times \text{CV} \times \text{the true mean at the reference site}$. For more details, see McBride, 2005).

One sees all kinds of other rules of thumb related to CVs and RSDs, but perhaps too often important assumptions are not checked. In one example:

It is a common practice among agricultural scientists that they usually ignore some strict and sensitive parametric assumptions e.g., homogeneity, normality and no outlier assumptions prior to undertaking the combined experiments. Any misunderstanding of statistics and statistical methods might lead to misleading and invalid inferences. In one example, agricultural scientists use the coefficient of variation (CV) as a measure of population variability. They use the CV to accept or reject the validity of trials. Use of CV is valid only when the simple regression coefficient b value of the regression of the natural log of the environments' MSE on the natural log of the environments' mean equals 2.0...Use of CV may affect the validity of trials due to heterogeneity of MSEs or because of existence of outliers in individual data across heterogeneous environments (Shah et al, 2009).

Are There Nonparametric Alternatives To The RSD and CV?

There appear to be none commonly used for environmental data. One might think of alternatives like the following, but many of these share some of the limitations of the CV and RSD. Several of them still rely on means and/or normal distribution or at least symmetrical patterns in the population being sampled. Sometimes, investigators heroically try too hard to force parametric statistics to work on skewed data. Although one could think of somewhat similar or analogous nonparametric alternatives, why stretch that far? So perhaps one should use none of the following at least partly nonparametric alternatives but instead should rely on more common nonparametric methods to deal with variability:

Instead of using the standard deviation in an RSD calculation, one might consider using the sample-data deviation (F-pseudosigma). This is a nonparametric statistic analogous to the standard deviation that is calculated by using the 25th and 75th percentiles in a data set. It is resistant to the effect of extreme outliers. Specifically, to get

the F-pseudosigma, one subtracts the 25th percentile from the 75th percentile to obtain the inter-quartile range (IQR) magnitude and divides the result by 1.349. Presumably one might then use the F-pseudosigma in place of the standard deviation in the RSD or CV equations. The result could essentially be nonparametric RSD (derived from f-pseudosigma).

An average deviation divided by the mean times 100 would produce an “average deviation percent”, which would be type of nonparametric version of a relative standard deviation (RSD). An average deviation (mean average deviation = MAD) is not used in higher moment statistics like the standard deviation but could be reported as a more intuitively understood stand-alone summary statistic.

It is not usually done, but to get away from over-reliance on means, one might also calculate a median deviation divided by the median times 100 to produce a “median deviation percent”,

One at least partly nonparametric alternative might be to report something along the lines of a dimensionless (percentage) “nonparametric coefficient of variation” (NPCV) hereby defined as the (15.87 percentile minus the 84.13 percentile)/divided by 2/divided by the median, all times 100. The reason behind using a 15.87 percentile at the lower end and the 84.13 percentile at the upper end is to be consistent with the standard deviation interval about the mean (a 68.26% tolerance interval). In cases where data is drawn from a (perfectly) normal distribution, one hundred % minus 68.26% = 31.74%. Dividing that value by two gives one 15.87 percent. Subtracting 15.87% from 100% = 84.13%... However, dividing by two hints at symmetrical pattern, which may not describe the underlying target population being sampled. Also why try to replicate standard deviation based tolerance interval methods when the distribution is not normal?

Don't Read Too Much into Correlation Coefficients

When comparing two environmental variables during early EDA, correlation coefficients are often reported as an initial EDA summary statistics. Remember though that correlation does not equal causation.

The nonparametric Kendall tau correlation coefficient has some advantages over the more ubiquitous parametric Pearson Correlation coefficient. It is more resistant to minor amounts of outliers, missing data, and nondetect data, and estimates non-linear monotonic correlation better. For identical data, it tends to be lower than the Pearson correlation coefficient, so a relatively strong Pearson correlation of 0.9 might be more like 0.7 for a Kendall tau correlation coefficient (for more detail see Internet discussion of Helsel and Hirsch, 2002).

Include Single Data Point Uncertainty

During early EDA, it is appropriate to compare uncertainty in each data point to make sure it is [trivial](#) compared to variability in measures from the open environment. Single data point

uncertainty can be estimated as an EDA and QC summary statistic called alternative measurement sensitivity (AMS). AMS is basically a special case of NIST/ISO/APHA measurement uncertainty. To make it functionally analogous to a method detection limit (MDL) but two-sided, it is based on sample size seven and 99% confidence based on sample size seven and 99% confidence. Two-sided AMS sensitivity values are logical for uncertainty when (unlike detection limits) we are interested more in how different two measures would have to be to be truly different from each other and we are less interested in the detection limit issue of different from zero (Irwin, 2008, Thoma et al. 2011).

Single measurement uncertainty simply adds to overall variability estimates and makes confidence intervals (and standard deviations, and tolerance intervals, and prediction intervals) bigger. The questions to be answered are 1) whether or not it adds enough to signify a QC problem with the measuring system and 2) whether or not it adds non trivial amounts of variability. How does one determine trivial amounts of variation? There are various complex ways to decide if it is trivial, but one simplified rule of thumb that has been used in Europe is that if one standard deviation is five times larger than another, then the larger one becomes dominant and the smaller one becomes (relatively) insignificant (Irwin, 2004).

Level One EDA Plots:

Rather than relying too much on a very small set of summary statistics, in EDA one should look for visual clues about patterns in the data first, since the greatest value of a picture is when it forces us to notice what we never expected to see (Tukey, 1977; Hartwig and Dearing, 1979). Always look for visual clues about patterns in the data and hand before employing formal procedures (McBride, 2005). Since environmental data is usually not normally distributed, the nonparametric versions of EDA explained by Tukey (1977) and summarized on the Internet by Statgraphics (2009a) are of particular value.

Very Common Simple EDA Plots:

Non parametric box and whisker plots are often more useful than histograms (Ellison, 1993).

Other commonly used simple plots EDA plots include scatter plots, time series plots, line plots, and probability plots (including normal probability plots). For examples of all these plots see the Internet summaries of EPA (2009) and NIST (2006).

One gets hints from these early plots not only about whether relationships are linear or non linear, but also about stationarity. Is there a hint in early plots of a trend or not?). Like decisions about normality, decisions about stationarity and linearity are important in deciding what statistical analyses should come next.

A common first step is simply visually look at the pattern in a line plot or X/Y scatterplot or line plot or smoothed scatterplot to see what pattern it seems to resemble most closely. One can choose between the example patterns in the NIST (2006) summary at <http://itl.nist.gov/div898/handbook/eda/section3/eda33q.htm>. At that website, one can click on

each of the following options, and then simply visually inspect the pattern at hand to see which of the following patterns your plotted pattern seems to most closely resemble:

1. [No relationship](#)
2. [Strong linear \(positive correlation\)](#)
3. [Strong linear \(negative correlation\)](#)
4. [Exact linear \(positive correlation\)](#)
5. [Quadratic relationship](#)
6. [Exponential relationship](#)
7. [Sinusoidal relationship \(damped\)](#)
8. [Variation of Y doesn't depend on X \(homoscedastic\)](#)
9. [Variation of Y does depend on X \(heteroscedastic\)](#)
10. [Outlier](#)

NIST (2006) also provided similar Internet example plots for histogram shapes at <http://www.itl.nist.gov/div898/handbook/eda/section3/eda33e.htm>. Here are the basics:

Histogram Interpretation: Normal. In this case it shows a single peak classical bell-shaped curve, see example at <http://itl.nist.gov/div898/handbook/eda/section3/eda3661.htm>). However, remember that some plots of some normal distributions curves are flatter than others (Zar, 1984; Helsel and Hirsch, 2020).

Histogram Interpretation: Symmetric, Non-Normal, Short-Tailed (shows most of the data in the middle with a short-tailed distribution so that the tails approach zero very fast—little in the way of tails on either side).

Histogram Interpretation: Symmetric, Non-Normal, Long-Tailed (looks a bit like a normal bell shape but the middle is skinnier and the tails on either side longer).

Histogram Interpretation: Symmetric and Bimodal (shows two peaks)

Histogram Interpretation: Bimodal Mixture of 2 Normals (shows two symmetrical peaks, each of which has the classic bell shape)

Histogram Interpretation: Skewed (Non-Normal) Right (long tail to the right, not much tail and few data points on the left)

Histogram Interpretation: Skewed (Non-Symmetric) Left (long tail to the left, not much tail and few data points on the right)

Histogram Interpretation: Symmetric with Outlier (symmetric with a bar or two by themselves separate and off to the right or left).

Less Common Simple EDA Plots:

These include: Stem and Leaf plots, Quantile (Q-Q) Plots, Ranked Data Plots, regression techniques to plot one variable versus another, Empirical Quantile-Quantile Plots for Two Variables, Index plots, scatterplot matrices, X-Y conditioning plots, strip charts, plots of squared residuals, and plots data that handle nondetects in acceptable ways. For examples of all these plots see the Internet summaries of EPA (2009) and NIST (2006).

The stem-and-leaf plot is really a histogram or bar graph turned on its side, with the added feature of displaying data values directly. Although box and whisker plots often tend to be among the most helpful and most common early EDA plots, there are many different versions of box-and-whisker plots and for that reason it is wise to always state conventions used (25th to 75th percentiles for the box and 1.5 times the IQR for the whisker lines or something else?) to calculate them and to designate outliers (McBride, 2005).

Less often, a histogram is used to portray a normal distribution probability density function (PDF).

There is Always a Model:

Even if the model is not stated, almost always some model is assumed. The model may be as simple as assuming the data is from a normal distribution. In recent times, nonparametric statistics are more often used in water quality than parametric statistics (McBride, 2005). Those who immediately transform variables to try to force them into normality are assuming things like lognormal distributions. However, there are numerous back transformation and other pitfalls that are often not fully considered.

Software Options for Simple EDA Plots:

Virtually any statistical software (even some common spreadsheets) can be used for simple level one EDA. In R software some early EDA steps are often done in the R Commander GUI.

ProUCL 4.0 free software developed for EPA can properly draw [Q-Q plots](#) and box plots for data sets with and without non-detect observations, as well as upper and lower tolerance limits and prediction limits (Singh et al., 2010).

Two software packages were designed specially to ease EDA tasks: Data Desk and Data Desk/XL, an add-in for MS Excel (Data Description, 2011). One dedicated statistical software package, Statgraphics, is designed to make simple graphing tasks easy and specifically does many of the EDA methods suggested by Tukey (1977), including nonlinear smooths based on running medians (Statgraphics, 2009a).

The wq package in R software draws side-by-side box and whisker plots in a time series and also will draw time series in the form of a line plot or an anomaly plot (Jassby and Cloern, 2010).

Others use downloading and organizing and Quality Assurance checking software for some simple EDA tasks. Here are a couple of examples from the NPA Vital Signs Monitoring Program:

In the Southwest Alaska Network, we use AQUARIUS along with SigmaPlot software for EDA. We first usually draw time plots (scatter plots of parameter vs. time) in Aquarius since the data are already there after using AQUARIUS to import it from sondes. We simply look at the pattern and see if anything looks unusual or out-of-place (spikes, potential [outliers](#), etc). For comparisons with water quality standards or other resource protection thresholds, we use AQUARIUS in a two step process. In the first step, we flag the values based on given criteria using the Threshold Flagging toolbox. In the second step we use the Math and Stats toolbox to report the number of data points exceeding the threshold value. When using the Threshold Flagging toolbox, there may be multiple steps depending on how many parameters one is screening. One could probably use that same Toolbox as another way to help identify outliers, but we don't use it that way. Instead we first look for outliers in simple plots. Most of our EDA graphing is done in SigmaPlot, which easily plots Box and whisker plots and other simple plots. We look for outliers using box and whisker plots and also look for potential outlier values outside of the 95th/5th or 90th/10th percentiles SigmaPlot is a STSTAT statistics software product that is user friendly and menu driven that makes graphing easy. We use it to produce all graphs for reports and presentations. We find its graphing capabilities to be vastly superior to MS Excel (Jeff Shearer, NPS, Personal Communication, 2011).

In the Upper Columbia Basin Network, we find that AQUARIUS software makes it easier to process time-series data from continuously monitoring sondes. We also use AQUARIUS to calculate the following summary statistics and related data: Start Date, End Date, Data Points, Minimum, Maximum, Sum, Mean, Median, Range, Standard Deviation, Kurtosis, Skewness, Inter Quartile Distance, Points Above [defined thresholds--we usually use state DEQ TMDLs], Average Above, Duration above threshold, Points Below Threshold, Average Below Threshold, Duration Below Threshold. Our threshold and rate of change flagging is done on a parameter by parameter basis using the same process that Jeff described above. We flag outliers for temperature, specific conductance, dissolved oxygen, pH, and turbidity, based on the hourly rate of change. More substantial QA is best done in an external statistical program. If the rate of change exceeds general thresholds it is flagged for further review. There are a number of factors that go into determining if flagged data is "real and actually important." Here are a few issues considered (Eric Starkey, NPS, Personal Communication, 2011):

- 1) What does the data look like on either side of the data point? Is there a steady increase or decrease over the previous/following hour(s)?
- 2) Does other data suggest this is a real value? We look at discharge data, other parameters, weather stations, etc.
- 3) Are there other factors that may have contributed to the outlier? We look at upstream irrigation/discharge, construction, restoration, interpretive programs in the stream, etc.
- 4) Are there hints of fouling or has there been wiper parking or have Sensor limits been exceeded?

Smoothing Basics:

In the remainder of this document, the reader will see numerous mentions of smoothed trend lines or other smoothing concepts. Therefore, we will explain some smoothing basics here before going farther.

Adding a smoothed line through the middle of the points in a busy scatter plot draws ones attention to the smoothed line and makes it easier to visualize trends and cycles. If plotted against a variable other than time, smoothing helps one visualize patterns in the relationship of the one variable to another variable (Helsel and Hirsch, 2002).

Having both regular X/Y plots of individual data points (the rough structure) and smoothed plots (smoothed structure of the underlying simplified structure) helps us understand both short and longer term patterns in the data more completely (Hartwig and Dearing, 1979). Both patterns are often important, so don't focus on only one. In other words, if one is going to add a smoothed line to a scatter plot, keep the data points too, and don't transform the data points to smoothed points on the line before doing hypotheses tests or other advanced statistics. Don't discard the original data. Smoothing done wrong can be a big problem and tends to be a point of contention in climate change debates (see discussion of potential misuse of smoothing at the end of this section for more detail). On the other hand, EDA smoothing done right (which includes the right reasons and transparency, and optimal methods for the particular issue at hand) can be very helpful, especially in the context of long term monitoring and avoiding bad practices (such as trying to use a strictly linear method as the only line of evidence on a non-linear pattern).

Often smoothing is done early on to find a pattern such as a trend. Thus many think of it as an early step, to be followed up later with a more complex trend test or other complex model-based analyses.

There are times when smoothing is a better predictive method than some of the very complex models such as an Autoregressive Integrated Moving Average Model (ARIMA). For example, one has less than 39 data points or if the data is highly volatile, then a smoothing method may perform better than an ARIMA model even for the relatively difficult task of predicting what will happen next (Morrison, 2011).

A good starting point for a span (scaling) factor is 0.75. One would usually not start with a span factor of less than 0.5. However, to see both the rough and smooth patterns, one might use a lower value (say 0.2) for the rough span and 0.75 for the smooth span.

THE ALPHABET SOUP OF SMOOTHERS:

SMA = Simple Moving Average (for normal data)

EMA = Exponential Moving Average (for more variable data)

MM = Moving Median (for most data)

RM = Running Median = MM

LOWESS = Locally Weighted Scatterplot Smoothing (one predictor variable, no inference statistic).

LOESS = Locally estimated scatterplot smoothing (This can handle one or more predictor variable). An inference statistic is calculated.

GAM. = Generalized Additive Model smoothing

Google Hits on Various Kinds of Smoothers in Water Quality Work, February 3, 2011: This comparison is admittedly a crude estimate of popularity of various ways to smooth trend lines or messy time-trend scatter plots in water quality, and some of these hits (especially for moving averages) relate to applications other than smoothing trend lines. Some related to water quality standards more than to trends, but the following is still of interest:

Googling "moving average", "water quality", "trend line" returned 2530 results.

Googling "LOWESS", "water quality", "trend line" returned 399 results

Googling "LOESS", "water quality", "trend line" returned 766 results.

Googling "moving median", "water quality", "trend line" returned 33 results.

Googling "exponential moving average", "water quality", "trend line" returned 14 results

Googling "GAM", "water quality", "trend line" returned 832 results, though not all were about Generalized Additive Model smoothing.

Googling "Generalized Additive Model", "water quality", "trend line" returned 34 results.

Just for a simple comparison moving averages also play a big role in stock market settings, but exponential moving averages (which tend to downplay the effects of [outliers](#)) seem more popular than for water quality:

Googling "moving average", "stock market", "trend line" returned 491,000 results.

Googling "LOWESS", "stock market", "trend line" returned 74 results

Googling "LOESS", "stock market", "trend line" returned 586 results.

Googling "moving median", "stock market", "trend line" returned 36 results.

Googling "exponential moving average", "stock market", "trend line" returned 21,500 results

Yahoo finance and other finance software often provide an easy way to add a SMA and EMA smoothed trend lines to stock prices over time.

For many users, the software on hand is probably thought of as a black box doing something to produce a middle-smooth. However, one still needs to thoughtfully choose the scaling factors. Also keep in mind that the lines drawn by different software packages, even under a single name (say LOESS) will not always look exactly the same.

Cautions Related To Smoothing

Don't Just Force It To A Straight Line.

Avoid the too common practice of simply using ordinary linear regression to fit a line on a pattern that is not clearly linear. Doing this does not insure a line that best fits the middle of the data pattern (Garson, 2009; Helsel and Hirsch, 2002). Most of the other smoothing techniques discussed herein, including moving median smooths, LOWESS smooths, LOESS smooths and “best fit” line plotting by Excel or of the software will do a better job of fitting a trend line closer to the middle of the data in a relationship that is not clearly linear.

Don't Ignore Assumptions or Discard Raw Data.

Kittel (2009) explained smoothing filters using simple and weighted 3 and 5 year moving averages used in climate data and provided the following cautions about smoothing with moving averages:

Though smoothing has some uses in careful and complex modeling, smoothing should not be used like transformations prior to standard statistical analysis such as hypothesis testing. Instead, think of smoothing as an exploratory tool, to aid visualizing temporal patterns. After smoothing, the resulting series has severely reduced variance and elevated serial correlation and so should not be used in statistical testing. Smoothing with 3 and 5 year moving averages also should not be done in spectral analysis due to the introduction of side lobes and any end-effect corrections. Keep these caveats in mind when obtaining processed data series from other sources, such as teleconnection indices. To this end, check documentation to see if processing included smoothing and, if so, check for the availability of unsmoothed versions.

Smoothing climate data may become so routine for monitoring mid- and long-term behavior that we miss or discount unusual, yet critically important events, thus possibly missing important details. Smoothed datasets may make it easier to miss tracking the effects of important but unusual changes related to single events, such as volcanic eruptions, in regional and local climate records.

One should avoid using non-centered (e.g., prior-moving) averages such as those found in the moving average option in MS Excel's 'trendline' graph function, since this can offset peaks and troughs with respect to actual timing. On the other hand, MS Excel (and many other software packages) also includes nonlinear line fitting routines that can aid in exploring possible transformations. These are also helpful in identifying polynomial patterns (quadratic or other) in curved but still monotonic patterns.

The first two cautions given by Kittel above would be also be relevant to moving medians and to more complex LOWESS or LOESS smoothing.

Don't Throw Out the Baby with the Bath Oil

Climate change is a controversial (minefield) topic and some involved with climate change statistics and debates have put anti-smoothing blogs on the internet (for one example, see <http://wmbriggs.com/blog/?p=735>). Some basically say **never to smooth data**, but looking

closer many of points seem to say don't do smoothing wrong. The cautions seem to emphasize some of the points already made above, such as

Don't transform data with a smoother and then use the new and less variable and now more highly correlated data in subsequent tests and modeling and correlation analyses,

Never throw away the original data and always look at plots of it too since, the original data is the true data, not the "fictional" smoothed data,

Don't throw away true outliers when they are important to the question to be answered (as they often are).

Don't just choose options based on the goal of trying to help support a pet theory. If one chooses higher and higher "span" values for scale, even for a LOESS smooth, one will get a straighter and straighter line that resembles the original data less and less.

Don't use smoothers to overhype climate change certainty or climate forecasts.

These are all potentially important points for typical climate change debates. However, perhaps the take home message **should be don't do smoothing the wrong way and for the wrong reasons** more than don't ever do smoothing. Some of those adding comments to these blogs also point out that although one should use smoothing with some caution, smoothing still has important value in picking out potentially important patterns in EDA and finding important signals. This brings to mind our earlier discussions of avoiding bad practices and the caution of Anderson and Burnham (2002) that any methodology can be misused and that potential misuse of methods is not a reason to never use them in any situation. The criticisms aimed at smoothing in these climate-change-related debates tend to be mostly aimed at final data analyses and forecasting based on using smoothing adjusted data rather than the original data. Many of the points are important though.

For example, using smoothing only at certain span width scale factors to help prove a preconceived pet theory is simply an unethical/bad practice. It is analogous to data dredging, where one might keep transforming data using different transforms until a pre-conceived answer supporting a pet theory finally appears. This should be thought of as a non-transparent and unethical bad practice to be avoided rather than an all inclusive reason never to use smoothing for more legitimate and defensible EDA purposes.

To thoughtfully make sense of one's data, one needs to decide if true-value [outliers](#) logically need to be de-emphasized or emphasized to answer a particular question at hand. One would not say medians should never be calculated because they de-emphasize outliers, and one would not say that means should never be used because the influence of outliers on means is greater than for medians. In the same way one would not say smoothed trend lines should never be plotted because some might misuse them. Instead, use them carefully and correctly for logical and valid EDA purposes.

Stock market technical analysts have not been so foolish as to throw out smoothing just because it can be misused or misinterpreted, and they have a keener interest in correctly identifying

patterns and likely trend reversals than many of us, because they have large sums of money at stake.

Simply avoid using smoothers in non-transparent, unethical, or bad practice manners, no matter what kind of smoother is used

However, do keep in mind that any smoother can be a bit arbitrary or misleading depending on the choices one makes for weighting and/or scale (span) factors. Smoothing also produces a loss of information on outliers (Ellison, 1993).

Should One Never Test for Trends in Seasonal Medians?

As discussed above, other things being equal, it is usually better not to test for trends using smoothed data.

Wait, don't some Kendall methods essentially test for trends in smoothed data?
And isn't this an example of the kind of thing frowned upon in the climate debates?

Sort of, but this can be an exception where the pros can sometimes outweigh the cons, making it basically OK to do it this way. Here are some things to consider.

One way to minimize serial correlation issues without resorting to very complex time series analyses would be to group the data into time periods (e.g., weeks, months) and compute a summary statistic for the period then use these summary statistics in regression-based trend methods (Helsel and Hirsch, 2002).

At first glance, the popular seasonal Kendall (SK) trend test may seem somewhat analogous to doing a test on smoothed data, at least if it is implemented by testing for trends on seasonal medians. However, that is not how all software options handle it. Just as not all software do the same thing for what they call LOWESS or LOESS smooths, not all software packages give the same result for something they call a SK test. Consider the following four ways to perform a seasonal Kendall test. The first two involve less smoothing and/or potential loss of information than the last two, but all four might be considered valid depending on the data set and important common sense factors. In other words, there are cases where smoothing before testing (as in option 4 below) is both commonly done and is considered valid.

- 1) Option one: There is only one observation per period, so that is all that is used. The one measurement may happen to be on a day that is unusual, and there are sample size concerns and it may be impossible to find any trend unless a great many years have passed. This is an option that can be specified in Time Trends software for the SK test (Jowett, 2009).
- 2) Option two: There are multiple values per season, and the SK code uses all the values, not just the median. SK is a 'blocking' procedure in statistics, of which there are many, including the Friedman test and others. It 'blocks out' seasonal variation by only conducting a test on data within the same season. The SK code from Practical Stats computes the test on the original data, not smoothed in any way. As with any SK version, it computes several tests, one on data in each season (block), and then combines the

several test results into one overall test. No smoothing of data is involved. Doing that is preferable in my opinion to selecting only one value to represent a year and season. This does mean that if one year has many more observations than another within a season, the results are weighted towards that year. So I understand why some code strips off some of the observations and uses either the median or the middle value of the season [as is the case in the other options], but strictly speaking there isn't a requirement to do so with the SK test (Dennis Helsel, Practical Stats Inc., Personal Communication, 2011).

- 3) Option three: The software uses the middle value of the season, not all of the values in the season. This is what S-ESTREND software does to implement the SK test. If the season is July, S-ESTREND will find the value closest to July 15th and use that. Using the single middle value would bring up a potential outlier issue to consider. However, there is always a chance that the one value from July 15 may be an outlier (say influenced by an unusual flood event) and not that representative of July in general. In that case, using the median as per the example in TimeTrends software (next option) would tend to deemphasize the possible influence of an outlier that just happened to occur on July 15.
- 4) Option four. The program uses the median of the season when computing the SK test. Time Trends software gives users the option of using this approach (Jowett, 2009).

Either using the “single middle value” or a summary statistic for a season is simply designed to get a value considered generally representative of that season, but using a single value comes at the potential disadvantage of information loss that can impact statistical power (van Belle and Hughes, 1984).

The potential loss of information is partly what climate change critics of smoothing data and then using the smoothed data rather than the original data in subsequent tests complain about. Using seasonal medians (or even single middle values (instead of all the values in a season) would be at least somewhat analogous to using smoothed data for the final trend. Those who say one should never use smoothed data in statistical tests would probably be happiest with option two above.

However, the seasonal median probably has a better chance to be representative of the season than a single value in the middle of the season. In fact, the seasonal median may be most representative way to test for changes in typical conditions over a large number of years, and representativeness is a very important but often not fully considered concern.

Consider the case where one is looking for trends only in seasons with no missing data. If there were logical reasons to do so (say to de-emphasize some unusual [outliers](#) that might otherwise obscure the underlying trend pattern), one might choose to look at trends in yearly medians from the July to August time periods. If the medians provide the most representative values typical for the July to August time period, it may be running a Mann Kendall test for trends based on the medians even at the risk that some might complain that in this example, one is basically testing trends on smoothed data rather than raw data. In a sense that is true, but if the median is most representative of typical conditions and the monitoring question to be answered is : How are typical conditions changing"?, a case could be made for the medians being logical and also having the advantage of lessening the risk from problems associated with autocorrelation.

In Time Trends software (Jowett, 2009), the Mann-Kendall test is a menu item, but one can use the seasonal Kendall test to get a Mann Kendall analysis. If one limits the data to one season,

either by specifying that only that season is to be used in the season specification or by using the filter, the seasonal Kendall test is the Mann Kendall and the heading on the output is Mann Kendall even though one selects the Seasonal Kendall menu (Ian Jowett, NIWA, NZ, Personal Communication, 2011).

More Detail on The common Smoothing Techniques:

Moving Median Smooths

Although moving medians don't seem to be used as much as LOWESS and LOESS and for water quality data, the EDA textbooks of Tukey (1977) and Hartwig and Dearing (1979) both suggested moving or running medians as a good starting point for initial smoothing. If the software package on hand has any easy way to do that, this is typically a good choice. To solve "end value" issues with running medians, make the first and last values the same as the data points in the series, make the first running median the median of the next three values and make the last median the median of the last three values before the end value.

Another even simpler first step no matter what software is at hand is to line up box and whisker plots in a time series, then simply connect the median lines to get a rough moving median smooth, and one can easily do this without sophisticated software.

Tukey's resistant nonlinear smoothers (such as a running or moving median) are very useful for displaying the trend in noisy time series data. In the Statgraphics Time Series Smoothing procedure, the smoothers are often used as preprocessors before application of a weighted moving average (Statgraphics, 2009a).

Simple Moving Average (SMA) Smooths

Moving averages are also sometimes plotted with environmental data, and 3 and 5 year moving averages used in that manner (along with some cautions) were discussed in detail by Kittel (2009).

For many environmental datasets, I tend to suggest avoiding simple moving averages in favor of LOESS because of the problems with averaging. Moving averages are inherently strongly influenced by outliers. Moving averages are often a series of averages of say 5 points within a window. One outlier and the average can be higher than the other 4 points. So a moving average is only useful in well-behaved data devoid of outliers that are true values rather than mistakes. Environmental data are rarely that well-behaved (Dennis Helsel, Practical Stats Inc., Personal Communication, 2011).

Not much space seems to be devoted to Moving Average Smoothers in water quality statistics text books, moving average smoothers and most moving average statistics are used more than some realize. One potential reason for using them in early exploratory data analysis (EDA) steps is that later in level three EDA and subsequent complex data analysis steps, one may attempt to

see whether or not different types of autoregressive (AR) models or moving average (MA) models or combination models (ARMIA) fit the data best (McBride, 2005).

Exponential Moving Average (EMA) Smooths

The issue with [outliers](#) is one reason that folks sometimes use exponential moving average smooths instead of simple moving average smooths. The exponential version tends to be preferred with parameters that show very high variability and/or when one wants to de-emphasize the effects of outliers. Though simple, one still has to pick the scale (span) factors thoughtfully (a 360 day exponential moving average produces a smoother trend line than a 10 day exponential moving average. That is one reason that in the stock market the crossing points of the 50 day and the 200 day moving averages are inspected as possible change points in an earlier up or down trend, whether one is considering the smoother exponential moving averages or the slightly less smooth simple moving averages.

LOWESS and LOESS Smoothers:

LOWESS and LOESS are basically nonlinear regression methods that work well on nonlinear patterns and thus are sometimes used to get an early hint of whether or not there is a non-linear trend in variable plotted vs. time or a possible other non linear pattern between two variables plotted against each other (say flow versus turbidity in water quality).

In developing LOWESS Cleveland (1981) called it a Robust Locally Weighted Regression method that could be used with one predictor variable. In developing LOESS 10 years later, Cleveland and Grosse (1991) provided a related method that could be used with more than one predictor variable.

Both of these, but particularly LOESS, are now more commonly used than before the development of powerful personal computers. Both LOWESS and LOESS are thought of as nonparametric regression methods. Unlike standard linear regression analysis, no assumption is made that the relationship is represented by a straight line

The LOWESS algorithm estimates a curve that represents the main trend in the data, without assuming a specific mathematical relationship between Y and X. This is why LOWESS is called nonparametric. For details, see Helsel and Hirsch (2002). LOWESS smoothers are locally weighted (the LOW in LOWESS).

Although a LOESS might be a better match for a quadratic trend line, LOESS is not even mentioned in some prominent water quality statistics text books (Helsel and Hirsch, 2002; McBride, 2005).

LOWESS and LOESS Are Not Exactly the Same

LOWESS and LOESS are fairly commonly used as virtual synonyms (NIST, 2006; Wikipedia, 2010) and some even say LOWESS is an older term that has been replaced by LOESS.

These are over-simplifications. As originally defined, they are not the same. Efforts to contrast the differences between LOWESS and LOESS include the following:

W.S. Cleveland, who was key player in the origination of both LOESS and LOWESS methods, gives a technically more precise distinction between the two on a personal website (<http://stat.bell-labs.com/wsc/smoothsoft.html>):

LOWESS smooths tend to be a function of one predictor variable while LOESS tends to be a function of one or more predictor variables. LOWESS smooths are for data with normal errors or for data with long-tailed symmetric errors (robust fitting). Statistics for inference are not computed. LOWESS is based on different code than LOESS. Code for LOWESS was first written in the 1970s in FORTRAN. In modern statistical software, S-Plus, R, Systat, XploRE, Gauss, and SAS have interfaces to LOWESS (Cleveland, 1981).

LOESS smooths carry out local fitting for normal and long-tailed distributions and unlike LOWESS can either use one or **more than one** predictor variable and unlike LOWESS it also computes the a statistic for inferences. LOESS is based on different code and was first written in C and FORTRAN subroutines and first implemented in modern statistical software in S-Plus (Cleveland and Grosse, 1991).

In more plain language, LOWESS is a one dimensional smoother, while LOESS develops multi-dimensional models. LOESS has a number of options for model development and to get agreement with LOWESS one must select the appropriate options. A one dimensional model with LOESS should produce a smoothed trend line that is the same as LOWESS but I find minor differences. Both LOWESS and LOESS seem to be equally sensitive to outliers (Jowett, 2009), confirmed by Ian Jowett, NIWA, NZ, Personal Communication, 2011:

For water quality time series of a single variable versus, simpler LOWESS smooths tend to be used with only one predictor variable (time), but a moving median might be a more intuitive starting point for beginners.

Matlab statistical software agrees that LOWESS is for one explanatory variable and LOESS for one or more, but explains it in different terms using the term weighted for both (Mathworks, 2010):

LOWESS = Local regression using weighted linear least squares and a 1st degree polynomial model. .

LOESS = Local regression using weighted linear least squares and a 2nd degree polynomial models.

Note: Many would associate a 1st degree polynomial with a linear pattern, a 2nd degree polynomial with a quadratic pattern. Even MS Excel has a simple line

fitting routine for both of these, but it is unclear if that would draw the same lines as Matlab would for LOWESS (1st degree) and LOESS (2nd degree).

In R software documentation, LOWESS and LOESS are contrasted as follows:

In R software, the function `lowess {stats}` is used to perform the computations for the LOWESS smoother which uses locally-weighted polynomial regression. The span is given as `f` (a fraction) where 2/3 (0.66) was given as a typical starting point. LOWESS is defined by a complex algorithm by W. S. Cleveland. Normally a local linear polynomial fit is used, but under some circumstances (see the file) a local constant fit can be used. See details from the R manual at

<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lowess.html>.

In R software, the function `loess {stats}` is used to fit a polynomial surface determined by one or more numerical predictors, using local fitting. A span of 0.75 is recommended as a starting point. See details from the R manual at

<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/loess.html>.

However, in one R software discussion, LOESS and LOWESS were wrongly treated as synonyms. Span values of 0.75 seemed to work well in trials for LOESS, confirming the generic recommendation in R. For details, see

<http://research.stowers-institute.org/efg/R/Statistics/loess.htm>

In the latest version of Statgraphics Centurion software, the LOWESS smoother lets the user interactively change the width of the smoothing window (the span) and immediately see the result. This is said to be a good way to visualize the trend-cycle (Craig Polhemus, StatPoint Technologies, Inc., Personal Communication, 2011).

LOWESS and LOESS Software

LOWESS and LOESS tend to be used for similar purposes and I suspect that many non-experts simply use whichever one their software at hand provides as an easy default. However, not all LOWESS and LOESS smoothed lines will look identical.

In one comparison, The S-PLUS software linear-option for LOESS and MINITAB LOWESS produced identical slightly curved trend line plots for a typical total phosphorus water quality data set. For contrast, using the same data set, the S-PLUS LOESS option based on a quadratic polynomial (rather than simple linear regression) produced a vaguely similar trend line but one with much more pronounced curves. (Dilip Shinde, National Park Service, Personal Communication, 2010).

Since 1991, more software packages now include LOWESS and LOESS routines, but not all seem to use the exact original Cleveland code. S-plus methods, however, are based on the original Cleveland code (see Cleveland discussion above).

More Complex Smoothers

Though LOWESS and LOESS are commonly used, there are many other complex smoothers that could be used. For example, the use of both Holt's exponential smoothing and Holt-Winters exponential smoothing in R software was summarized by Coghlan (2011a) in her discussion of time series tasks.

In another example, Time Trends freeware gives users the option of smoothing trend lines with either Generalized Additive Model (GAM) or with simpler smoothers (Jowett, 2009), confirmed by Ian Jowett, NIWA, NZ, Personal Communication, 2011.

Nonparametric GAM approaches are relatively new, having been used since 1990. Like LOESS, GAM smoothers may be more appropriate than some simpler smoothers for complex multidimensional models.

GAM smoothers and GAM trend and predictive models tend to be too complex to explain fully herein and may be too complex for most of our beginner target audience. However, for those who want to learn more, Sanger et al. (2002) have an lengthy example on the Internet, of using GAM methods to model a water quality process with not only regular diel cycles but also with complex tidal cycles. Among their conclusions were the following statements:

Classical harmonic regression analyses of water level (depth) typically fit the data remarkably well and provide near-perfect predictions. Occasionally, shallow-water settings create complexities in tidal signatures that are not well modeled by a weighted sum of sine waves using the classical constituents. Moreover, "noisier" variables like dissolved oxygen and salinity require a much more flexible approach than the classical sine-wave-based models, and there is certainly no reason to believe that the very important diel signature will be well approximated by sine waves for any variable. These considerations, and an unusual abundance of data, led us to use GAM model approach. The GAM model fits the noisier variables much better than a classical sine-based model, and is very competitive for modeling depth. Despite its complexity, it lends itself very well to graphical depiction, which is key for interpretation.

TYPICAL QUESTIONS TO BE ANSWERED WITH LEVEL ONE EDA TOOLS:

Is the Distribution Normal?

This perhaps should not be the first question but it often is because statistical training in the past has tended to focus heavily on parametric methods best used on normally distributed data. There is also a tendency to try too hard to get parametric methods to work when really nonparametric methods would be superior and in fact have more power when the distribution is not normal. Tests for normality have quite low power with small (less than 30) sample sizes (Helsel and Hirsch, 2002).

So, in early EDA, one often looks for hints that distribution being sampled is either a normal distribution or something close. Normally distributed data is unusual for environmental data, but if sample size is adequate and the sample covers the full range of conditions of the underlying

population, early hints that it might be include the following. All of these work better at higher sample sizes, but these may be all one has if sample size is too small (<30) to allow for one to use tests for normality.

The mean, median, and mode are very close to each other in magnitude.

The standard deviation is much smaller than the mean.

Both skewness and kurtosis are close to zero or at least between -0.1 and +0.1 for larger sample sizes. The farther that skewness or kurtosis are from zero, the more likely the distribution is not normal (NIST, 2006).

In the case of the standard error of kurtosis (standardized kurtosis, or SEK) as calculated in Statgraphics Software, Data that come from a normal or Gaussian distribution should yield standardized skewness and standardized kurtosis values between -2 and +2 (Craig Polhemus, Statgraphics, Personal Communication, 2011):

A standard normal probability density function plot of the data shows a relatively symmetrical distribution on either side of the mean, and the tails (extremes) get smaller as they move out from the mean. In other words, a plot of the data suggests the classic bell shape curve or something close (see NIST 2006 example at <http://itl.nist.gov/div898/handbook/eda/section3/eda3661.htm>).

.A normal probability plot reveals a (primarily) straight line hint of a normal distribution? See NIST example plot at <http://www.itl.nist.gov/div898/handbook/eda/section3/eda331.htm>)?

The relative standard deviation (RSD) is often smaller than 50%. If the RSD is greater than 100% it is a relatively strong hint of non normality (Doug McLaughlin, Western Michigan University, Personal Communication, 2011).

.None of the above should be used as stand-alone strong evidence. Use them along with other hints above and as time goes along bring in other lines of evidence such as [Q-Q plots](#).

What is a Probability Density Function (PDF)?

The acronym PDF in this context has nothing to do with adobe acrobat files. PDF plots and also should not be confused with distribution function ([CDF](#)) line plots, though distribution patterns are often recognized either through a PDF pattern or a CDF pattern, as illustrated by the NIST (2006) gallery of common distributions at <http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm>.

For a continuous function, the probability density function (pdf) is the probability that the variate has the value x. Since for continuous distributions the probability at a single point is zero, this is often expressed in terms of an integral between two points (NIST, 2006). In less technical terms, a PDF is a plot that helps us determine the shape the distribution and provides users with a

graphical representation that a random variable may be observed within a specific sub-set of the space.

At very early stages of EDA one often uses simple probability histograms based on raw data instead of PDF plots to as a first look at the shape of the distribution (EPA, 2009). One can also use simple probability density plots using raw data. If the distribution is normal or close to it these plots will at least look symmetrical, though not always exactly like the classic bell shape curve, since some plots will be flatter than others (Helsel and Hirsch, 2002; Zar, 1984).

However, to look closer and to more fully understand the next section, one needs to understand curved line probability plots and especially those that are produced with probability density functions. For an internet discussion of probability plots in general and how to screen with them see EPA (2009) and NIST (2006).

Probability distributions are typically defined in terms of the probability density function. However, be aware there are a number of probability functions used in various applications (NIST, 2006).

In a PDF, the area under the curve tends to sum to one. Like a normal distribution line plot, one can glance at the plot and immediately see that the values in the middle of the distribution are more likely to occur than the values at the extremes (tails of a normal distribution).

Other than that, probability density functions are somewhat complex to fully explain to beginners but are discussed and illustrated in detail by McBride (2005) and NIST (2006).

What are Cumulative Distribution Function (CDF) Line Plots?

CDF plots are also used to look for hints of normality. Sections 1.3.6.2 and 1.3.6.7 of NIST (2006) explain cumulative distribution function (CDF) line plots. A CDF is the probability that the variable takes a value less than or equal to x . A CDF tends to be less frequently used than simple scatterplots and histograms or even quantile plots in early EDA and also tend to be less easily understood by beginners. Unlike a probability density function (PDF) histogram, the vertical axis in a CDF line plot can DIRECTLY represent the probability that a certain value is exceeded.

However, distribution patterns are often recognized either through a PDF pattern or a CDF pattern. NIST (2006) provided typical example of both for each common distribution in its gallery of common distributions at <http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm>.

If not Normal, What is the distribution?

For a single variable, just compare the shapes in your early single variable probability density function plots versus the NIST gallery of shapes on the Internet at <http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm>.

The following distributions are common in water quality work (McBride, 2005):

The normal distribution is used for symmetric continuous data, and also is used in the form of the unit normal distribution as the distribution of a test statistic. (The unit normal distribution has zero mean and unit variance).

The lognormal distribution is used for right-skewed continuous data, as is common with water quality variables (especially microorganisms).

The gamma distribution is also used for right-skewed continuous data (it too is always right skewed). It is seen as a more versatile two-parameter (alpha and beta) right-skewed distribution than the lognormal distribution. For technical reasons it is preferred over the lognormal distribution to describe skewed data in Bayesian analyses.

Gamma distributions are often relevant in water quantity work, where extreme events may important even if rare (Sharla Stevenson, NPS, Personal Communication, 2011).

Is there an unusual distribution? Although considered perhaps less often in typical early EDA than the distributions above, other distributions for discrete random include binomial distributions for dichotomous data, Poisson distributions for symmetrical distributions mimicking randomness, negative binomial distributions for right skewed discrete data, hypergeometric distributions, and multinomial distributions. Each of these distributions was introduced as potentially one to consider in exploratory work in Chapter 2 of McBride (2005).

What is the Relationship Pattern between Two Variables?

When considering how any two variables vary compared to each (often time versus another variable in the case of trend work) how what hints to do we get from early visual inspections of plots? Do these plots hint at a linear pattern, a curved pattern such as a quadratic pattern, an exponential pattern, or something else?

Some patterns other than trend are expected. For example, in water quality work, we expect a relationship between flow (discharge) magnitude versus turbidity or total phosphorus in the water column. Early hints about potential relationships are usually seen in line plots, box and whisker plots lined up to compare two variables, or in simple or smoothed X/Y scatterplots. Another early line of evidence would be found in initial EDA summary statistics such as the Kendall tau [correlation coefficient](#), which tends to be better than the more common Pearson (parametric) correlation coefficient when data is not normally distributed.

Hints from Visual Inspection:

A common first step is simply visually look at the pattern in a line plot or X/Y scatterplot or line plot or smoothed scatterplot to see what pattern it seems to resemble the most closely. One can

choose between the patterns illustrated by NIST (2006) at <http://itl.nist.gov/div898/handbook/eda/section3/eda33q.htm>. At that website, one can click on each of the following options, and then simply visually inspect the pattern at hand to see which of the following patterns your plotted pattern seems to most closely resemble:

[No relationship](#)

[Strong linear \(positive correlation\)](#)

[Strong linear \(negative correlation\)](#)

[Exact linear \(positive correlation\)](#)

[Quadratic relationship](#)

[Exponential relationship](#)

[Sinusoidal relationship \(damped\)](#)

[Variation of Y doesn't depend on X \(homoscedastic\)](#)

[Variation of Y does depend on X \(heteroscedastic\)](#)

[Outlier](#)

It is safest to use linear methods only when there is a **strong** linear relationship. When the pattern is non-linear, a quadratic pattern tends to be fairly common in water quality and biology.

A quadratic pattern tends to have only one hill or valley. If a quadratic pattern is mostly downward as in the lower half of a C-shaped pattern, it may still be thought of as a mostly downward trend, it is just not a linear trend. Keep in mind that it may also be part of a longer term cycle if monitoring were to continue.

However, the terminology used related to quadratic patterns must be chosen carefully. If a longer term pattern totally reverses into a more classical U or upside down U – shaped pattern, the data may still be thought of as a quadratic pattern in cases where it fits quadratic equations and/or models well (Regnier and Ryan, 1999).

If a longer term pattern totally reverses into a more classical U or upside down U shape, it will not fit the definition of a monotonic (one way) trend and most trend tests, such as the Kendall tests, will conclude there was no trend.

In an exponential pattern, there is often a curve in the middle but a tendency for the ends of the line to be relatively straight, and the relatively straight lines at each end of the data do not tend to go in different directions as would be the case in a sigmoid S-curve.

A sigmoid S-curve or sinusoidal pattern will also not fit the definition of a monotonic (one way) trend and should just be thought of as a regular cyclical pattern. Most trend tests, such as the Kendall tests, will conclude there was no trend.

When looking at potential patterns in trends, perhaps the most common first questions would be does it look linear or non-linear? If non-linear, does it look most like a quadratic pattern or a cyclical pattern (sinusoidal) or something else? The something else includes a large number of other options.

Hints from Simple Curve or Line Fitting Software:

Many who have no dedicated statistical software do have access to MS Excel. As will be discussed in more detail below, there are better programs to use for this task, but if MS Excel is the only tool a beginner has handy, it can provide an additional line of evidence to supplement other visual clues discussed just above.

One can plot a quadratic trend line by choosing the order 2 polynomial trendline (same as a polynomial of degree 2 or second order, or a 2nd degree polynomial), which often tends to be a good fit the pattern has only one hill or valley. This is relevant because many automatically think of quadratic functions when they see a curved line and since NIST states that many curvilinear functions are possible, but the data analysis principle of parsimony suggests that we try fitting a quadratic function first. A polynomial with a degree of 0 is simply a constant and a polynomial with a degree of 1 is a line (a linear pattern). A polynomial with a degree of 2 is a quadratic pattern. A polynomial with a degree of 3 is a cubic pattern (see discussion of polynomial trend lines, including quadratic and cubic functions in section 4.8.1.1 of NIST (2006).

Even some who use much more sophisticated methods such as generalized additive model (GAM) methods and principal components analysis sometimes also look at simpler lines of evidence from MS Excel. For example, Sanger et al. (2002) first looked at patterns in simple scatter plots in MS Excel and also plotted mean hypoxia and super-saturation for each deployment duration interval in MS Excel and a polynomial (quadratic) trend line fit to the data, before going on the more complex GAM and principle component methods.

In Excel, it is relatively easy to use a simple line fitting routine to see which is a best-fit. Here is the step by step:

Line-fitting/curve fitting routines in Excel can be used by selecting chart and then selecting “add trendline.” In the add trendline dialog box, only if the pattern is clearly linear, choose linear. If the pattern in the scatterplot does not appear to be linear, choose the appropriate pattern (polynomial, logarithmic, power, exponential, or moving average (Sincich et al. 1999). One can compare the patterns seen visually to shapes recognized as linear or quadratic, etc. When in doubt, plot all the options and then as an additional line of evidence, choose the one with the highest R² value. A MS Excel help screen clarifies that “A trendline is most accurate when its R-squared value is at or near 1. When one fits a trendline to your data, Excel automatically calculates its R-squared value. If desired one

can choose to have Excel display this value on the chart.' One can then choose the pattern and plot that gives an R-squared value closest to one, reflecting the best initial fit of the data to the options for fitted trendlines. Using R-squared like this as one aid in early EDA pattern may be OK. However, it is not ideal for final data analysis, and complex modeling (especially for predictive models).

Cautions on Curve Fitting in MS Excel:

Though probably adequate for most summary statistics and simple plots and simple curve fitting in early EDA steps, MS Excel will usually not be the optimal program to use for later EDA and especially for final data analyses or model fitting. MS Excel is sometimes simply not accurate enough for final data presentation and final conclusions. Some would probably argue it is not even ideal for early EDA pattern recognition. Many criticisms of MS Excel use in statistics have been summarized and are on the Internet. Here is one example:

MS Excel has been shown to be inferior to more accurate programs like Minitab in fitting polynomial trend lines, including quadratic trend lines. In fact Excel will "fit" nonsense trend lines to data presented on column and line charts, and can report an inadequate number of significant digits for polynomial trend lines (including quadratic trend lines). In 2009, there were no apparent Excel 2007 workaround solutions for the forced-intercept polynomial trend line errors in MS Excel 2003 and 2007 (Hargreaves and. McWilliams, 2009).

So for recognizing curved patterns such as the quadratic pattern, rather than relying totally on MS Excel for curve fitting, one should usually look from additional angles. . At minimum, one might follow up the trendline line fitting in MS Excel by also plotting a moving median, LOWESS, or LOESS smoothed trend lines. However, even in MS Excel the curve fitting routines may result in a line that fits the middle of the data better than a linear regression fit if the trend or other relationship is not linear.

If the hints of trends do not seem linear (there seem cycles or curves) and if all one has is MS Excel (and there are few nondetects), one can plot a quadratic trend by choosing a second order polynomial from the line plot wizard and it will plot your data as a quadratic. However, Excel does not run a quadratic regression, only plots the data and spits out coefficient of determination and equation. Users won't be able to tell if the quadratic is significant or not (Matthew Hall, State of Maryland, Personal Communication, 2010).

Better statistical programs like SPSS and R have similar curve fitting routines to help one choose between the same common patterns. In a discussion of using SPSS software for similar curve fitting purposes Garson (2009) stated that

Curve-fitting tests whether a best fit line may be created based on any of a number of common models, such as linear, exponential, logistic curve, or others

A best curve fit determination may be made graphically.

While selecting the model with the highest R-squared is tempting, it is not the recommended method. Much more complicated methods are needed to really do it right.

The “pick pattern with the R-squared value closest to one” rule of thumb may be OK for early EDA pattern recognition, but for advanced predictive modeling it is not adequate. For example, in a discussion of curve fitting in R software, Maindonald (2008) explained that a predictive model is adequate when the standard errors of predicted values are acceptably small, not when R^2 achieves some magic threshold. The full remedies using R are complex however, when needed, consult your local statistician.

Better albeit often less user friendly options are discussed below

Hints from a Good Smoother:

Due to the issues with curve and line fitting routines in MS Excel, some might prefer to smooth the data and then comparing the patterns visually. In other words, for recognizing curved patterns such as the quadratic pattern, rather than relying totally on MS Excel for curve fitting, one should usually look from additional angles. . At minimum, one might follow up the trendline line fitting in MS Excel by also plotting a moving median, LOWESS, or LOESS smoothed trend lines. However, even in MS Excel, the curve fitting routines may result in a line that fits the middle of the data better than a linear regression fit if the trend or other relationship is not linear.

If one has access to one the better smoothing options (moving median, LOWESS, or LOESS), one should probably use those as another line of evidence to see the underlying patterns more clearly, then compare them visually to pattern recognitions choices such as those provided by NIST (2006) at seems to resemble the most closely. One can choose between the patterns illustrated by NIST (2006) at <http://itl.nist.gov/div898/handbook/eda/section3/eda33q.htm>.

Hints from Dedicated Statistical Software:

Though MS Excel line fitting and simple smoothing may be OK for getting early hints about relationship patterns, as data analysis progressed to more definitive stages, it is better not to rely only MS Excel and/or trying to identify patterns visually. Many prefer to do line fitting in better dedicated statistical software, which is probably better if one understands how to handle the program and what it doing. When it comes to final data analysis, complex modeling, and especially predictive models, better programs usually preferred.

When possible use dedicated statistical software such as Minitab for “best fit” line fitting (Hargreaves and. McWilliams, 2009).

EPA’s ProUCL (version 4.0) free software has goodness-of-fit (GOF) tests for normal, lognormal, and gamma distributed data sets and can properly handle data with or without nondetects. It is also free (EPA, 2010).

For forecasting, a more difficult task than initial data exploration, Statgraphics (2009b) has relatively easy options for forecasting based on moving averages, trend models, and simple, linear, quadratic, or seasonal exponential smoothing,

Matlab software has a curve fitting tool box. Therein, one can also choose between smoothing with localized regression using either a first-order polynomial (lowess) or a second-order polynomial (loess). See details in Mathworks 2010 and at <http://www.mathworks.com/products/curvefitting/description5.html>.

More detailed discussions of options for curve fitting in r software are in Maindonald (2008) and in scattered Internet summaries that are quickly evolving.

Is it Linear?

Although when sample sizes are adequate, tests for linearity are sometimes used, one often gets the first hints about linearity from time plots, and smoothed trend lines can help determine the pattern in busy (many data points) scatter plots. Of course one can also use the curve fitting routines discussed in the sections just above.

It is important to decide early on if the trend is linear vs. non-linear or the relationship between two variables of interest is linear or non linear. If it is non linear, it is important to decide if the pattern is a quadratic pattern or some other non-linear pattern (Zuur et al, 2007; McDonald et al, 2009). Garson (2009) and Helsel and Hirsch (2002) explain the hazards to trying to apply a linear regression method to a non linear pattern. Applying linear techniques to nonlinear data will underestimate relationships and increase error of estimate. The researcher must test or otherwise determine linearity (Garson, 2009).

Monitoring specialists sometimes transform the X and/or Y axis to try to achieve linearity before using linear regression. However transformations bring up potential pitfalls related to back transformation bias and/or “changing the question” so that overall, beginners should probably just use non linear methods instead.

If EDA suggests relationships are not linear and the distribution being sampled is not normal, resist the temptation to try to use linear methods on non linear patterns and resist the temptation to rely too heavily only on normal distribution statistics. At the end of this section some typical questions to be answered with level one EDA and some specific EDA tools to answer them are presented.

Linear patterns are unusual for environmental data, but if sample size for each variable is over 25 and the samples is fully representative of the full range of conditions of the underlying population, early hints of linearity include:

A scatter plot of the variable vs. time, suggests a fairly straight line (see NIST 2006 and example plot at <http://www.itl.nist.gov/div898/handbook/eda/section3/scatter2.htm>.

In level one EDA smoothed trend lines might first be done with moving medians or moving averages or even with a simple line fitting routine in MS Excel discussed earlier.

:

Is There A Non-Linear Trend Or Other Non-Linear Pattern?

It is fairly common for environmental data to have a curvilinear pattern that may or may not also suggest a monotonic (one-way) trend. For example, take a look at the quadratic pattern from NIST (2006) at <http://itl.nist.gov/div898/handbook/eda/section3/eda33q5.htm>. Many would look at that scatter plot and conclude there was a trend, but like NIST would call that a curvilinear trend. Others might even say it was not that different than a straight line, and might use linear trend tests like the Kendall tests as one line of evidence. However, it is not strictly a straight line pattern, so even though the Kendall tests are more tolerant of the lack of strict linearity than is a linear trend test, others might prefer to follow up with a quadratic test. Still others (Maryland and British Columbia) look at the results of both quadratic tests and a Kendall Trend test.

Early hints of a curvilinear but still monotonic pattern can be seen in early EDA line plots, box and whisker plots, histograms, and/or smoothed scatter plots. If one is trying to document a pattern between two variables (say flow and turbidity) rather than between one variable and time, the pattern seen in the plots may also be curvilinear or linear. Of course the pattern may also be totally reversing (as in a U shape) signifying an important pattern but "no trend."

To help with this question, one would ideally have a sample size of over 25 (10 at minimum) for each variable being compared. The sample should be fully representative of the full range of conditions of the underlying population. Early hints of a curvilinear pattern include:

In spite of some shorter term ups and downs along the way for a scatter plot, a smoothed trend line of the variable vs. time over the longer term suggests a curved line, but the trend does not totally reverse and go from up overall to reversing just as far down.

If box and whisker plots lined up by successive time periods such as years tend to have continually rising or falling medians, there is probably some kind of monotonic trend pattern, either linear or curved. As mentioned in our discussion of moving median smoothing, in this case connecting the medians with a line is a rough way to get a moving median smooth, and one could do the same with the 25th and 75th percentiles to see if they also have parallel or divergent patterns. If the pattern is consistently up or down but not a straight line, it is one hint of a non-linear monotonic trend. Looking at trends in extreme percentiles such as the 5th and 95th percentiles might reveal a different pattern for biologically important extremes than one might see for medians.

Is it Stationary?

This is another way to say: Is there a trend? We have touched on this repeatedly in our discussions above, but this question is so important that it is discussed in more detail here. If there is variability but no clear upward or downward change over time, the pattern may simply be stationary.

One often looks for early hints from time series line plots, time series scatter plots, box and whisker plots lined up side by side in a time series, or an X/Y scatter plot with a smoothed trend line of a best fit trend line.

A time series is stationary if there is no systematic change in the mean (i.e., no trend) or variance across time (EPA, 2009b; Morrison, 2011). In relatively simple terms, a plotted stationary times series looks flat and therefore suggests no strong hint of a trend. It has a constant variance over time, a constant autocorrelation structure over time, and no major seasonal fluctuations (NIST, 2006, section 6.4.4.2).

Deciding whether or not the data show hints of stationarity is especially important if one is going to later use the data later in complex Autoregressive Integrated Moving Average Model ([ARIMA](#)) calculations (Morrison, 2011). Likewise, if autocorrelation is present, before one can optimally analyze patterns of autocorrelation and clearly separate these patterns from the trends or cycles, one may first have to remove trends and cycles from the data (Chatfield, 1982).

Since our overall focus herein is on long term monitoring, keep in mind that in that context “Stationarity is dead.” In other words, there is almost always a trend in environmental data if one collects data long enough (Milly et al, 2008). Therefore identifying all important patterns, and not just trends, is important.

Are There Cycles?

The smoothed trend line plotted suggests no longer term trend overall but shows cycles that reverse direction completely (it is and up and down pattern reversing in regular cycles). What do I do next?

Since there is no trend, use EDA to describe the pattern. One does not always need a trend test, particularly a linear test if the relationship is non-linear data (Irwin, 2010).

So first describe the regularity and the frequency of the oscillations. Use the plot to estimate how long the cycles last (what is the distance between peaks) and whether or not they seems to correspond to known short term patterns (diel or seasonal) in other variables (time of day or month) or longer term cycles (Atlantic decadal oscillation or El Nino or La Nina weather patterns, etc.). One might first line up plots of the known climatic cycles versus the pattern in the environmental variable being measured to see if there are similarities. If the relationships between the cycles in the environmental variables are not fairly simple, the logical next steps in documenting the relationships tend to be complex (See the Level Two EDA discussion below and/or consult your local statistician).

Are There Outliers?

Outliers are important enough to warrant a separate discussion here. Outliers are most commonly identified as data points that seem to lie outside of the common patterns seen for other data. Hints of outliers can be seen in many of the EDA plots discussed herein.

I

Initial identification of outliers is usually an important part of first level exploratory data analysis (EDA). Outliers out of common ranges can often initially be spotted in time-plot scatter plots of a variable versus time or in box and whisker plots.

Outliers may also be identified as points outside of normal ranges in [QC control charts](#) based on upper and lower percentiles such as a simplified tolerance interval bounded by the 5th and 95th percentiles. When data are normally distributed (**rare in environmental work except for precision or sensitivity QC replicate samples**), the mean plus or minus three times the standard deviation is sometimes used in control charts as a rough 99% tolerance interval to help identify outliers that lie beyond the interval.

A normal Q-Q plot is another of several EDA tools that can be used to help identify outliers and multiple populations that may be present in a data set. On a Q-Q plot, observations well separated from the majority of the data may represent potential outliers, and jumps and breaks of significant magnitude may suggest the presence of observations from multiple populations in the data set. It is suggested that analytical outlier tests (e.g., Rosner test) and goodness-of-fit (G.O.F.) tests (e.g., SW test) should always be supplemented with the graphical displays such as Q-Q plot and box plot...In other words, there is no substitute for graphical displays of a data set, but don't throw out outliers for statistical rule of thumb reasons without first consulting subject matter experts. ProUCL 4.0 can draw Q-Q plots and box plots for data sets with and without nondetect observations (EPA, 2009).

Outliers are usually addressed in EDA in two steps: 1) Identifying them, and then 2) Deciding if they are junk outliers (mistakes) or if they are possibly real values. If they cannot be confirmed as mistakes they may be very important (information-rich) true-value outliers. Many experts strongly emphasize that one should not throw away true outliers when they are important to the question to be answered, as they often are (Helsel and Hirsch, 2002; Gilbert, 1987; Kittel, 2009), see additional discussion in next section).

Don't perform EDA tasks including outlier identification only on transformed values due to the EDA basic of staying close to the original data, the desire not to change the question, and due to potential complications such as back-transformation bias.

Kittel (2009) provided a summary of outliers in climate data, including a discussion of multivariate outliers. The Rosner Test for detection outliers and discussions of using [control charts](#) to help identify outliers are included in a chapter on outliers in Gilbert (1987).

Don't Just Throw Out all Outliers!

Again (for emphasis), statistical outliers are not necessarily cause for alarm and should not necessarily be thrown away. If outliers are just automatically deleted, one risks only seeing what one expects to see and disregarding the most important data (Helsel and Hirsch 2002). In other words, however outliers are identified, once they are identified, further investigation should be initiated to determine if they are mistakes or potentially true and important values.

Do not throw out outliers based only on statistical tests or rules of thumb based on ranges or control charts. The Antarctic ozone hole discovery was delayed 10 years because automatic data processing routines were throwing out extreme values that the routines assumed must be wrong (Helsel and Hirsch, 2002; NIST, 2006).

If outliers are set aside or de-emphasized as a data analysis strategy to answer a certain question (such as what is the yearly median), the raw data itself should not be discarded unless it is very clear that the outliers are mistakes. A few outliers don't impact a large-sample-size median much. However, later someone might want to calculate yearly averages, some extreme but true-value outliers tend to influence means more. Keep all raw data not clear identified as mistakes, as someone might want to look from a different angle later.

How Does Software Identify Outliers?

It varies by software package, but if the software package has a default way to identify outliers, for example with [control charts](#), one still has to check the outliers to see if they might be true values or not.

ProUCL software provides a way to look for parametric outliers not only with simple plots but also using statistical outlier tests (including the Rosner Test and the Dixon Test) that give the analyst probabilistic evidence that an extreme value (potential outlier) does not “fit” with the distribution of the remainder of the data and is therefore a “statistical outlier” (Singh et al., 2010).

Aquarius software uses rate of change to help identify outliers.

More robust and complex outlier identification methods not found in most commercial software methods are available in EPA Scout 2008 software (EPA, 2008 and Singh et al 2010). A major part of Scout software deals with classical, robust, and resistant univariate and multivariate outlier identification, and robust estimation methods that have been available in the statistical literature over the last three decades. Outliers in a data set represent those observations which do not follow the pattern displayed by the majority (bulk) of the data. It should be pointed out that all of the outlier identification methods are meant to identify outliers in a data set typically representing a single population. Outlier identification methods are not meant to be used on clustered data sets representing mixture data sets, especially when more than two clusters may be present in the data set. On data sets having several clusters, other methods such as cluster analysis and principal component analysis may be used...For multivariate data sets with nondetect observations, Scout can compute mean vector, covariance matrix, prediction and tolerance ellipsoids, and principal components using the Kaplan-Meier method. For multivariate data sets with nondetects, Scout can also generate Q-Q plot of Mahalanobis distances and prediction and tolerance ellipsoids (EPA, 2008).

What is Data Verification?

Not all agencies have identical guidance. Validation and verification steps sometimes get mixed up, but identifying transcription or copying errors is usually done in data verification steps. Data verification is usually done before data validation. For a good example of clearly differentiating verification from validation as two separate steps, see example guidance from the Sierra Nevada Vital Signs Network at

http://science.nature.nps.gov/im/units/sien/datamgmt/chapters/SIENDMP_Ch7.pdf.

What is Data Validation?

The exact steps vary by agency, but identifying outliers that may or may not be mistakes is usually done as part of data validation steps. Some agencies emphasize finding and discarding junk (mistake) outliers than others.

NPS guidance for data validation for the Sierra Nevada Vital Signs Network at http://science.nature.nps.gov/im/units/sien/datamgmt/chapters/SIENDMP_Ch7.pdf is a fairly complete example which clearly places a strong emphasis on finding and eliminating mistakes or junk outliers:

Validation is the process of reviewing computerized data for range and logic errors and may accompany data verification *only* if the operator has sufficient knowledge of the data and subject. More often, validation is a separate operation carried out *after* verification by a project specialist who can identify generic and specific errors in particular data types... Invalid data commonly consist of misspelled species names or site codes, wrong dates, or out-of-range errors in parameters with well defined limits (e.g., pH). More interesting and often puzzling errors are detected as unreasonable metrics (e.g., stream temperature of 70°C) or impossible associations (e.g., a tree 2 feet in diameter and 3 feet high). These types of erroneous data are called *logic errors* because they produce illogical (and incorrect) results. The discovery of logic errors has direct, positive consequences for data quality and provides important feedback to the methods and data forms used in the field. Histograms, line plots, and basic statistics can reveal possible logic and range errors... An outlier is simply an unusually extreme value for a variable, given the statistical model in use.” Any data set will undoubtedly contain some extreme values, so the meaning of ‘unusually extreme’ is subjective. The challenge in detecting outliers is in deciding how unusual a value must be before it can, with confidence, be considered ‘unusually’ extreme. Data quality assurance procedures should not try to eliminate outliers. Extreme values naturally occur in many ecological phenomena; eliminating these values simply because they are extreme is equivalent to pretending the phenomenon is ‘well-behaved’ when it is not. Eliminating data contamination is perhaps a better way to explain this quality assurance goal. When an outlier is detected (via GIS, database, graphic, and statistical tools for ad-hoc queries and displays), the possibility of contamination will be evaluated and noted.

Are There Patterns in Missing Data?

If data is sparse, one can use line plots (basically scatter plots where the few data points are connected by lines) with breaks in the line for missing data. If there are many data points, consider using moving averages or other smoothers. The scale used on the vertical axis can influence utility of interpretation, so plot them over the range of scientific interest. Also indicate sample sizes and uncertainty of each data point when such uncertainties are non-[trivial](#) compared to real variability (Chandler and Scott, 2011).

It is important to notice these patterns in missing data early on, since when later analyzing for longer term trends, say with a Seasonal Kendall test, try to select seasons for which there is no missing data. For example if one had daily values for 10 years, and July to August was the only

period over the 10 years that had no missing data, one might draw a time series line plot of July to August medians to visualize trends considering only the July to August time period. One might then follow that up with Mann Kendall trend test using only the data from July to August. Since seasonality has basically been removed, the Mann Kendall can be used instead of seasonal Kendall test. See also the detailed discussion below on the topic of [testing for trends in medians](#).

When using the seasonal Kendall test for trends, an allowance for missing data can be made. In fact, non-parametric tests are sometimes chosen because 1) they are not affected when the distribution of data is not normal, 2) are insensitive to outliers, and 3) are less impacted (or not impacted) by missing or censored data. Decision rules included the following: “How many missing values are needed in a data set before it is necessary to collapse the data from monthly to quarterly?” After an extensive simulation study, Harcum et al. (1992) concluded that:

1. When applying the Mann-Kendall or Seasonal Kendall tests, collapse monthly data to quarterly values if more than 50% of the monthly data are missing.
2. When applying the Seasonal Kendall test with correction for serial correlation, collapse monthly data to quarterly values if more than 40% of the monthly data are missing.
3. When there are only five years of record and more than 50% missing values, there is not a good alternative.

Are There Patterns in Nondetects?

A related question is how should Nondetects be handled in early EDA?

Nondetect values are value below a detection limit. In chemistry, below that value, the measuring system cannot distinguish a chemical concentration from zero, so there no confidence that the chemical is present. These often come into play in water quality and contaminants monitoring. They are less often used for other types of monitoring, though there is no particular reason they couldn't be used. For example, when trying to identify the presence of bird or amphibian species by faint calls, at some point the signal becomes so weak that it is difficult to distinguish it from zero (background noise), so one could use detection limits in those scenarios. For a complete explanation of detection limits, see Irwin (2008).

In EDA, one often calculates the percent of nondetects and, if there more than just a very few rare nondetects, is also often helpful to make plots to see if there is a pattern to the occurrence of nondetects in time or space.

Acceptable methods for dealing with nondetects were documented in the Helsel book on non-detects (Helsel, 2005). Until the revised version becomes available, Internet-available discussions of many of the Helsel recommendations include EPA (2010) and Singh et al. (2010).

EPA's free Software ProUCL version 4.0 offers useful graphical displays for data sets with or without NDs, including: histograms, multiple quantile-quantile ([Q-Q plots](#)), and side-by-side box plots. The use of graphical displays provides additional insight about information (such as hidden data structures) contained in data sets that may not be revealed by the use of estimates (e.g., 95% upper limits) or test statistics, such as the GOF test statistics or the Rosner test. In

addition to providing information about the data distribution (e.g., normal or gamma), Q-Q plots are very useful in identifying potential outliers or the presence of mixture samples (e.g., data from different populations) in a data set. Side-by-side box plots and multiple Q-Q plots are useful to visually compare two or more data sets (groups), such as: site-versus-background contaminant concentrations, surface-versus-subsurface concentrations, contaminant concentrations of groundwater monitoring wells (MWs), and the areas of concern (AOCs) of a potentially polluted site (EPA, 2010).

It is important to understand which type of nondetect is being dealt with. For water chemistry and contaminants work, there are **quantitative detection limits** such as Minimum Level of Quantitation (MLs). Alternative names for quantitative detection limits (though some are used only in multi-lab scenarios) include practical quantitation limits (PQLs), minimum reporting levels (MRLs), reporting levels (RLs), limits of quantitation (LOQs), minimum quantitation limits (MQLs), sample quantitation limits (SQLs), Contract-Required Quantitation Limits (CRQLs), or an inter-laboratory quantitation estimate (IQE). All of these are considered quantitative limits. It is quantitative detection limits (and not **semi-quantitative detection limits** such as MDLs, LODs, IDLs, EDLs, or LT-MDLs) that should be used for most detection limit rules of thumb (EPA, 2009).

As part of spatial EDA, GIS Maps are sometimes used to plot the occurrence of nondetect data on maps and one could also plot nondetects vs. time or driver events (floods etc.) to look for temporal patterns. However, it is important to not put too much weight on mapping detections versus nondetects on GIS maps or vs. time if detection limits are not consistent over the whole geographical area and time period covered and/or if quantitative detection limits are not lower than resource protection comparison benchmarks or thresholds.

Those in long term monitoring programs should realize that detection limits almost always change over longer periods of time, so if one is plotting nondetects on a map, one may be plotting apples and oranges. How often should detection limits be re-calculated by a chemical lab? Most commonly they are calculated at least once a year.

Don't replace nondetect data with zero or with half the detection limit (EPA, 2010; Singh et al., 2010, and Helsel, 2005). Although EPA's unified statistical groundwater guidance (EPA, 2009) is a good statistics manual on many topics, it is less ideal when it comes to nondetect advice. There are cases where 10% nondetects strategically placed can wreck the results for linear regression by resulting in poor estimates of slope and intercept when substitution by the DL/2 (half the detection limit) is used (Dennis Helsel, PracticalStats, Personal Communication, 2011).

Level Two EDA:

Level Two Basics:

Those who use EDA optimally don't stop with Level One summary statistics and plots to put into progress reports. Instead, they tend to approach data analysis with a continuing goal of repeatedly trying to understand more clearly the patterns in the data and the potential processes behind the patterns. In long term monitoring, that is a better approach than just presenting a few

key summary statistics for a progress report and then quickly moving to classical methods fitting preconceived notions (NIST, 2006).

As longer term data processing continues, there should be a continuing side by side process to **explore** and then try to **confirm** patterns that appear. One should not be too quick to use a linear regression model or test on a relationship that is not linear. A linear regression result might be significant, but the linear model may not have been the ideal model to use, so jumping to linear methods too soon may obscure the more important conclusion that the relationship was non-linear. Linear regression should not be used to numerically summarize relationships that are not linear (Hartwig and Dearing, 1979).

In other words, during level two EDA, one is beginning to think more about the statistics to be done next, including tests. A premature stand-alone trend test might suggest no trend, while a smoothed trend line might reveal an important pattern, say that trend was up and then down with no overarching trend. Smoothed trend lines can help identify trends in otherwise busy scatter plots and can also help one look for presence or absence of stationarity (lack of trend). Smoothed trend lines can also identify the duration and frequency of both short and longer term cycles. They can also provide hints about whether relationships are linear or not, which helps drive future statistical choices.

As additional data from longer term monitoring or additional studies becomes available, it is common to continue to plot temporal change patterns using side by side box and whisker plots, and/or scatter plots of a variable versus time with smoothed trend lines. Less often, confidence intervals about means are lined up over time as part of level two EDA, but resist the temptation to conclude that no overlap always means significantly different (run the hypothesis test to be sure).

Level Two EDA Software:

Most Level Two (and Three) EDA is done in dedicated statistical software such as Minitab, SAS, SYSTAT, STATGRAPHICS, MatLab, or R. TimeTrends freeware allows one to easily perform seasonal Kendall trend tests (the ones done most commonly in water quality) and also helpfully allows one to display trend lines with smoothing done either by LOWESS or by generalized additive models (Jowett, 2009).

In Statgraphics software, LOWESS stands for *Locally Weighted Scatterplot Smoothing* and is a technique that can be applied to any X-Y scatterplot to help visualize the relationship between the variables plotted on each axis. Robust LOWESS is an option in Statgraphics and is made “robust” or resistant to outliers by smoothing the data twice, down-weighting points far from the first smooth when the second smooth is performed. Other smoothing options in Statgraphics include running means, linear, and locally weighted regression without the robust option (Polhemus, 2005).

The Tukey (1977) robust line fitting routine in R software is explained at <http://projetos.inpa.gov.br/i3geo/pacotes/r/win/library/stats/html/line.html>.

Some earlier versions of Minitab software had ready methods to smooth data with moving averages and exponential moving averages and ‘resistant smoothers’ but no easy option for moving median,

LOWESS or LOESS smoothing. In some later versions one could get a LOESS smooth by selecting Graph > Scatterplot, then choosing “with Regression.”

Level two EDA concepts are explained in more detail in the form of typical questions and answers as follows:

TYPICAL QUESTIONS TO BE ANSWERED WITH LEVEL TWO EDA TOOLS:

First all the questions listed above for Level One should be revisited with more plotting and some additional and/or more sophisticated EDA tools. As one example, to further assess whether or not the distribution being sampled may be normal, one might supplement the simpler lines of evidence discussed under level one with other lines of evidence such as [Q-Q plots](#).

How Often and When Should I Collect One Variable?

In the past those planning future monitoring have tended to answer this question based cost and logistics. In today’s world is better to use past data from similar sites along with EDA and autocorrelation analyses and power analyses to help answer this question.

Now that auto-loggers and remote sensing makes getting more frequent measurements more feasible. It is now more common to base answers carefully on the questions to be answered and deciding how autocorrelation in time and/or space will be handled, Autocorrelation issues to be considered related to frequency of collections in time and how close together in space sites are discussed separately herein.

Statistical power to detect trends within a reasonable period of time should also be considered. For details, see Irwin (2008).

There is one post-hoc issue to bring up related to using power partly to decide how often to measure. Although post-hoc power analysis has been frowned in science, in planning future environmental monitoring, it is acceptable to use past data because (Ken Gerow, University of Wyoming Statistics Department, Personal Communication 2007):

- (1) Any data set can be thought of a “preliminary” in that it stands as fodder for future work.
- (2) It is indeed better to have (our best available) estimates of the standard deviation (which requires preliminary data).
- (3) Post-hoc simply means (“after”) which sounds in fact like the *right* thing to do (if one believes my points (1) and (2) are valid. So it is perfectly fine (good, even) to do post-hoc power analyses in this particular scenario of looking at past monitoring estimates to get the best possible estimates of variability to help plan for future monitoring.

Using simple EDA plots, one can also plot different frequencies in a time-series line plot as described in the section below and think through which important questions could only be answered with more frequent monitoring. Older questions tended to be about means. Which is more important to fish populations, the average annual pH or the number of consecutive 30 minute periods that pH exceeded

a water quality criteria? Thus, the newer and more biologically relevant questions tend to be more like these (Irwin, 2008; Irwin, 2010):

During the hottest period of the year, is the number of 15-minute periods during which the temperature exceeded a water quality standard or alternative benchmark (such as an ecological threshold or for endangered species X, Y, or Z) increasing?

During the worst case season (usually summer) and at the worst case time of day (usually just before sunrise) is the number of 30-minute periods during which the oxygen is below a water quality standard or alternative benchmark (such as a critical ecological threshold for species X, Y, or Z) increasing?

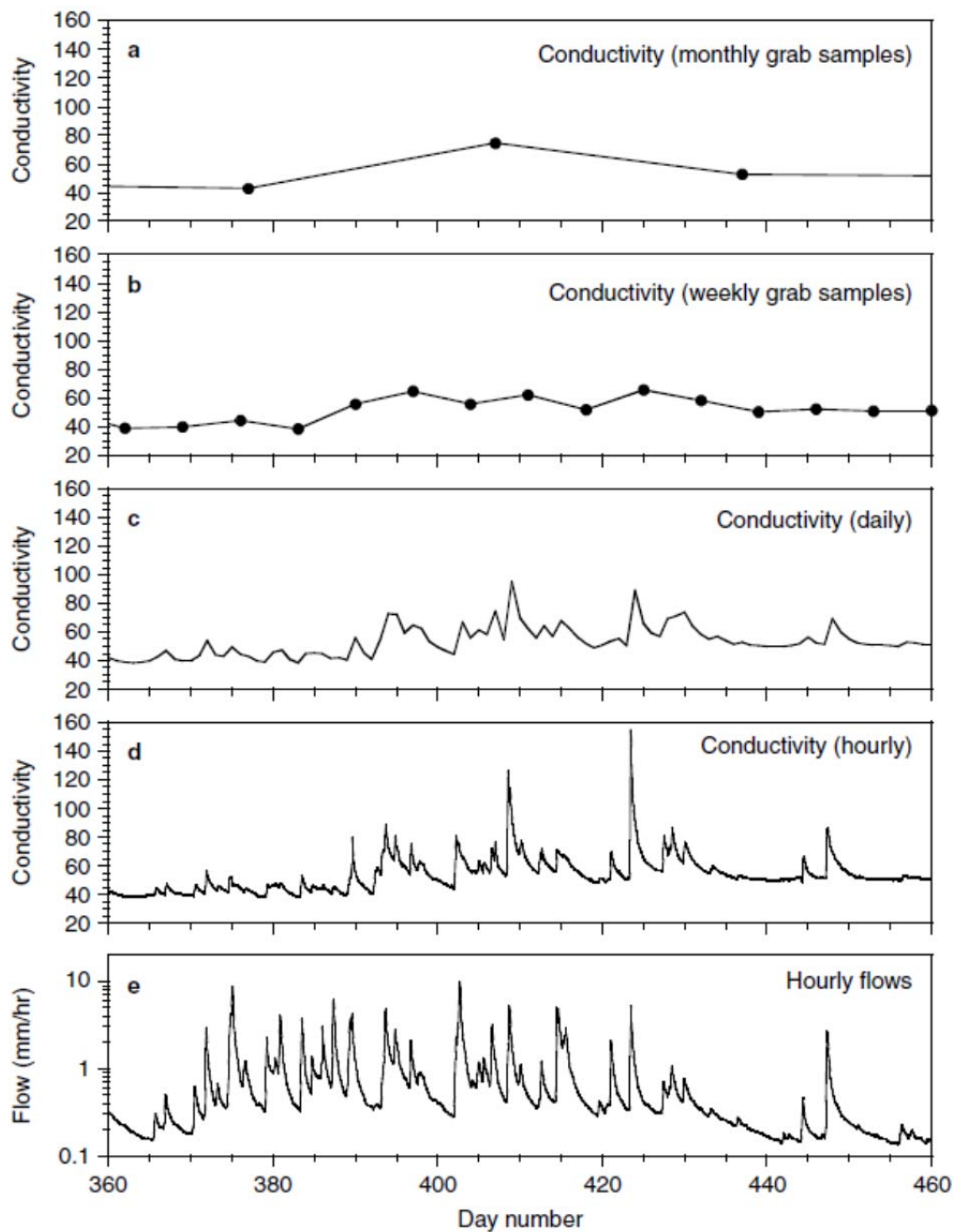
Is the number of one day periods when the stream is not flowing increasing?

Is the number of 30-minute periods when the stream is totally dry (with no pools) increasing?

In past long term monitoring, the tendency has been to monitor as often and as many places as possible and then let a statistician help sort out the data when done. A newer and better approach is to use some past data or pilot-scale data to help describe the patterns discussed herein, carefully think through the most important questions to be answered, consider power and autocorrelation, choose an appropriate survey design, then talk to a statistician about remaining questions before the monitoring design is completed.

How Often and When Should I Collect Two Variables I Want to Compare?

To begin to answer this question, one should typically think through the single variable issues discussed above, but also decide if the two variables should be collected at the same time. Often the same time and frequency is ideal, but if there is lag of time for a dependent variable to respond to an independent variable, that needs to be factored into the decision. For example, suppose that total phosphorus in the water column tends to rise after an increase in flow but that it takes a period of time before the increase in TP shows up. An easy EDA step would be to make a time series line plot for each variable with the same timing on the horizontal axis, then place the plots on top of the other and see how the patterns tend to match up and whether or not there is a lag time the response (see example plot below from Kirchner et al (2004, used with the author's permission). From this figure, one can see that one might be able to link conductivity to flow if both measures were done hourly, but trying to see the relationship between hourly flow and monthly conductivity would be a problem. One can also get an initial graphical hint about the amount of information lost going to less frequent measures. As pointed out by Kirchner et al (2004): 1) one loses information rich patterns by going to less than daily measures, and 2) the correlation between hourly flows and hourly conductivity is stronger during the later parts of the record.



In rivers and streams, conductivity tends to be a response variable that changes in response to changes in flow (discharge). Other similar response variables that tend to act in this manner include total phosphorus and turbidity. It is also true to some greater or lesser degree for many other water column variables. This is why we look at both exploratory data analysis patterns and trends in two ways: concentrations and flow-adjusted concentrations. On the other hand, more water soluble parameters (such as nitrate) tend to be less impacted by changes in flow than is

phosphorus or other substances that have a greater tendency to attach to soil or suspended sediment particles.

In comparing two variables, one might follow up initial frequency plots plot with correlation summary statistics, and one might also then look for instructive patterns in lag plots. Lag plots plot data against itself lagged by certain differences. A lag 1 Lag Plot is a plot of each data point against the very next data point. A lag 2 Lag Plot is a plot of each data point against the second data point. These plots provide one line of evidence about whether or not a data set is random. EPA (2006) and NIST (2006) provided illustrations of the following patterns using Lag Plots

Lag Plot: Random Data

Lag Plot: Moderate Autocorrelation

Lag Plot: Strong Autocorrelation and Autoregressive Model

Lag Plot: Sinusoidal Models and Outliers

One should also think through which important questions could only be answered with more frequent monitoring. For two variable comparisons, a newer and more biologically relevant question might be something like this:

Based on monitoring flow using an auto-logger stream gauge at the same short-time frequency as water quality variable X, is the [flow adjusted concentration](#) of X trending up or down, and how often would both measurements have to be taken (and how close together in time and for how many years?) to optimally define the relationship between the two or to develop a reliable equation so that a measure of X could be used as a surrogate for measuring flow or vice versa? In this case, X might be something highly driven by flow magnitudes or first flush events, such as total phosphorus, turbidity, or certain bacteria measures. The 'how many years' part of the question is designed to more fully explain the relationships over a full range of conditions (wet years, dry years, early snow melt, late snow-melt), etc.? To cover the full range of conditions, the sampling needs to be done long enough to cover important cyclical changes (diel, seasonal, or multi-year), or the conclusions need to be limited to the limited times actually sampled. In space, to cover the full range of conditions, the sampling needs to be done broadly enough to cover the full range of conditions in space, or the conclusions need to be spatially limited to the areas actually sampled.

Do More Complex Lines of Evidence Also Suggest Normality?

Probability plots are used to determine how well data fit a theoretical distribution, such as the normal, lognormal, or gamma distributions. This could be attempted by visually comparing raw data histograms of sample data to density curves of the theoretical distributions...However, research into human perception has shown that departures from straight lines are discerned more easily than departures from curvilinear patterns. By expressing the theoretical distribution as a straight line, departures from the distribution are more easily perceived. This is what occurs with

a probability plot. Q-Q plots are one type of probability plot used to help identify normality (Helsel and Hirsch, 2002).

Does a Q-Q plot suggest a normal distribution (see NIST, 2006 and accompanying example plot at <http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm>)?

Does a crude tolerance interval formed by the mean plus or minus twice the standard deviation cover about 95% of the data. In other words, is the length of this crude empirical-rule-based tolerance interval about the same as a crude nonparametric tolerance interval between the 2.5% percentile and the 97.5% percentile?

Is My Data Is Independent?

My data has been collected close together in either time and/or space, or I am collecting from a relatively stable (not very variable) medium such as deep groundwater. Are there EDA tools that would help me decide if the data is independent and can thus be used in the normal way for trend tests, confidence intervals, and other statistics?

Yes, it is important to identify autocorrelation patterns during EDA since data collected close together in time and/or space can be **autocorrelated** (not independent) and can bias statistical results.

What is Autocorrelation?

If one measures environmental variables too close together in either time or space, one will often not be measuring independent samples, but instead will be measuring autocorrelated samples. Independent samples are assumed for many statistical tests and calculations.

For data in many environmental time series data sets, the value of any given datum is largely determined by the value of the preceding datum in the series. This autocorrelation must be controlled before inferences may be made about correlation with other variables. Failure to control autocorrelation is very apt to lead to spurious results (Garson, 2009).

In water quality temporal autocorrelation (serial correlation) is a more frequent problem than in the past since now sondes or other auto-loggers are now often used to measure water quality very frequently. Since the samples are so close together in time they are usually temporally autocorrelated and thus not independent samples. If samples are not independent but instead are autocorrelated, the sample size (n) is essentially artificially too high and p-values tend to be artificially too low. For a very simplified plain language introduction related to water quality and the kinds of temporal autocorrelation issues often encountered in frequent collection of water quality samples, start with the summary on the Internet provided by Irwin (2010).

Autocorrelation can be thought of as an indicator of persistence in behavior of data, or how similar one data point (whether raw data or residual) is to other data points taken in nearby space or time. If autocorrelation is present, either it must be removed before doing most statistical tests or the tests must be adjusted. If autocorrelation is present, each observation is not independent of

the others and we do not have as much information as we thought we had on the basis of sample size. If autocorrelation is present in raw data, it may be the result presence of a regular cycle or a trend pattern. If autocorrelation is still present in residuals after such factors are removed, then trend analyses must explicitly account for autocorrelation (Reckhow et al., 1993).

A 2011 book on environmental trends states that discriminating between true trends and autocorrelation, and analyzing trends correctly in the presence of autocorrelation, is often one of the most difficult tasks in trend analysis. Therefore they prominently introduce autocorrelation basics in chapter one of their trend analysis text book and keep bringing it up in more complex ways in subsequent chapters. When using regression method for trend analyses, most of the difficulties are caused by the potential presence of autocorrelation or other factors that impact the variable of interest (Chandler and Scott, 2011).

As will be discussed below in more detail, one (relatively) simple way to deal with autocorrelation is to calculate a new “[effective sample size](#)” and use that in calculations instead of the larger number of samples actually measured.

If the data is autocorrelated, and it is analyzed in the usual way with no accommodation for autocorrelation, it will bias the results of the results of trend or other hypothesis tests, standard deviations, standard errors, normal variance estimates, standard errors about a mean, f-tests, t-distribution or x-distribution confidence intervals about a mean, nonparametric confidence intervals about a median, many if not most ANOVA applications, many regression applications (including regression options for trend analyses), prediction intervals, tolerance intervals, and many other complex statistics.

In regression analysis used relative to potential trends, if temporal autocorrelation (serial correlation) is present but is not accounted for, several nasty things will happen, including (Kirchener, 2001):

Spurious--but visually convincing--trends may appear in the data. These may occur in many different forms, including linear trends, abrupt steps, periodic or aperiodic cycles, etc.

Although the regression coefficients will still be unbiased, there will be a tendency to underestimate their uncertainties, potentially by large factors. For serial correlation coefficients of $\rho \approx 0.6$ and above, the actual standard error of the regression slope will be more than twice the estimate provided by conventional regression techniques. This effect increases with sample size, and increases drastically with the true correlation between residuals.

Because uncertainties will be underestimated, confidence intervals and prediction intervals will be too narrow.

Estimates of "goodness of fit" will be exaggerated.

Estimates of statistical significance will be exaggerated, perhaps vastly so.

What EDA Plots Help Us Understand Autocorrelation Patterns?

The first step is typically to plot an autocorrelation function (ACF) correlogram. Hints from the correlogram are typically used to decide whether or not any autocorrelation is present. Correlograms can also give us a rough first estimate of how far apart (in regular increments of time or space) samples have to be before they are considered independent so that no special steps have to be taken to deal with autocorrelation (EPA, 2009).

The number of periods apart is usually called the "lag". For example, an autocorrelation at lag 1 measures how values 1 period apart are correlated to one another throughout the series. An autocorrelation at lag 2 measures how the data two periods apart are correlated throughout the series. Autocorrelations may range from +1 to -1. A value close to +1 indicates a high positive correlation while a value close to -1 implies a high negative correlation. These measures are most often evaluated through graphical plots called "correlograms". A correlogram plots the autocorrelation values for a given series at different lags. This is a graphical representation of the ACF ---"autocorrelation function" (Morrison2011).

Either the rough approximation version ACF correlogram or the better large lag standard error ACF correlogram will help identify whether or not any autocorrelation is present. As will be discussed in more detail, below, the large lag standard error ACF correlogram is at least a bit better to answer a different question: At what lag does the autocorrelation disappear?

No matter which type of ACF correlogram is plotted, to get the cleanest answer for autocorrelation alone, the data plotted often needs to first be de-seasonalized and/or de-trended before final correlograms are plotted. If there are no strong seasonal or trend patterns this extra step may not be necessary. When in doubt try the correlogram on the raw data first. If no autocorrelation is indicated, one is done with the autocorrelation EDA. If autocorrelation is hinted in the correlogram and there are also hints of a trend or of seasonality or other strong cycles in the data. Extra steps may be needed. One may need to first de-trend the data. If autocorrelation is still present, one may then need to de-seasonalize the data or take out other cycles in other ways. Cycles will appear as alternating up and down patterns in the correlogram. If autocorrelation hints do not disappear after de-seasonlizing and de-trending the data, underlying autocorrelation is still present.

What Kinds Of Datasets Tend To Have Fewer Problems With Autocorrelation?

Data collected far apart in time and space tends to be less autocorrelated than data collected closer together in time and/or space. One would worry more about autocorrelation in water quality data collected from an auto-logger every 15 minutes (more common today) than about data collected once per year (as was common in the past). However, very stable groundwater samples might still be autocorrelated even if sampled once a year. Often, it is not clear so one needs to check with correlogram tools.

There is also less worry about autocorrelation in certain kinds of complete census surveys, as detailed below:

What EDA is Done for a Complete Census?

Most often summary statistics such as mean, medians or percentiles are calculated as a starting point. Beyond that, some of the options and considerations can become complex.

Companion Question: My sonde measures every 15 seconds, so wouldn't my average for a month be more accurate than if I measured say once a day or once a week? Wouldn't the monthly average or monthly median be more accurate and not less accurate? So in this case isn't there less need to be concerned with autocorrelation?

There is truth in this but the issues can be a bit tricky so one has to be careful with terminology and thoughtfully decide what to do. The summary statistics will tend to be more accurate; just don't calculate a standard deviation or a confidence interval about the mean or median in the normal way. A standard deviation as well as a confidence interval will be biased because the large amount of serial correlation will mean that most of the data points are not independent and the sample size used in the calculation would thus be greatly inflated compared to the case where one instead had used only independent data. Confidence intervals and many other advanced statistics require independent samples.

On the other hand, if all one did was report the mean and median, there would be no problem with autocorrelation. One might also look at trends in median from one season over the years, choosing say July as the season. The resulting medians would be far less autocorrelated than the original data from every 15 seconds, so that is one relatively simple way to handle autocorrelation, and the advantages of doing that might outweigh objections that one should not smooth data before doing trend tests.

However, in the case of data from every 15 seconds one could make the case that this very dense data set is basically a complete census, so the mean and median are perfectly accurate and therefore one need not calculate a confidence interval based on multiple measures from the environment to express uncertainty in the summary statistic. This basically amounts to the "I don't need no stinking inferential statistics" argument. The mean is the exact mean and the only uncertainty would be measurement QC uncertainty relevant to each single data point (Irwin, 2010). That single measurement uncertainty could be expressed either as NIST uncertainty or AMS uncertainty (Thoma et al. 2011; Irwin, 2008).

How Do I Plot an ACF correlogram?

Are there different types of correlograms, and if so, which ones are best?

Different types of software offer different types of correlograms as the default option. The two most common types are the **rough approximation correlogram** and the **large lag standard error correlogram**.

Several often quoted advanced time series text books seem to prefer routinely using the large lag standard error derived confidence interval instead of the rough approximation equation when routinely plotting ACF correlograms (Box et al, 1994; Anderson, 1976).

Some statistical programs or calculators, including Wessa (2008); Statgraphics (2005), Statistica Software (Statsoft (2008), Statgraphics, and Minitab; routinely generate a large lag standard error ACF correlogram with expanding error bars (using equations other than a correlogram based on the rough approximation equation. Meko (2009) discusses both types of correlograms in a relatively plain language in an Internet available discussion. He explains that the large lag standard error-based correlogram is better than the rough approximation equation based correlogram to help answer the question: At what lag does the autocorrelation disappear? However, the rough approximation version is probably adequate to answer the question, does my data contain any autocorrelation at any lag (Meko, 2009).

Strictly speaking, the rough approximation equation based correlogram is only used when the assumption is that the process is white noise, showing no autocorrelation (Statsoft, 2008; Box et al. 1994). One doesn't know that ahead of time.

Less commonly, if one has SPSS (SPSS, 2004) statistical software, or [STATISTICA advanced software](#), or some versions of Minitab software (Minitab, version 13.31 help screens from 2000, see also <http://www.minitab.com/en-US/support/>), be aware that those software packages may be plotting a confidence interval about zero on a correlogram based on Box and Ljung test rather than a confidence interval based on either the rough equation or the large lag standard error.

For those who have Statistica software, Statsoft (2008c) does not recommend either the large lag standard error or the moving average assumption or the rough approximation equation for the white noise assumption for data with missing values. Instead, specialized equations are recommended.

Which kind of ACF Correlogram is my Software Drawing?

If it is drawing an expanding confidence interval as the lags move from right to left, it is probably drawing large lag standard error correlogram. The rough approximation version will have a non expanding confidence interval always equal distant from zero correlation line parallel to the x-axis.

In R software, one can generate a large lag standard error correlogram by adding an “ACF function to an lme model,” but if the ACF function is chosen, the rough approximation correlogram will be drawn.

What is the rough approximation ACF correlogram?

It is also sometimes referred to as a “white noise” ACF correlogram because it helps one answer the question: Is there any autocorrelation in my data at any lag? In other words is the pattern random white noise only? This plot shows the auto correlation coefficient at multiple numbers of lags. If there is no significant correlation at any lag, none of the lag lines will extend above or below the confidence interval lines, and the indication is that the series is random.

Some software (Aquarius and some ACF script in R for example) provide only this option. It is easily calculated in spreadsheets too, since the equation is simply 95% confidence interval = zero

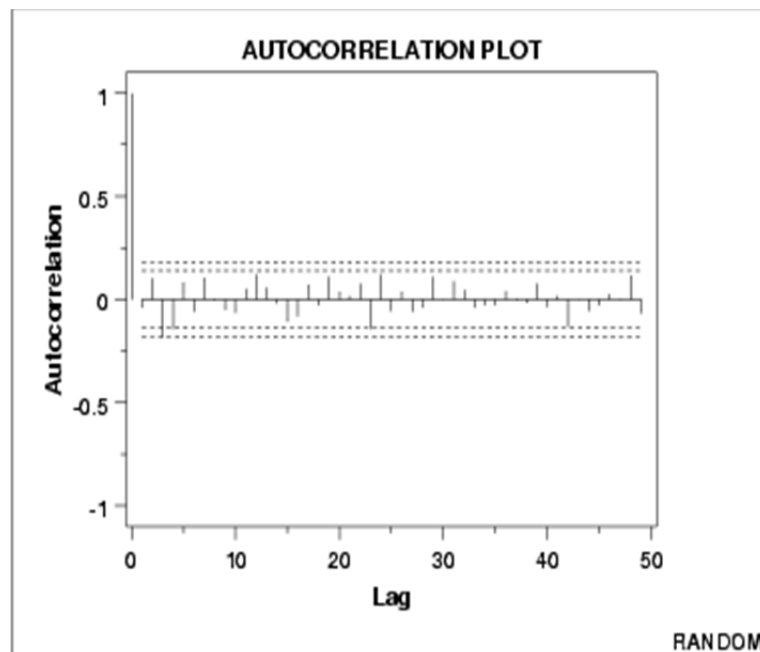
plus or minus $1.96/\sqrt{n}$ (the rough approximation). EPA (2009) says that one can simplify 1.96 to 2 and use the result in the following way: As a first approximation, one can count any autocorrelation as significantly different from zero if its absolute magnitude exceeds $2/\sqrt{n}$, where n is the sample size. The smallest lag (k) at which the autocorrelation is insignificantly different from zero is the minimum sampling interval ensuring temporally uncorrelated data (EPA, 2009).

All in all however, a generally better way to do this is to instead create a large lag standard error ACF correlogram, so doing that would generally be better, even to answer the simple question about whether or not there is any autocorrelation. Again, however, the rough approximation is better than nothing and considered good enough by many.

How do I interpret the results of a rough approximation correlogram?

Both the rough approximation-based correlogram and the large lag standard error correlogram are used to determine if there is any correlation in the data. If none or 5% or less of lags extend beyond the 95% confidence line, the conclusion there is little (5% or less) or no autocorrelation in the pattern.

NIST (2006) provides the following example of a rough approximation correlogram (they call it an autocorrelation plot) which hints there is no autocorrelation:



Although a few lags extend slightly outside the 95% (inner limit) and 99% (outer limit) confidence limits, the plot above would necessarily indicate non-randomness. For a 95% confidence interval, we might expect about one out of twenty lags to be statistically significant due to random fluctuations. Therefore, NIST (2006) concludes that plot above suggests that:

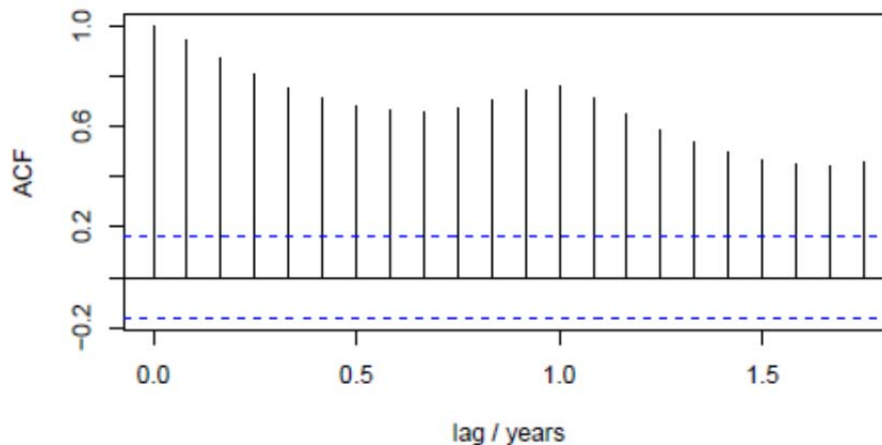
There are no significant autocorrelations.

The data are random.

Most software will draw only the 95% interval lines and not also include the 99% interval. The NIST freeware program Dataplot will draw both (NIST, 2002).

How Does a Correlogram Look When There is a Trend?

A raw-data correlogram can give us hints that there may be a trend in the data. Consider the following example of a rough approximation correlogram calculated in R software based on raw data (not de-trended, not residuals, not differenced). This ACF correlogram shows a gradual decay of [correlation coefficients](#) as lags increase, but no correlation coefficients that are non-significant or reach zero is typical of a time series containing a trend, as suggested in this example rough approximation ACF correlogram (Cowpertwait, 2006).



The same data would show a similar pattern in a large lag standard error (with an expanding confidence interval line through time) correlogram. Whenever one sees a slowly decreasing pattern over quite a few lags (say the first 10 lags), it is a hint of a trend (a non-stationary series). The mean has probably changed over the sampling period.

For the example plot above, one might first confirm that this is indeed a rough approximation or white noise ACF correlogram, since the dotted lines representing 95 are parallel to the zero line. Seeing the pattern above, one would probably conclude that the data should be de-trended and another correlogram done on the de-trended data to see if any autocorrelation remains.

How Does a Correlogram Look When There Are Strong Cyclical Patterns?

If one sees a regular up and down pattern in the correlogram, it is a strong indication that there is a regular cycle in the data (seasonal, diel, or longer). Seeing that, one would probably next take the cycles out (for example by de-seasonalizing the data) and then run the correlogram again to see if there is still remaining autocorrelation in the data.

At What Lag Does Autocorrelation Disappear?

We tend to get our first hints of an answer to this question not from tests but from EDA plots such as ACF correlograms. As explained just above, if there are strong trends or cycles in the data, one needs to get rid of them first before drawing the correlogram.

Hints from a Rough Approximation ACF Correlogram

Some use the rough approximation ACF correlogram as an initial EDA tool to help answer the question: at what lag does autocorrelation disappear?

EPA (2009) states that one can count any autocorrelation as significantly different from zero if its absolute magnitude exceeds 2 divided by the square root of n (the rough approximation equation), where n is the sample size. In planning the frequency to be monitored in the future, if a goal was to avoid autocorrelation, the smallest lag (k) at which the autocorrelation is insignificantly different from zero is the minimum sampling interval ensuring temporally uncorrelated data.

In a correlogram an estimate outside the limits would allow us to reject the hypothesis that the autocorrelation at that particular lag is equal to 0 (NWP Associates, 1997). Chatfield also suggested one could answer the question with the rough approximation correlogram. For monthly average air temperatures, there is a seasonal pattern so would typically have to de-seasonalize the data first (by subtracting the appropriate monthly average from each data point) to get the clearest correlogram. If the result was that the first three lags extend beyond the confidence interval line, those lags are the ones that are autocorrelated, indicating lack of independence. In this example, each monthly average will tend to be similar enough in temperature to the next few months, so we are only relatively sure that monthly averages separated by four months are independent/not autocorrelated (Chatfield, 1982)..

Hints from a Large Lag Standard Error ACF Correlogram

Although if that is all one had, one could use the rough approximation correlogram to decide how many lags need to go by before serial correlation disappears, when possible it is probably better to routinely use the large lag standard error ACF correlogram to answer this question instead. Using the (more precise) large lag standard error ACF correlogram, one can answer two questions: 1) Is there any autocorrelation in the series, and 2) At what lag does the autocorrelation disappear? The rough approximation correlogram is best used only in conjunction with the first question, though some also use it for the second.

Meko (2009) drives this point home with the following example from his figure 3.4 that has a correlogram of raw values that includes both 1) a constant width confidence interval line drawn from equation 1 (which he calls the rough approximation and 2) an expanding width confidence interval line based on equation 6 (which he calls the large-lag standard error. More detail:

Figure 3.4 in Meko (2009), reproduced in Irwin (2010) at http://acwi.gov/monitoring/conference/2010/C3/C3_Irwin.pdf shows an example for a tree-ring index time series illustrates the slight difference between the confidence interval computed from the large-lag standard error correlogram and computed by the rough approximation $\pm 2/\text{the square root of } n$, where n is the sample size. This equation is totally dependent on sample size. The alternative confidence intervals differ because the null hypotheses and questions to be answered differ. Looking at the rough approximation (white noise) correlogram, we conclude that there is autocorrelation in the data. The hint from the rough approximation correlogram is that the last lag where autocorrelation is significant is lag 5. Thus, in the rough approximation correlogram, the autocorrelation at lag 5 is judged significant under the null hypothesis that the series is random. If all one had was the rough approximation correlogram, one might conclude that the autocorrelation does not die out until lag 5. However, the large lag standard error is really a better way to begin to answer the question of how many lags have to go by before the autocorrelation dies out. Figure 3.4 shows that using the curved interval line of the large lag standard error equation, the autocorrelation dies out after lag 4 rather than lag 5.

That is a better answer, but others might counter that using a rough approximation based correlogram is better than the previous common practice of not considering autocorrelation at all. The results of the two types of correlograms will often not be that different. In example above the answer was lag 4 for the better large lag standard error and lag 5 for the rough estimate. Since both are only rough estimates, perhaps the difference is not often that important, so some argue that choice between the two is not that critical. To fully document autocorrelation patterns in more detail would require much more complex methods. The rough approximation version is sometimes the only one discussed in hydrology books (Gordon et al (2004) and many software programs only calculate the rough approximation version.

Bottom line: the large lag standard error correlogram is the better tool to answer the question. The rough approximation correlogram was really designed only to answer the question: is there any autocorrelation at all? However, if the rough approximation correlogram is the only one available, it still gives us a rough hint to how the question might be answered.

Hints from More Complex Lines of Evidence:

The Box-Ljung test can't answer the question about at what lag autocorrelation disappears, since instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags. This method tests the significance of autocorrelation at each lag. The Durbin-Watson coefficient tests only first-order autocorrelation, only at lag time-minus-one, so it doesn't help answer the question if it is not first order autocorrelation only (Garson, 2009).

Methods for using Box-Ljung test R software was summarized by Coghlan (2011a) in her discussion of time series tasks. She also has some examples of plotting the rough approximation ACF correlogram. .

How Do I Plot The Large Lag ACF Correlogram?

An easy way to do draw the correlogram not matter what software is at hand is to use the free Wessa (2008) R-based large lag standard error-based ACF correlogram calculator on the Internet. Unlike most other software, that one also simultaneously produces a partial autocorrelation function (PACF) correlogram that has other uses (to be discussed later). and is also the default version in certain software, For example Statgraphics software includes this large lag standard error as the default choice for drawing an ever-widening (as lags increase) 95% confidence interval across a correlogram, while using equation 1 for PACFs (Statgraphics, 2005).

Autocorrelation patterns are often initially identified as part of Level Two EDA correlogram plots. These include autocorrelation function (ACF) correlograms (for Internet examples, see NIST 2007 and Meko, 2009). Most simple correlograms including some ACF defaults in R software are estimated with a simplified and shortened equation. More defensible correlograms are provided by using instead using a large-lag standard error correlogram (Irwin, 2010). A free Internet calculator for the large-lag standard error correlogram was provided by Wessa (2008). If the PACF correlogram indicates the presence of first order autocorrelation only, then relatively simple “[effective sample size](#)” equations can be used to calculate an alternative sample size to be used in subsequent statistical calculations of hypothesis tests or confidence intervals to ameliorate the effects of autocorrelation (Kirchner, 2001, Wessa, 2008, Meko, 2009).

NWP Associates (1997) STATLETS software also calculates an expanding error bar for an ACF correlogram and a horizontal (non-expanding) error bar for PACF correlograms.

What does the PACF tell me?

The partial autocorrelation function (PACF) plot can help provide answers to the following questions:

- Is an AR model appropriate for the data?
- If an AR model is appropriate, what order should we use?

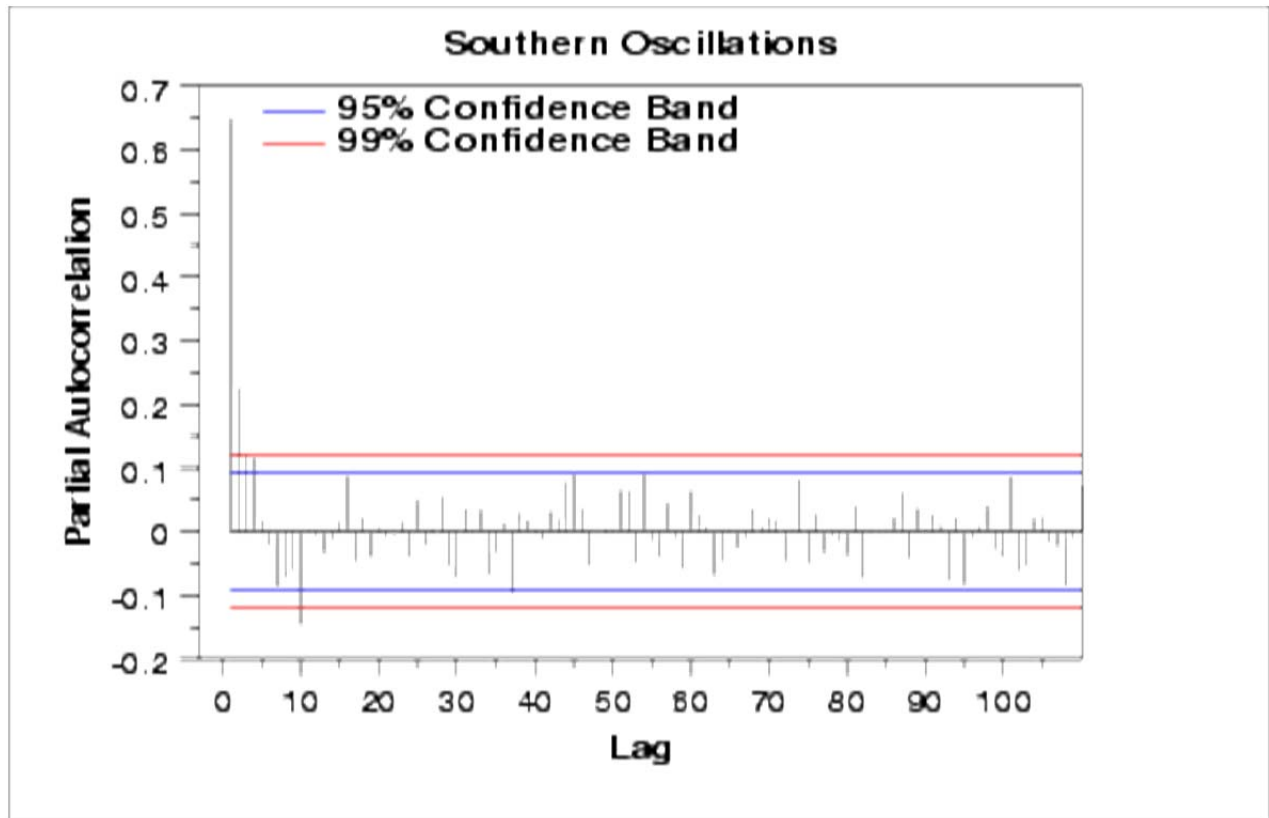
If it is first order autocorrelation only, as it often is, one can also use the relatively simple option of using an [effective sample size](#) rather than the number of observations in advanced statistics.

As explained in more detail, NIST (2006, Section 6.4.4.6, at <http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc446.htm>, the PACF can use used help identify which kind of Box-Jenkins Model might be appropriate if one were to use those methods following early EDA.

PACF plots tend to be more difficult to interpret than the ACF correlograms and a complete discussion is beyond our goals herein.

NIST (2006, Section 6.4.4.6.3 at <http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4463.htm> suggests that for the following PACF correlogram plot, one interpretation would be that this partial autocorrelation plot shows clear statistical significance for lags 1 and 2 (lag 0 is always

1). The next few lags are at the borderline of statistical significance. If the autocorrelation plot indicates that an autoregressive (AR) model is appropriate, we could start our modeling with an AR(2) model. We might compare this with an AR(3) model. .



After looking at a PACF correlogram, one would often begin to think about what kind of [ARIMA](#) model might best fit the data. Morrison (2011) explained in (relatively) plain language that a (mixed) ARIMA model is usually stated as $ARIMA(p,d,q)$. This represents the order of the autoregressive components (p), the number of differencing operators (d), and the highest order of the moving average term. For example, $ARIMA(2,1,1)$ means that one has a second order autoregressive model with a first order moving average component whose series has been differenced once to induce stationarity... Typically one's data represents only a sample of the underlying process. This means that sampling errors (outliers, measurement error, etc.) may distort the theoretical identification process. That is why traditional ARIMA modeling is an art rather than a science.

Rules of thumb on hints one gets from a PACF correlograms while practicing this "art" include the following (Garson, 2009):

ARIMA (0,1,0): Random walk model. The only effect is a non-seasonal [differencing](#) to remove a linear trend. ACF is either constant or is balanced between positive and negative. PACF is spiked only at lag 1.

ARIMA (p,0,q): ACF and PACF both decline slowly toward 0. PACF declines erratically due to shock effects.

ARIMA (1,1,1): A mixed model. *Warning*: normally one does not include both autoregressive effects and moving average effects in the same model because this may lead to over fitting the data to noise, and may reduce the reliability of interpretation of the significance of individual components in the model.

As a rule of thumb, one may wish to start with $p=1$ and/or $q=1$ and then increase the p and/or q values if the ACF and PACF for the residuals display spiking.

How Do ARMA and ARIMA Models Fit Into The Grand Scheme Of Things?

Autoregressive Moving Average Models (ARMA) and Autoregressive Integrated Moving Average Models (ARIMA) time series methods are often associated with very complex predictive models. Although one gets hints about potential later ARIMA model fitting in EDA (Garson, 2009), actually working with ARIMA and ARMA models is typically beyond EDA and beyond the skill set of our beginner target audience.

In relatively plain language, one might contrast ARIMA and ARMA methods with simple Kendall family trend tests as follows (Dennis Helsel, Practical Stats Consulting, Personal Communication, 2011):

Parametric methods (including ARIMA and ARMA models) excel when the model itself is important, as with prediction. One can include complex terms such as serial correlation, seasonal terms, along with the trend slope. It takes time to learn how to use these models well. Once learned, it takes time to apply them correctly to a given dataset.

Nonparametric tests such as the various Kendall trend tests excel when it is the test that is of importance. These tend to be more flexible and forgiving and harder to mess up. They are also fast and can be used over a wide variety of types of data.

ARIMA models are often used when trying to fit the pattern of a single variable over time (in a time series). When one series is seen as a function of one or more other data series it is now common to use procedures associated with linear mixed models (LMM). When one series is seen as a function of one or more other data series but the focus is on time-to-event, it is common to use a parametric event history procedure or the semi-parametric Cox regression procedure (Garson, 2009).

If an ACF correlogram indicates that several of the first few lags are significant but gradually decay to non-significant after that, while at the same time a PACF correlogram (sometimes called a partial-correlogram) indicates that only the first lag is significant, there is no hint of long term trend and the underlying process might be an AR(1) process having first order serial correlation (Cowpertwait, 2006).

Those who want to learn more could start with taking a time series class and by reading the seminal text book on time series analyses (Box and Jenkins, 1976). A few ARIMA Internet summaries include Morrison (2011), Garson (2009), and NIST (2006) Otherwise, consult your local applied statistician.

What is Differencing?

This is a complex subject, so we will just introduce some basics here. One of the (relatively) plain-language Internet explanations of differencing found to date was provided by Garson (2009) who clarified that:

Differencing is a data pre-processing step which attempts to de-trend data to control autocorrelation and achieve stationarity by subtracting each datum in a series from its predecessor. Single differencing is used to de-trend linear trends. Double differencing is used to de-trend quadratic trends. Differencing will drive autocorrelation toward 0 or even in a negative direction. As a rule of thumb, if single differencing yields autocorrelation spike in an ACF plot $>.5$, then there has been over-differencing. Over-differencing is also indicated by a residual plot where there is a change in sign from one time period to the next. As another rule of thumb, the optimal level of differencing will be the one with the lowest standard deviation. Sometimes adding moving average (q) terms (see below) to the model will compensate for moderate over-differencing. Note that with each degree of differencing, the time series is shortened by one

For those using r software, Coghlan (2011a) explained sequential differencing in (rough approximation) ACF correlograms in her discussion of time series tasks. For example, one might try plotting the correlogram for once-differenced data as one step.

There are different kinds of differencing, for example, one might choose to use first order Markov differencing (Matalas and Sankarasubramanian, 2003). Others might try other types of differencing.

Differencing is done as one way to get rid of trends before proceeding to more advanced correlogram or later ARIMA model steps. However, remember to use the raw data and not the de-trended data for trend tests and trend line plotting. In other words, there is one potential catch related to de-trending. Serial correlation is a potential problem for most regression analyses. To look at serial correlation in a complete way, one first needs to take out seasonal signals (by de-seasonalizing) and longer term trend signals (by de-trending) and other obvious signals and then look at the residuals left over to see what serial correlation is still left in the data. Again, however, it makes no sense to next look for trends in data that has been de-trended.

For those interested in more technical detail, Matalas and Sankarasubramanian (2003) discussed (complex) fractional Gaussian noise (FGN) processes, provided an Internet explanation that things are different depending on whether or not the exact specification of the Markov process is well documented

What is Pre-whitening?

Both differencing and pre-whitening are sometimes done before drawing correlograms and especially before getting too far into ARIMA models. There are several types of differencing and several types of pre-whitening, so don't assume all authors mean the same thing using these terms, and don't assume all software programs are doing the same things either.

Whereas a common goal for differencing is getting rid of trends, a common goal for pre-whitening is to get rid of autocorrelation. Other authors have stated they are using pre-whitening partly to remove seasonality. Both pre-whitening and differencing seem to be done at times at least partly to help deal with autocorrelation. So if a raw data correlogram indicates the presence of considerable autocorrelation, one might pre-whiten the data to get rid of the autocorrelation and cycles and try the correlogram again. One might also try differencing the data to get rid of trends, and then draw the correlogram again.

In the case of working with ARIMA models, SAS statistical software gives one the option of using a similar step by step process. If the need for differencing is indicated the time series can be differenced before the pre-whitening filter is applied.

The Box-Jenkins ARMA and ARIMA modeling approaches are sometimes used partly in the contexts of pre-whitened data.

CAUTIONS ON PRE-WHITENING:

When done alone, pre-whitening attempts, and fails, to remove serial correlation (temporal autocorrelation) while preserving trend. It often at least partially removes both. Over the years, some have tried to adjust Mann Kendall tests for serial correlation using "pre-whitening." At times this can actually make matters worse, inadvertently stripping out some of the trend while trying to perform a correction for serial correlation (Helsel, 2011).

The following is more detail on the relationship between pre-whitening and the Mann Kendall Tests from Matalas and Sankarasubramanian (2003)

Sometimes one pre-whitens data to get rid of strong trends. Yeu and Wang (2002) showed that pre-whitening water quality time series is not always a simple cure all for nonparametric trend testing. They showed that pre-whitening is not effective in eliminating the effect of serial correlation on trend detection through the Mann-Kendall (MK) test when sequences do in fact have trending observations. When sample size and magnitude of trend are large enough, they demonstrated that serial correlation no longer significantly affects the MK test statistics. A white signal is a signal of independent samples, and a pre-whitened filter is used to change a signal that is not white into one that is. In other words, these are complex ways to deal with serial correlation, but one can find the "pre-whitening" terminology used in certain complex examples of ways to analyze water quality data, often associated with complex predictive models and/or Monte Carlo simulation. Yeu and Wang (2002) also showed that:

Removal of positive AR(1) from time series by pre-whitening will remove a portion of trend and hence reduces the possibility of rejecting the null hypothesis

while it might be false. Contrarily, removal of negative AR(1) by pre-whitening will inflate trend and leads to an increase in the possibility of rejecting the null hypothesis while it might be true. Therefore, pre-whitening is not suitable for eliminating the effect of serial correlation on the MK test when trend exists within a time series.

Pyper and Peterman (2011) clarified that the choice between using pre-whitening or differencing when dealing with autocorrelation is not always easy and should be done very thoughtfully and with attention to the time scales of variation: In their abstract, they stated that:

Autocorrelation in fish recruitment and environmental data can complicate statistical inference in correlation analyses. To address this problem, researchers often either adjust hypothesis testing procedures (e.g., adjust degrees of freedom) to account for autocorrelation or remove the autocorrelation using pre-whitening OR first-differencing before analysis... Our Monte Carlo simulations show that pre-whitening and especially first-differencing decrease power in the common situations where low-frequency (slowly changing) processes are important sources of covariation in fish recruitment or in environmental variables. Conversely, removing autocorrelation can increase power when low-frequency processes account for only some of the covariation. We therefore recommend that researchers carefully consider the importance of different time scales of variability when analyzing autocorrelated data.

Is the autocorrelation first order autocorrelation only?

The PACF correlogram is the best one to answer this question (Box and Jenkins, 1976). If the first lag is the only one to extend beyond the 95% confidence line, it is first order autocorrelation.

In first order autocorrelation, the second measure is correlated to the first, the 5th to the 4th, etc., and in very frequently measured water quality data, there is usually at least first order autocorrelation so the effective sample size should be reduced compared to n at least this much (if not more, if serial correlation is more persistent).

Why Do I Need to Know if First-order Autocorrelation is the Only Kind Present?

If it is first order autocorrelation only, one can use fairly simple equations to calculate an 'effective sample size' that can be used in statistical equations instead of the larger sample size of data points. This is an important substitution for autocorrelated data points that are not independent.

Knowing the order of autocorrelation also helps identify advanced models that fit the data the best. See separate discussion of ARIMA models below.

What is Effective Sample Size and How is it Used?

Serial correlation leads to underestimates of uncertainty (and thus overestimates of statistical significance) because when residuals (or raw values for that matter) are not independent, the

effective number of degrees of freedom can be much smaller than the sample size would indicate. In principle, one should be able to account for this loss of degrees of freedom by calculating the "effective" sample size--that is, the number of independent measurements would be functionally equivalent to your set of non-independent measurements. You then use this "effective" sample size, N_{eff} , in place of n in calculating standard errors and performing significance tests (Kirchner (2001)).

In other words, the effective sample size is the equivalent number of independent observations that would provide the same expected variance as the autocorrelated samples, if the variables were independent rather than autocorrelated. There are a large number of potential equations that could be used to determine effective sample size, some more complex than others (Bayley and Hammersley, 1946).

Potential Complications with Effective Sample Size

Like many autocorrelation issues, there are potential complications to keep in mind when considering effective sample size. For example, positive dependence means that the individual observations include information present in other observations, so that the effective sample size, say n' , is less than the number of observations, n . Negative spatial dependence can have the opposite effect, making the effective sample size greater than n . The overall effect of dependence on the effective sample size depends on the balance between positive and negative over all possible pairs of observations in the study. This characteristic makes the issue of special importance to ecologists, who often encounter data with waves of alternating high and low values arising from patchy phenomena, producing autocorrelation that fluctuates with distance between positive and negative (Dale and Fortin, 2009).

Once one has determined from PACF correlogram hints that first order autocorrelation is only the kind present, one can next determine effective sample size before doing hypothesis testing, or calculating confidence intervals, or many other complex statistics. For one thing, the simplified equations above only deal with first order (between lag 0 and lag 1 only) autocorrelation. The calculations for more severe autocorrelation are more laborious and probably not amenable to beginners (Reiher and Huzzen, 1967).

A precursor to most of the effective sample size equations was provided by Lattenmaier (1976). Statistical power is extremely important in the trend detection problem, since it gives the probability, at a fixed confidence level of trend detection. The power of a given test will vary with record length, trend magnitude, and the marginal probability distribution of the data in an independent time series. For dependent time series, power will also vary with the form of the dependence of the observations. The method presented by Lattenmaier (1976) for detecting trends serially correlated data (a dependent time series) was based on establishing equivalence between power curves for dependent and independent observations. A method of summarizing the three independent variables (record length, trend magnitude and standard deviation for data with a Gaussian marginal distribution) into a single index is first established for the independent case. This approach is then extended to dependent time series. The method is basically a t test applied to both two sample step trends and to linear trends. Linear trends in this case refers to the scenario where mean levels vary in a linear manner over the period of record.

Relatively Simple Equation Options for Effective Sample Size

Several simplified equations found to date for first-order autocorrelation only are listed below in general order from simplest to more complex:

Kirchner (2001) stated that for first order autocorrelation, the effective sample size N_{eff} (some use n') can be roughly estimated (based on WMO Technical Note 79, 1966): with the following equation.

$N_{\text{eff}} = n * [1 - \rho / (1 + \rho)]$. The drawback to this approach is that one does not usually have the true (underlying) population ρ but instead has a sample [correlation coefficient](#), r , which underestimates ρ and therefore leads to overestimates of effective sample size.

Meko (2009) used r instead of ρ in basically the same equation:

When the times series exhibits first-order autocorrelation (dependence on lag-1 only), the governing process is first-order autoregressive, or Markov. Computation of the effective sample size requires only the sample size and first-order sample autocorrelation coefficient. The “effective” sample size is given by: $n' = n * (1 - r_1 / (1 + r_1))$, where N is the sample size, N' is the effective sample size, and r_1 is the first-order autocorrelation coefficient. For example, a series with a sample size of 100 years and a [first order](#) autocorrelation of 0.50 has an adjusted (effective) sample size of $N' = n * (1 - r_1 / (1 + r_1))$. So if $n = 100$, and $r_1 =$ correlation coefficient at lag 1 = 0.5, then $(1 - r_1 / (1 + r_1)) = (1 - 0.5 / (1 + 0.5)) = 0.5 / 1.5 = 0.333$. and n' (100) times that = $0.333 * 100 = 33$ = adjusted sample size if the only serial correlation is at lag 1.

This same equation was also suggested by Chandler and Scott (2011). They also state that more complex methods may be better since explaining a justification for the simplified equation fully is difficult and since its implications may not always be fully clear. It is relevant to first order autocorrelation only and is based on the notion that a simple way to deal with autocorrelation issues in regression models is to adjust the (wrong) standard errors after fitting a model using least squares. When adjusting standard errors in regression models, it would preferable to start with considering the exact expression of the covariance matrix of the parameter estimates, given with much more complex equations (Chandler and Scott, 2011).

There are several other more complex but still relatively simple first order effective sample size equations discussed below, so it is easy enough to estimate effective sample size by at least two different lines of evidence. If one is drastically different than the other, an input or calculation mistake may have occurred in one the calculations.

For lag 1 autocorrelation (first order autocorrelation) Helsel (2009) suggested a slightly different and a bit more complex version while citing some of the same documents, but still relevant only to [first order](#) autocorrelation between lag zero and lag one:

When dealing with hypothesis tests, first the values compared and/or sample sizes are changed to their 'effective' values, the values representing the number of independent pieces of information rather than the number of (correlated) measurements. This idea is not new, and was presented for water resources data by Matalas and Langbein in a 1962 paper of the Journal of Geophysical Research (vol 67, no 9). The effective sample size N_{eff} will be less than the observed sample size N when serial correlation is present. For the serial correlation coefficient r , the effective sample size is $N_{\text{eff}} = N / \{ (1+r)/(1-r) - (2/N)*r*((1-r)^{2N})/(1-r)^2 \}$, where $**$ represents 'raised to the power'. The value between the brackets will be greater than 1 when r is greater than 0. Using the effective sample size instead of the observed sample size is a simple way to adjust for serial correlation in hypothesis tests and confidence interval calculations. For regression applications, more formal time series methods are warranted.

That last sentence is a reminder that there are various more sophisticated approaches than most discussed herein and that some of these more sophisticated approaches to dealing with serial correlation may be superior to some relatively simple approaches discussed herein.

For the average spatial autocorrelation coefficient r , the number of independent samples $N^1 = n / [(1+r)*(n-1)]$. It was clarified that N^1 is the number of independent samples required to give the same variance as some number of n having a mean inter-correlation coefficient of r , and N^1 is the number of equivalent independent samples which would provide as much information about the variance. It can be shown that as n becomes large, N^1 approaches $1/r$. The example given for 100 stations and an average correlation of 0.25, was that $N^1 = n / [(1+r)*(n-1)] = N^1 = 100 / [(1+0.25)*(100-1)] = 3.9$, so only about 4 independent stations would be needed to provide as much information (say about the true variance) as the original 100 (Gordon et al, 2004).

What Equation Is Used in a PACF Correlogram?

The rough approximation equation is universally used in a partial ACF (PACF) correlogram. The 95% confidence interval lines are usually placed at $2/\text{square}/\text{square root of } n$.

Box and Jenkins (1976) recommend using this equation only in a PACF correlogram and to instead use the large lag standard error equation in an ACF correlogram

What Insights Do My ACF and PACF Correlogram Give Me About Model Fitting?

There is always a model, but whereas early EDA line plots and histograms hint whether or not a distribution may follow a normal distribution model, ACF and PACF correlograms give advanced modelers more complex insights.

Most of the following is about complex predictive models, beyond standard EDA as practiced by beginners, so only a few basics will be mentioned for those who want to learn more. Whereas

simple trend tests such as those of Kendall family might be part of advanced EDA, most of the advanced models mentioned just below would be part of final data analysis.

The main problem in classical Box-Jenkins time-series modeling approaches is trying to decide which Autoregressive Integrated Moving Average Model (ARIMA) specification to use -i.e. how many autoregressive (AR) and/or moving average (MA) parameters to include. This is what much of Box-Jenkins (1976) was devoted to the "identification process. It depended upon graphical and numerical evaluation of the sample autocorrelation (ACF) and partial autocorrelation function (PACF) as depicted in ACF and PACF correlograms (Morrison, 2011).

Picking out the correct specifications is as much art as science. Among the rules of thumb one uses when looking at ACF and PACF correlograms while practicing this "art" include the following (Garson, 2009):

Autoregressive models. AR models are indicated when PACF cuts off sharply at lag x but ACF declines slowly. To determine tentatively the value of p , look at the PACF plot and determine the highest lag at which the PACF is significant.

Moving average models. MA models are indicated by a rapidly declining ACF and PACF. If the ACF does not decline slowly but rather cuts off sharply at lag x , this suggests setting $q=x$, thereby adding a moving average component. If autocorrelation is negative at lag-1 then this also indicates the need for an MA (q) term higher than 0.

Hints from correlograms that would help one with the art of picking the correct Autoregressive Integrated Moving Average Model (ARIMA) pattern include the following:

ARIMA ($p,0,0$): ACF is spiked at lag p and declines toward 0. PACF is spiked at lag 1 to lag p .

ARIMA ($0,1,0$): Random walk model. The only effect is a non-seasonal [differencing](#) to remove a linear trend. ACF is either constant or is balanced between positive and negative. PACF is spiked only at lag 1.

ARIMA ($1,1,0$): First-order autoregressive model. There is non-seasonal differencing to remove a linear trend, and lagging the dependent variable by 1.

ARIMA ($0,1,1$): Simple exponential smoothing model. There is non-seasonal differencing to remove a linear trend, and lagging shock effects by 1.

ARIMA ($0,0,p$): ACF is spiked at lags 1 to p , declining sharply thereafter to 0. PACF is spiked at lags 1 to p , declining more slowly toward 0.

ARIMA ($p,0,q$): ACF and PACF both decline slowly toward 0. PACF declines erratically due to shock effects.

ARIMA ($1,1,1$): A mixed model. *Warning:* normally one does not include both autoregressive effects and moving average effects in the same model because this may lead to over-fitting the data to noise, and may reduce the reliability of interpretation of the significance of individual components in the model.

As a rule of thumb, one may wish to start with $p=1$ and/or $q=1$ and then increase the p and/or q values if the ACF and PACF for the residuals display spiking

Similar rules of thumb were given by Coghlan (2011a) in her discussion of using R software for time series model fitting. There are many other related guidance documents on the Internet (Google “using R for time series”).

Are There Also Tests for Autocorrelation?

Are there tests that would be useful as another line of evidence in addition to indications from correlograms? What other EDA tool(s) do I use to assess whether or not autocorrelation may be a big enough problem to require special different than normal steps in subsequent data analysis?

The Durbin Watson test can be used to test for the presence of first order autocorrelation (Garson, 2009; Helsel and Hirsch, 2002). However, during early EDA one does not yet know if first order autocorrelation is the only pattern present and it is more common to first look at autocorrelation patterns with correlogram plots rather than to jump too quickly to relying only on a first order serial correlation statistic such as Durbin Watson. Some say not to use Durbin Watson with a lagged dependent variable

The other reason for not relying on the Durbin Watson test alone is that it can be inconclusive and the result can be reject, fail to reject, or the test is uncertain. It should just be considered one line of evidence to go along with hints from correlograms. All of these methods are approximations, so (like trend tests), it is best to use more than one and see if they agree.

If autocorrelation in a series is first-order autocorrelation, it means the dependence is present at lag-1 (the very next measure) only. In other words, the dependence is only to the preceding data point in either time or space, rather than to two or more preceding data points. In first order autocorrelation, the governing process is said to be a first-order autoregressive, or Markov pattern (Meko, 2009). Since the ordinary Durbin-Watson statistic only tests for first-order autocorrelation, in SAS software, the Durbin-Watson statistics for higher-order autocorrelation are called generalized Durbin-Watson statistics (SAS, 2011).

If one is collecting water quality data with sondes or other auto-loggers every 15 minutes, one will almost always have considerable temporal autocorrelation and probably it will continue over several successive measurements (lags). Thus one will have more than only first order autocorrelation. If dealing with very frequent measures, first read Irwin (2010), an introduction on the Internet.

Another test option would be to use the `bgtest` command in R software to calculate a Breusch-Godfrey test.

What Autocorrelation EDA Tools Work For Data Not Regularly Spaced?

Can I use correlograms to get a first estimate of autocorrelation if data is NOT regularly spaced in time or space?

No, for irregularly spaced data and spatial issues, **variograms** tend to be the initial EDA plots of choice. ACF and PACF correlogram methods could only be for spatial data if the spacing is regular. However, they seem to be seldom used for spatial work perhaps partly because the spacing is often not regular.

If the spacing in either time or space is not regular, spatial variograms are often used (Chandler and Scott, 2011). Variograms are frequently used in spatial to assess autocorrelation related to three dimensional issues or soil type/land use strata. They could also be used in irregular or three dimensional spacing in an aquatic setting, where one might be dealing with latitude, longitude, and depth.

In GIS work, a Moran statistic is sometimes plotted on map subdivisions and used as an alternative to variograms in spatial analyses. At least one freeware program GeoDa is designed specifically for map based spatial EDA. It makes it easy to plot different Moran results on map subdivisions to help visualize clustering and to begin to understand spatial autocorrelation magnitude using Moran statistics (Anselin, et al. 2007).

Spatial autocorrelation should not be thought of as all bad, since having stations close together can be desirable from several angles, including 1) better estimations of missing data between stations than one would get from rough interpolations from widely spaced locations, and 2) more accurate estimates of regional means. For contrast, serial correlation (temporal scale rather than spatial) can reduce accurate information needed to estimate an annual mean if seasonality is not considered (Gordon et al (2004).

What EDA Helps Identify a Longer Term Trend In Spite of Shorter Term Cycles?

Cycles can be seen in simple line plots, smoothed trend lines, correlograms, and variograms. These are discussed separately herein. One might first simply describe the pattern seen in a line plot based on rough (not smoothed) line plots compared to a smoothed line plot.

As one advances into later EDA steps, it is common to begin to separate trends and cycles from the other patterns. See discussion of [decomposition plots](#) in Level Three EDA discussions.

That could be followed up eventually with a [trend analysis on medians](#) of each cycle as well as more complicated trend analyses.

One recommended approach might be use at least three lines of evidence, a line plot or scatter plot with a fitted line or smoothed trend line, as well as a seasonal Kendall trend test and a quadratic trend test, then report the trend hints one gets from all three. This multiple lines of evidence approach is used in Maryland and British Columbia and goes along with the notion that no single trend analysis method is ideal for data sets, so use several and see if they all conclude the same thing.

What Are Some More Advanced EDA Tools for Cyclical Patterns?

When regular cycles are present, more advanced periodograms are also sometimes constructed. Periodograms can be useful when the data contain oscillations at specific frequencies.

A periodogram can be thought of as a linear filter of the data. However, be aware that different authors and different programs suggest slightly different equations for periodograms. ACF correlograms and periodograms generally contain a similar amount of information regarding the length of cycles (Chandler and Scott, 2011).

Therefore; many just construct ACF correlograms instead, since they add information of autocorrelation. Several of the more complete EDA summaries (NIST, 2006b; EPA, 2009b) and several prominent environmental statistics textbooks [including Helsel and Hirsch (2002), McBride, 2005, and Gilbert (1987)] don't cover periodograms. Thus, we will not spend much additional space on periodograms either. However, as part of better documenting cycles periodograms could be constructed as part of Level two or Level three EDA.

Some software programs (including but not limited to Statgraphics Centurion, Systat, and Statistica) make drawing periodograms relatively easy.

There are many complex steps that some use in addition to or instead of periodograms. Again, similar clues about cycles will typically show up in time series line plots, smoothed scatter plots, correlograms, and other EDA plots, so one need always draw a periodogram.

More complex EDA steps such as spectral analysis and spectral plots tend to be beyond our beginner target audience. For those who want to learn more, some basics were discussed by NIST (2006) and Kittel (2009). For an introduction to later trend analysis using either sine-based models and more complex Generalized Additive Models (GAMs), steps that could logically follow EDA at times, see Sanger et al. (2002).

Is There A Simple Way to Handle Spatial and Temporal Autocorrelation Together?

Yes, there is one way available to beginners, though statistical and modeling experts tend prefer more complex methods that utilize all the data.

However, even some true experts who have tried some the very complex modeling solutions to factor in both time and space autocorrelation at one time have sometimes decided the added complexity may require additional data (that may not be available) or that the marginal increase in information may not justify the added effort. In practice, it may be more pragmatic to use simpler methods that are robust to spatial autocorrelation such as re-sampling methods which collapse the data from more frequent to less frequent values (Jeffrey Morissette, USGS, Personal Communication, 2011).

So one option is to collapse the data in a relatively simple two-step re-sampling method:

First use a large lag standard error correlogram as discussed above to find out how far apart data needs to be in time to be independent. If the sites are in regular units of space, one can next use the same kind of correlogram to find out how apart sites need to be in order to be independent. If not evenly spaced, use the variogram or Moran's I remedies discussed above.

Then re-sample out of the larger data set choosing only data points that meet the criteria of being far enough in both time and space so that autocorrelation is not present in either time or space.

One should thoughtfully decide which analyses make the most sense and one of the most commonly used water quality text books recognizes the validity of simpler approaches as one option.

For example, one might group the data into time periods (e.g., weeks, months) and compute a summary statistic for the period such as a time-weighted mean or median, a volume-weighted mean or median, and then use these summary statistics in the regression. This should only be done when the sampling frequency has remained unchanged over the entire period of analysis. Alternatively, one might sample from the data set. For example, if the data set is quite large and the data are closely spaced in time (say less than a few days apart), then one option for certain subsequent statistical procedures is to simply not use some of the data in a regular pattern. The dependence that exists is an indication of considerable redundancy in the information, so not a great deal is lost in doing this. Simpler approaches such as these can be an acceptable way to handle serial correlation without having to resort to Box and Jenkins (1976) modeling approaches to handle serial correlation in much more complex ways (Helsel and Hirsch, 2002).

However, re-sampling a data set to get a smaller data set as one strategy for data analyses does not mean one should discard the raw data totally. Always keep all raw data, it may have other uses and there may be a later need to analyze it in other ways. .

If sub-sampling is done in ways that do not separate samples enough in time and space, difficult issues can arise not only related to autocorrelation but also to potential pseudoreplication issues (van Mantgem and Schwillk, 2009).

Again, advanced statisticians or modeling experts tend to favor more complex modeling solutions, partly due to worries of a loss of information associated with re-sampling (collapsing) to a smaller dataset. These solutions tend to be beyond our beginner target audience.

Should I Calculate Confidence Intervals as Part of EDA?

These are less often plotted during early EDA, where box and whisker plots are more often plotted and are often more immediately helpful. One could plot them if one had a specific reason to emphasize means. Just don't line up confidence intervals about a mean to see if there is overlap and use that as a pseudo-hypothesis test about a difference in means. A real test can give a different result. Means tend to be overused, so consider a nonparametric confidence interval about a median instead.

For confidence intervals about the mean, EPA (2009) includes the following caution “Using a back-transformation to bracket the location of the true arithmetic population mean will result in an incorrect interval” and suggests that “a back transformation of this type be considered a confidence interval around the geometric mean of the raw (untransformed) data. For the lognormal distribution, the geometric mean is equal to the median, but it is not the same as the arithmetic mean.... “It is far better to have an approximate answer to the right question than a precise answer to the wrong question — John Hauser.”

For contrast, there is no back transformation bias introduced in constructing a nonparametric confidence interval around the median, aka the 50th percentile (Kirk Cameron, MacStat Consulting, Personal Communication, 2011).

Should I Calculate Prediction Intervals or Tolerance Intervals as Part of EDA?

Tolerance intervals are used at times **to help define background levels**, and during level two EDA one might recheck early preconceptions about logical data categories, background or unimpacted reference areas, and potential explanatory variables, in the EDA spirit of “letting the data do the talking.”

Both parametric and nonparametric [control charts](#) are often based on tolerance intervals, so many using control charts are essentially using tolerance intervals.

Other than that, tolerance intervals are less often done as part of EDA, but there is no reason why one could not use them more often, especially if one calculated a different and better tolerance interval, which would include the % confidence in the length of the interval. These could be calculated more often as part of level two EDA to answer specific questions, for example related to background (control) levels. Since tolerance and prediction intervals are often misunderstood and/or confused with confidence intervals, and since we use the term tolerance interval herein, a bit more detail about them is provided as follows.

A key point to understand is that prediction intervals and tolerance intervals are calculated the same way, but are just used for different purposes. In fact, often the identical statistic can either be used as an upper prediction limit or an upper tolerance limit, with only a difference in statistical interpretation. A tolerance interval is about a range of values covering a certain proportion of the population, whereas the emphasis with a prediction interval is about our confidence that the very next measure will fall within that same range (EPA, 2009).

Tolerance intervals are often confused (even in some reputable literature) with a confidence interval. Hahn and Meeker (1991), the definitive text book on intervals, use the phrase “statistical tolerance interval to contain a proportion of a population” to help avoid this confusion and make it clear that a tolerance interval is a type of enclosure interval that is meant to enclose the middle proportion of the population being sampled. For contrast, they use the phrase “confidence interval for a percentile” to make it clear that is what they are talking about when the focus is a percentile rather than to a two-sided tolerance interval meant to control the center of the distribution. Unless otherwise stated, herein we use the phrase tolerance interval to mean a

type of enclosure interval that describes the “statistical tolerance interval to contain a proportion of a population.

What are Measurement or Manufacturing Process QC Control Charts?

NIST (2006) discusses the use of control charts in EDA in general.

We will discuss simple control charts such as those used in quality control (QC) applications in this section. For more complex control chart EDA applications see separate sections below on [complex control charts](#) used to indentify differences from background levels or to identify trends.

The simplest control charts are based on tolerance interval concepts, but in the context of manufacturing products, tolerance limits tend to be used to determine whether individual products are within limits, whereas control charts are used to control a measurement process or a manufacturing process. That is the basic difference between tolerance limits and control limits.

Since there are multiple kinds of control charts discussed herein, when one uses the phrase control chart, one should make it clear which type one is discussing and how they are calculated. Common options include the following.

In the first case described in this section, control charts are also used related to QC and maintaining quality control relevant to every single data point.

Control charts are also used for outlier identification (see separate discussions herein on [outliers](#)).

Some use CUSUM or other complex control charts when looking for trends or when looking for differences from background. See discussion farther below under the topic of [complex control charts](#).

In other words, those using the phrase control chart should make it clear which level of concern is being discussed, and how the control limits of concern are calculated.

QC Control Charts: Controlling a Measurement or Manufacturing Process:

As part of controlling the measurement process, historically chemical labs have documented that they are keeping measurement processes “in control” by plotting control charts of duplicate or replicate precision QC estimates and flagging those that seem out of control based on multiples of either standard deviations or % differences (APHA, 1998). ” In one simple type of control chart, if too many “out of control values” beyond the standard deviation or % interval lines appear on a control charts, recalibration or other remedies are applied to the measurement process until the measurement process comes back into control. This is one way that a chemical lab can demonstrate that they are keeping the measurement process in control (within documented quantitative bounds).

The concept is similar to keeping manufacturing “under control. Historically no such efforts have been made in field measurements, though now that continuous monitoring sondes are being used

more often, this will be easier to do and will probably be done more frequently (see AMS discussion below). Simple control charts of this type could be considered Level One EDA.

For lab measurements, NIST (Taylor and Kuyatt, 1994) and ISO (ISO, 2008) recommend the measurement uncertainty be estimated and reported for all data points using multiples of standard deviations. The Standard Methods book recommends basically the same thing using slightly different terminology (APHA, 1998).

For field water quality measures of variables such as pH and temperature, in the NPS we have recommended using a similar thought process. In this case, one plots slightly more complex measurement uncertainty intervals over longer periods of time in control chart format to keep track of changes in alternative measurement sensitivity (AMS). For small sample sizes, both NIST and ISO suggest one can use two-sided t-values to determine the K factor to multiple times the standard deviation to calculate expanded uncertainty. In the NPS, we use the same method in suggesting a method to calculate special case expanded uncertainty (called alternative measurement sensitivity or AMS) using a sample size of 7 and 99% confidence. These two special case values were used as analogous as possible to MDL detection limits, but using the two-sided t-value as the K factor (instead of the one sided t value used in MDL calculations) to be consistent with NIST/ISO/APHA expanded measurement uncertainty (Irwin, 2008, Thoma et al. 2011). Two sided values are logical for uncertainty because unlike detection limits we are interested more in how different two measures would have to be to be truly different from each other and we are not interested in the detection limit issue of different from zero.

If one has a good estimate of bias for each logical batch of measures, Phillips et al. (1997) and Thoma et al. (2011) suggested a way to construct similar uncertainty intervals adjusted for bias for that batch.

In long term monitoring, it would be a good idea to plot uncertainty (or AMS, or NPS special case of expanded uncertainty) over time as part of EDA plotting. Though this has not been done much in the past, this would allow one to see if there are changes over time in measurement uncertainty/sensitivity.

Why would such changes matter? If measurement sensitivity (uncertainty) becomes poor, it will artificially inflate variability, which in turn will artificially inflate EDA summary statistics such as standard deviations, IQR distances, variances, and range. This will also inflate the magnitude of confidence intervals about means or medians and make it more difficult to detect true trends over time.

Poor measurement sensitivity should not be allowed to be a major factor in comparisons with water quality standards or other benchmarks. This is why it is appropriate to check not only QC performance for precision and bias, but also for sensitivity. As noted by Chandler and Scott, 2011, one should transparently indicate sample sizes used to calculate various things and uncertainty of each data point when such uncertainties are non-trivial compared to real variability (Chandler and Scott, 2011).

In other words, not only is there uncertainty about summary statistics such as means or medians (estimated with confidence intervals) from different environmental samples, but there is also an uncertainty interval about each single data point (estimated with uncertainty intervals).

One could also construct control charts based on alternative ways to estimate uncertainty. Harmel et al (2009) suggested that measurement uncertainty be estimated and reported for all data points but instead of using standard NIST/ISO/APHA methods, they used a simpler combined % method to get a type of combined uncertainty with root mean square errors (RMSEs) methods after considering multiple uncertainty sources. The authors stated that whereas errors were assumed to represent the 0.68 significance level in previous RMSE work, the 0.9999 significance level (3.9 standard deviations) was instead assumed in their work. The authors also concluded that

Although the uncertainty in sample collection is often ignored, it can be the dominant source in environmental investigations. This was certainly the case for sediment and total N and P sample collection in the present study. Collection of representative sediment and total N and P samples is quite difficult because streams can have substantial cross-sectional and vertical gradients in particulate-associated concentrations. In contrast, collection of dissolved N and P samples is quite easy since these constituents are typically uniformly distributed within the channel... This difference was exemplified by greater sample collection uncertainties for sediment and total N and P concentrations, which ranged from 12 to 26%, than for dissolved NO₃-N and PO₄-P concentrations, which ranged from 6 to 17%.

Level Three EDA

Although some of the EDA tools discussed below tend to be more complex than would easily be understood by our beginner target audience, one sees many of these topics discussed in other EDA summaries, so it seemed imprudent to leave them out totally. Therefore, the discussions below summarize a few basics and also provide some leads for those who want to learn more.

In the spirit of continuing EDA in more complex ways as long term monitoring data accumulates, there are a large number of relatively complicated EDA plots explained with example plots on the Internet by NIST (2006). We will not list them all here and some are used more than others for field environmental data. A few of the more complicated plots are 4-Plots, 6-Plots, Bihistograms, and Weibull Plots. Spectral Plots can suggest random data, or a strong autocorrelation and autoregressive model, or a sinusoidal model.

TYPICAL QUESTIONS TO BE ANSWERED WITH LEVEL THREE EDA TOOLS:

How Do I Use Decomposition Plots in EDA?

Do line plots, scatterplots, box and whisker plots, or smoothed trend line plots suggest there are regular or irregular up and down cycles in the data? Do the plots also suggest trends? Often the

early plots suggest both. Box and whisker plots are sometimes used to help portray quarterly or seasonal patterns and they can be an early hint that decomposition would be helpful.

In water quality data, one often encounters regular up and down cycles, both over very short diurnal cycles and longer term seasonal patterns. Sometimes there are even longer term climate patterns. Plotting frequently collected data in scatterplots of the variable vs. time sequentially with different degrees of smoothing will begin to show these patterns, and hints of them will also show up in more advanced EDA tools such as correlograms. Keep in mind that a curved pattern over a shorter period of time may turn out to be part of a cycle when longer time periods are monitored. Also keep in mind that a severely smoothed pattern will get rid of shorter cycles, such as the diel cycles.

The different patterns in the data can obscure each other. Trends and cycles can be difficult to disentangle from autocorrelation patterns (Chandler and Scott, 2011). In EDA decomposition steps, one often tries to separate the cyclical patterns and trends from remaining patterns with **decomposition plots**. These help one visualize the different patterns more easily. In later steps, one may need to de-trend and de-seasonalize data before proceeding to longer term model-based trend analyses and autocorrelation analyses.

The following is repeated for emphasis: Do not use de-trended data in later trend tests. Sometimes one also identifies important short term events and include additional differencing is done as part of decomposition.

The following example of a decomposition plot shows regular seasonal cycles as well as other important patterns (Jassby and Cloern, 2010):

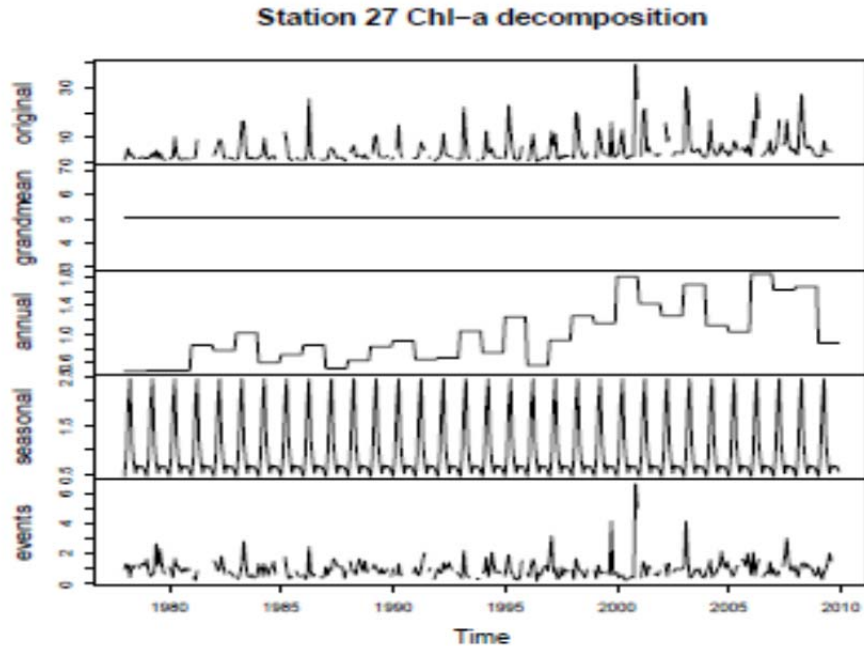


Figure 10: Multiplicative decomposition of chlorophyll at Station 27 in San Francisco Bay.

One can find similar decomposition plots in most time series summaries Coghlan (2011a) provided a similar plot to the one above in her discussion of using *r* software for time series tasks. Like the plot above, the top line plot was the original data, but the others below it included trend, seasonal, and random patterns.

Chandler and Scott (2011) summarized how to use LOWESS smoothing in nonparametric seasonal-trend decomposition (STL) tasks to decompose signals into seasonal components vs. trend components vs. remaining irregular component (residuals).

Sanger et al. (2002) used complex Generalized Additive Model (GAM) methods to deal with extremely variable data that was complicated by both tidal and regular diel cycles.

Decomposition Software.

Most statistical software packages have decomposition options. For example, Minitab and SAS software have easy decomposition routines.

In *R* software, patterns in time series of water quality data are often looked at in the [wq](#) package, where the [decompTs](#) function decomposes a time series into a long-term mean, annual, seasonal and "events" component (Jassby and Cloern, 2010). Coghlan (2011a) provides a discussion of decomposition and other times series tasks in *r* software.

Can Relatively Simple Trend Tests be Part of EDA?

Yes. Although trend tests are typically thought of as confirmatory methods for final data analysis, the simpler trend tests (such as those in the Kendall family of tests), are easy enough to do as an additional line of evidence when doing EDA in long term monitoring.

What Should Be Kept in Mind About all Trend Tests?

No single trend test or plot is ideal for every scenario, but (especially if the trend is not clearly linear and/or one way only) if one does a couple of different types of trend tests and they both conclude the trend was up, and if simple line plots, scatter plots, or smoothed trend line plots also hint that the trend was up, then one can have relatively strong confidence that the overall trend was up (Robert Hirsch, USGS, Personal Communication, 2011).

One should always plot simple EDA time series plots in addition to more complex trend tests. If one makes a mistake with the complex method, the simple plots may help alert one to the problem.

What Are Seasonal Kendall (SK) Tests?

Various iterations of SK tests have the most commonly used trend tests for water quality data (Griffith et al. 2001). These have been mentioned briefly here and there already in this document but due to their importance are discussed in more detail here.

There is more than one variety of SK test. The seasonal Kendall trend test that includes an adjustment for serial correlation (serial dependence) within seasons (SKSD) is a modified SK test (Hirsch and Slack, 1984). Hirsch and Slack (1984) mentioned a power issue in their description of the SKSD test: The SKSD sacrifices power compared to the original SK test, but the SK test can lead to seriously wrong significance conclusions if seasonal or monthly scale serial correlation is present. These are tradeoffs to consider when choosing a test.

SKSD tests are also sometimes referred to as modified Seasonal Kendall Tau tests. The Seasonal Kendall's Tau assumes that data are serially independent - i.e., that values are not determined in whole or in part on the previous state in the sequence. To compensate for serial dependence in a data series, Hirsch and Slack (1984) proposed the (SKSD) modification to the seasonal Kendall's Tau that takes into account any co-variation between seasons in a data set. Either version of the Seasonal Kendall test is most appropriate if trends are consistent throughout a year. For example, a negative trend for six months followed by a positive trend of six months would yield a test statistic indicating zero trend. The two tests do not measure the size of any trends, only the direction (Regnier and Ryan, 1999).

Harcum et al. (1992) tested for serial correlation presence at lag 1 and if present, simply used the SKSD test.

Nonparametric tests such as the SK or SKSD tests are more appropriate than parametric tests if one is unsure about whether or not the distribution is normal (which is hard to decide based on small data sets). They are also most appropriate for data sets with missing values or censored

data, or any time that one decides that extreme values should not be given undue influence in the outcome of a test (Hirsch and Slack, 1984; Helsel and Hirsch, 2002).

For the SK test as done in ESTRENDS software, the assumptions of the SK test are that the sample data represent the target data and that the data for each season has the same distribution and no serial correlation. The seasonal Kendall test for serially dependent data (SKSD) is a bit different and it assumes the possibility that the data can be serially correlated, arising from a process no more complicated than can be described by a stationary autoregressive moving average [ARMA (1,1)] model as described by Hirsch and Slack (1984). As such, it is designed to analyze random samples at regular intervals and not systematic high-frequency samples. The problem comes from the higher-order ARMA processes in high-frequency data. We feel pretty comfortable that monthly water-quality sample data can be described as an ARMA (1,1) model process, but would not be so comfortable with data collected much more often. An ARMA (1,1) process is one in which the current value depends only on the previous value plus some noise, or random variation. (David Lorenz, USGS, Personal Communication, 2009).

By design the ESTRENDS and Time Trends software will likely give different results for the Seasonal Kendall test where there are multiple values per season. So to interpret results in an optimal way, it is important to understand what the software being used is actually doing. Overall, I like the options one has in Time Trends software and consider it to be my go-to water quality trend analysis program now, and it has the added advantage of being free (Jeff Runde, Crater Lake National Park, Personal Communication, 2011).

The wq package in R software offers two nonparametric tests, the Mann Kendall Test and the Regional Mann Kendall test plots, and also offers a routine to determine if the seasonal Kendall test might be warranted given the amount of missing data present. The Seasonal Kendall test is not informative when trends for different months differ in sign and doesn't work well with much missing data. The function 'seasonTrend' in the R wq package enables visualization of individual monthly trends and can be helpful for, among other things, deciding on the appropriateness of the Seasonal Kendall test. When aggregating monthly series to produce an annual series for trend testing, there is a utility function tsSub that allows subsetting the months beforehand (meanSub is actually more efficient when aggregation is the goal). It can be useful for avoiding months with many missing data, or to focus attention on a particular time of year (Jassby and Cloern, 2010).

How Do I Confirm A Monotonic Linear Trend?

If scatter plots and/or smoothed trend lines and [Q-Q plots](#) seem to hint there may be a linear and monotonic trend, what do I do next to analyze for trend?

Congratulations, this is a rare event in environmental data, and in this case one may be able to use the simple linear regression methods learned years ago in Statistics 101. However, one still has to deal with assumptions and potential complications like how to deal with lack of perfect normality, how to deal with autocorrelation. Resist the temptation to just transform the X or Y variables until you get a straight line, at least not without thinking through potential back transformation bias (Kittel, 2009; Koch and Smillie, 1986) and the problem that you may have changed to the question to be answered to something less relevant. To avoid pitfalls, most may

still be better off to supplement evidence from parametric linear regression lines of evidence with simple and different lines of evidence seen with a moving median smoothed trend line and other nonparametric lines of evidence, such as the linear seasonal Kendall or Mann Kendall Tests. These should probably be the go-to tests for linear trends rather than relying only on parametric ordinary least squares (OLS) linear regression methods.

EPA's free Software ProUCL version 4.0 offers lets one do both OLS and Mann Kendall tests in a manner that handles nondetects correctly (EPA, 2010). Time Trends software gives one several options for performing Mann Kendall or Seasonal Kendall Tests (Jowett, 2009).

The forthcoming version ProUCL4.1 of ProUCL can be used to perform One-way ANOVA, trend tests such as Mann-Kendall and Theil-Sen tests, OLS regression, and Time Series plots. Also, ProUCL perhaps is the most comprehensive software that has rigorous statistical methods for data sets with nondetect observations and multiple detection limits. For contrast, EPA's Scout software is an advanced software package that can analyze multivariate data sets (Principal component analysis, discriminant analysis) and has robust statistical methods, including Huber, Biweight, LMS, MCD, and others (Anita Singh, Lockheed Martin, Personal Communication, 2011).

What Simple Trend Test Confirms a Longer Term Trend In Spite of Shorter Term Cycles?

One way to approach a trend test that will handle the pattern to look for trends in medians, when each median spans a single cycle. Then, one can do a trend test on those successive medians rather than on the raw data. Read over the considerations and [cautions related to using medians in trend tests](#) first though.

The seasonal Kendall trend test as implemented in Time Trends software has an option for this approach if one chooses the option of using seasonal medians to represent the season. If there is only value in the season, one uses that instead of a median. If there is more than one value in a season, one can specify the option of using all the values in the season instead of a median in the SK test (Jowett, 2009) confirmed by Ian Jowett, NIWA, NZ, Personal Communication, 2011:

Where one looks for a trend in medians over the years only considering certain months or certain seasons, one is simply removing the seasonal pattern to allow one to focus more cleanly on the underlying long term trend. An example might be the scatterplot of CO₂ vs. time at Mauna Loa, one of the more famous and some say one the most important plots ever. This one shows regular short term cycles but there is still an apparent longer term trend in spite of the shorter cycles, see <http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm>. If there is linear trend in water quality based on very frequent sampling it might tend to look a bit like this due to regular seasonal cycles and the line might look more ragged due to diel cycles if sampling is done multiple times per day. This kind of pattern might prompt later decomposition steps in level three EDA before analyzing for trends and/or doing complex model fitting to counter autocorrelation.

How Do I Confirm A Non-Linear Trend Or Pattern?

If early Hints from time plots or simple EDA plots suggest a curved quadratic pattern, but the visual and/or curve fitting clues are somewhat inconclusive, deciding what to next can be complex.

If a quadratic pattern seems to be present, and the curve is not too drastic, and seems to be going mostly up or mostly down (say the lower half of a C) rather than totally reversing (as in a U or upside down U in the plots), there may still be a monotonic trend. In this case one relatively easy option is to use a linear Mann-Kendall trend test or (if seasonality is present), a seasonal Kendall test most commonly used for water quality data. These Kendall tests are not as strongly influenced by a moderate degree of non-linearity or outliers compared to an ordinary least square (OLS) linear regression, though they are still basically considered linear methods. This has some appeal though since it is relatively easy to do, and for borderline patterns not a bad thing to do **in addition to** more complex quadratic trend tests, which may require the help of a statistician.

If a pattern is totally reversing (as in a U or upside down U in the plots), there is not a monotonic (one way) trend, but there still may be a pattern that should be described. Some of these situations lend themselves to looking from several lines of evidence, as in the following examples:

Due to the commonness of curved trend patterns, the State of Maryland routinely reports both the quadratic and seasonal Kendall trend results for water quality data if they are significant after 10 years of data, presenting the results of both graphically on a map (Hall and Wazniak, 2006).

Likewise, Regnier and Ryan (1999) looked at trends in British Columbia using at least three lines of evidence: Simple time series plots such as line plots, a nonparametric and non-linear seasonal Kendall test for serial dependence, and a parametric quadratic non-linear model fitting method.

Both the Maryland and BC examples illustrate the need to look from different angles and to always supplement complex trend tests with simpler lines of evidence from simple line plots, scatter plots, time series lined up box and whisker plots, and/or smoothed trend lines evidence. This approach is consistent with the consensus view of many modern environmental statisticians, that if the pattern proved to be quadratic, one might later follow up with a quadratic model or other polynomial model trend tests. These can be complex, so beginners may need to consult statisticians.

How do I Use Control Charts for More Complex Purposes?

Morrison (2008) attempted a plain language summary (a difficult task) for various ways control charts are used. Although he doesn't use the words stationary or stationarity, in part he seemed to be suggesting using control charts as an alternative to hypothesis testing to determine if there is a trend or not. He briefly mentions plotting as simple graphs without the use of inferential statistics, but doesn't mention commonly used trend line smoothers such as moving medians, LOWESS, or LOESS smoothers. Although regression trend methods were discussed, there was no mention of nonparametric counterpart methods such as the seasonal Kendall tests or Mann

Kendall tests, the most common trend tests used for water quality data. In one comment, Morrison stated that: “Thus, assuming a normal distribution centered at the centerline, the control limits would encompass 99.73% of the distribution.” That is a tip off that Morrison is discussing a standard (3 sigma) parametric [QC control chart](#) similar to those used in chemical labs. This appears to use a calculation not that different from those used for crude tolerance intervals and crude prediction intervals. When a distribution is perfectly normal, the mean plus or minus three standard deviations contains 99.7% of the measurements. This tends to be more true in repeat measures in chemical labs, where distributions tend to be normal, than for our chemical or biological measurements made in outdoor environments, where distributions are seldom normal and sample sizes sometimes are small. Morrison mentions small sample size as a potential problem with hypothesis tests, but does not equally emphasize the same small sample size cautions related to control charts. Overall the Morrison summary is less complete and accurate than the others quoted above, especially EPA (2009), but Morrison’s summary has plain language appeal to beginners wanting an introduction to control charts.

Looking for Differences from Background with Control Charts

EPA (2009) summarized how to use control charts when comparing a possibly impacted site with an un-impacted reference site. They also explained the somewhat complex interrelationships between control charts, prediction intervals, and tolerance intervals. Rather than using the later, EPA recommended control charts as a somewhat more optimal method comparable to prediction limits, but which possesses some unique benefits. Control charts assume that the background mean is stationary over time. This means there should be no apparent upward or downward trend in the background measurements. However, when using control charts for background comparisons, a trend must not be present. A trend imparts greater-than-expected variation to the background data, increasing the baseline standard deviation and ultimately the control limit. The net result is a control chart that has less power to identify groundwater contamination

Looking for Trends with Control Charts:

One use of control charts is to look for hints of trends. Do indications from the control chart indicate a stationary or a changing process? The simpler types of control charts are similar to QC measurement process control charts already discussed under Level Two EDA. They are also somewhat similar to crude tolerance intervals and prediction intervals. However, there are also other approaches.

For example there are various non-parametric Cumulative Sum (CUSUM) control charts used to help identify trends (Gilbert, 1987; Manly and Mackenzie, 2003). An Australian software package for water quality analysis originally dating back to 1963 appears to use some simple trend analysis methods (Mann Kendall rather than seasonal Kendall), but evidently does include CUSUM and serial correlation analyses options (eWater CRC, 2005).

One can also use more complex parametric Shewhart-CUSUM control charts to allow compliance point data to be assessed graphically over time in a trend context. This kind of Shewhart-CUSUM control chart is a parametric procedure with no existing nonparametric

counterpart. It also has applications related to trends. More detail from EPA (2009): is summarized as follows:

For trend work, an advantage of control charts over prediction limits is that a control chart graphs the compliance data over time. Certain varieties can also evaluate gradual increases above background over the period of monitoring. Trends and changes in concentration levels can be easily seen since the sample observations are consecutively plotted on the chart. This provides the analyst an historical overview of the pattern of measurement levels. A gradual increase or trend is particularly indicated if the CUSUM exceeds the control limit but the compliance concentrations do not. The reason for this is that several consecutive, small, increases in x_i will not trigger the control limit, but may cause a large enough increase in the CUSUM. As such, a control chart can indicate the onset of either sudden or gradual contamination at the compliance point... The methodology for control charts assumes that the sample data are statistically independent. A control chart can give misleading results if consecutive sample measurements are serially correlated, (*i.e.*, temporally autocorrelated).

A less complete but more plain-language Internet-available introduction on the use of simpler tolerance interval-based control charts to identify trends was provided by Morrison (2008). Although he contrasts control charts with confidence intervals to estimate uncertainty around a parameter estimate and says that confidence intervals could be used in control charts; that would be unusual. Confidence intervals tend to be about means and not about individual data points or ranges of data points. A better discussion and comparison of confidence intervals, prediction intervals, and tolerance intervals in the context of control charts is found in EPA (2009).

How Do I Handle Nondetects in Advanced EDA and Trend Testing?

We have already discussed and defined nondetects in an early discussion of looking for patterns in [nondetects](#). When it comes to advanced EDA and trend work, more elaborate methods are needed, so the topic is summarized here.

Too many people still substitute half the detection limit or zero or some other fabricated value for each nondetect data point. This can cause problems not only in advanced EDA, but also in follow up trend work or confidence intervals. Even when only 5-10% or less of the data are nondetects one can compute biased confidence intervals. In fact, Singh et al (2006) concluded that incorrect confidence and other limits will result with as little as 10% nondetects. The ProUCL technical support guide (Singh et al., 2010), another USEPA publication, states "Based upon the results of the report by Singh, Maichle, and Lee (EPA, 2006), it is strongly recommended to avoid the use of the DL/2 method. Although the DL/2 method has been incorporated in ProUCL 4.0 for historical reasons (only), NERL-EPA, Las Vegas strongly recommends avoiding the use of DL/2 method even when the percentage (%) of NDs is as low as 5%-10%."

The best ways to handle nondetects have been summarized over recent years by Dennis Helsel. The basics were discussed in our discussion of [nondetects in early EDA](#) and in a free Internet available discussion (Singh et al., 2010). Additional details are in the next section.

Software for Handling Nondetects

The best ways to handle nondetects have been summarized over recent years by Dennis Helsel. Most of the Helsel (2005) methods are now available in R software, and updates of the R commands will be summarized in a 2011 revision of the book currently underway. In the meantime, here are some basics:

When there is more than one quantitative detection limit in a data set, the NADA survival analysis methods should be used to calculate summary statistics. In R and S-Plus, the summary command is typically used to briefly describe the characteristics of the data. In the NADA for R package, the censummary command fulfills the same role for censored data (Helsel and Lee, 2006).

Trend methods that include defensible handling of nondetects in R software have been developed by Dennis Helsel and are now available on the Internet at <http://practicalstats.com/nada/nada/downloads.html>. These include a parametric approach (MLE/TOBIT) equivalent to standard regression, and a nonparametric (Akritas-Theil-Sen or ATS) method, a censored analogue to the Mann-Kendall procedure that allows a slope and intercept to be computed for censored data in a fully nonparametric fashion. If there are considerable nondetects and they are at multiple quantitative reporting limits, the Akritas-Theil-Sen (ATS) Mann-Kendall nonparametric procedure that accommodates nondetects at multiple reporting limits would tend to be preferred (Helsel, 2005;).

For those not comfortable in working with R software, some other options are available:

ProUcl software includes many options that are generally compliant with many of the recommendations of Helsel (2005). In addition, the free Internet available discussion goes into some detail on certain methods discussed by Helsel. For example for the ROS MLE hybrid method, Singh et al. (2010) state that:

The hybrid estimates thus obtained are called robust ROS ML estimates (Helsel (2005). In this hybrid method, MLEs are computed using log-transformed data. However, Potential issues with the hybrid method include the following:

It should be noted that the performance of such a hybrid estimation method is unknown.

It is not known why this method is called a robust method.

The stability of the MLEs obtained using the log-transformed data is doubtful, especially for higher censoring levels.

The BCA and (%) UCLs based upon this method will fail to provide the adequate coverage for the population mean for moderately skewed to highly skewed data sets.

Meaning of the Word Robust in this particular case: The ROS ML estimates are 'robust' because it works well when data don't exactly follow the assumed lognormal distribution (Dennis Helsel, PracticalStats, Personal Communication, 2011).

Comparison between ProUCL and Scout software: The forthcoming version ProUCL4.1 of ProUCL can be used to perform One-way ANOVA, trend tests such as Mann-Kendall and Theil-Sen tests, ordinary least squares (OLS) linear regression, and Time Series plots. Also, ProUCL perhaps is the most comprehensive software that has rigorous statistical methods for data sets with nondetect observations and multiple detection limits. For contrast, EPA's Scout software is an advanced software package that can analyze multivariate data sets (Principal component analysis, discriminant analysis) and has robust statistical methods, including Huber, Biweight, LMS, MCD, and others (Anita Singh, Lockheed Martin, Personal Communication, 2011).

For those who already have Minitab software, some of the newer routines recommended by Helsel are available at <http://practicalstats.com/nada/nada/downloads.html>. The current Minitab updates, (version 2.9) include new and improved bootstrapping macros for censored data. These macros provide bootstrapped estimates of medians and means using several procedures. The advantage of bootstrapping is that confidence limits around the estimates can be provided without assuming that the statistic being estimate follows a distribution, such as the normal distribution.

When assessing trends, if one has daily or more frequent data that is serially correlated and there are relatively few nondetects or missing values, and they are not clustered in asymmetrical blocks of time within a month, one might also consider using the seasonal Kendall Test with Serial Dependence (SKSD) to test trends on monthly medians. This does not require R software usage as it can be done in other software, including [TimeTrends](http://www.niwa.co.nz/newr/tools#analysis) freeware, available at <http://www.niwa.co.nz/newr/tools#analysis> (Jowett, 2009). ESTREND software can be used by those with S-Plus to produce versions of the seasonal Kendall Trend tests. USGS staff developed S-ESTREND trend software, an updated version of an earlier program (ESTREND). The first S in S-ESTREND is a tipoff that it runs on an S-PLUS platform. Using S-ESTREND requires users to have Release 6.1 or later of S-PLUS for Windows (Sorenson, 2002).

How Do I Handle Spatial and Temporal Autocorrelation Together and Still Use all the Data?

We have already discussed the simpler ways to do this by [collapsing data to a smaller dataset that is not autocorrelated](#). However, advanced statisticians or modeling experts tend to favor more complex modeling solutions, partly due to worries of a loss of information with re-sampling to a smaller dataset. As one example, if the data tends to be very heterogeneous or clumped, using only a smaller sub set of the original larger data set may make it less representative of the full range of conditions of the underlying target population being sampled. Thus, sub-sampling (re-sampling) to a smaller and less representative sample would tend to impede the quality assurance basic of ensuring representativeness of the underlying target population being sampled.

However, explanations of the more complex ways to simultaneously handle both spatial and temporal autocorrelation tend to be tough reads that are probably beyond the capability of our beginner target population. Those who want to learn more could consult Legendre and Legendre (1998); Manly and Mackenzie (2003); Gilbert (1987); and Fortin and Dale, (2005). Consult your local applied statistician if necessary?

In ArcGis, one can construct maps which show both temporal and spatial status and trends, using time-aware GIS data tools summaries (see http://help.arcgis.com/en/webapi/flex/samples/index.html?#/TimeSlider_dynamic_service/01nq000003500000/). However, since needed calculations can be complex; some choose to portray complex spatial/temporal combination patterns in R software (Bill Monahan, NPS, Personal Communication, 2011).

Is There Multivariate EDA?

Yes, but we won't spend much space on it here due to its complexity. Multivariate EDA is often needed for ecological assemblages. Spatial autocorrelation using covariance functions and very complex exponential models are among EDA methods used. Simultaneous consideration of both spatial and temporal issues including EDA based on Bayesian or simulation methods are also used. Principal Components Analysis and other complicated EDA clustering and ordination methods are sometimes used as an alternative to autocorrelation analyses for spatial redundancy analyses or for clustering the data into logical and similar categories or strata (as part of monitoring optimization tasks).

Some software packages specialize in assisting with more complex multivariate EDA including the use of interactive stacked graphs. As one continues EDA in long term monitoring, one might try to partition variability according to one or more explanatory variables or factors, or measuring or testing interactions among factors. Many of the most helpful summaries on complex Level Three EDA are in books and journal articles tend to be not easily available for free on the Internet and/or not readily understandable by beginners. An exception is the summary by Kittel (2009), which is available for free on the Internet.

Methods for multivariate EDA plotting in R software summarized on the Internet by Coghlan (2011b) included the following:

- [A Matrix Scatterplot](#)
- [A Scatterplot with the Data Points Labelled by their Group](#)
- [A Profile Plot](#)
- [Calculating Summary Statistics for Multivariate Data](#)
- [Means and Variances Per Group](#)
- [Between-groups Variance and Within-groups Variance for a Variable](#)
- [Between-groups Covariance and Within-groups Covariance for Two Variables](#)
- [Calculating Correlations for Multivariate Data](#)

- [Standardising Variables](#)
- [Principal Component Analysis](#)
 - [Deciding How Many Principal Components to Retain](#)
 - [Loadings for the Principal Components](#)
 - [Scatterplots of the Principal Components](#)
- [Linear Discriminant Analysis](#)
 - [Loadings for the Discriminant Functions](#)
 - [Separation Achieved by the Discriminant Functions](#)
 - [A Stacked Histogram of the LDA Values](#)
 - [Scatterplots of the Discriminant Functions](#)
 - [Allocation Rules and Misclassification Rate](#)

How Does EDA Compare to FDA?

EDA is a toolset of techniques for exploring the data before it is modeled. **Functional data analyses (FDA)** is a toolset of (often multivariate) nonparametric techniques for analyzing and (in particular) modeling the data, usually based on some prior understanding of how the data should behave. In other words, after an EDA analysis, one can choose FDA (or some other method) to model it. However, both EDA and FDA are considered to exploratory and data mining tools by some. Both EDA and FDA fit into the “summarize and then analyze” good-advice paradigm, though EDA would typically be done earlier in the process.

How Do I Follow Up EDA with Trend Analyses When I Have Rotating Panels and/or a Probabilistic Sampling Design?

Very carefully! Most beginners will need to seek help. The focus of the trend discussions in this document relate to analyzing monitoring results from multiple years and fixed sites at reasonably regularly-spaced time intervals. More complex approaches may be needed when there are fewer temporal samples but more sites, more complex response variables, and (especially) rotating panel designs where sites are not always visited once a year and/or may change over time, and where such sites were picked originally with probabilistic sampling designs. Over very long term monitoring, one may need to re-do the sampling frame, which complicates things even further. Optimal methods to do complex EDA and complex follow-on trend analyses in some of these scenarios can be more complex than is easily explained in this summary for beginners (consult your local applied statistician).

Some Examples of Putting It All Together

There are probably many more complete and complex examples, but since this summary is for beginners, here are a few comparatively simple ones:

In analyzing water quality data, Maryland first looks at the data from different angles during early EDA plots. If autocorrelation is present they take steps to accommodate it before any trend test. Then they perform a seasonal Kendall trend test (accompanied by a Sen's slope estimator for the magnitude of the trend) to test for linear trends. Finally they also test for a quadratic trend. This is done partly because it is not always conclusive which type of trend test is ideal for borderline datasets (with some curve). Therefore, Maryland routinely uses both the SK test and a quadratic test and sometimes follows up with a more complex model-based trend analyses to help better identify and describe non-linear trends. At some sites the conclusions based on Seasonal Kendall test was no trend while the conclusion with the quadratic test was decreasing and then increasing. At other sites the conclusions based on Seasonal Kendall test was a significant decreasing trend while the conclusion with the quadratic test was decreasing and then increasing. At still other sites the conclusions based on Seasonal Kendall test was a significant decreasing trend while the conclusion with the quadratic test was no significant trend (Hall and Wazniak, 2006).

British Columbia has used a very similar approach. Regnier and Ryan (1999) did early EDA work on large water quality data sets in British Columbia by first looking at time series line plots and time series box and whisker plots (lined up side by side, see example at http://www.env.gov.bc.ca/wat/wq/trendstuff/bctrends/trendsbc-34.html#P600_30819). To look from different angles of interest, they often blocked by both month and year. This step is typical for what we call early Level One EDA. Monthly blocking is used to block out seasonal variation by only looking at data over the years based on single month. After looking at several both simple and complex lines of evidence during EDA, in final data presentation, for each parameter, they included three lines of evidence related to potential trends: 1) a simple time series, 2) the conclusion obtained from a nonparametric linear trend test, the seasonal Kendall test for serial dependence, and 3) the conclusion obtained from a quadratic non-linear trend test. An example for total phosphorus in the water column data can be seen at http://www.env.gov.bc.ca/wat/wq/trendstuff/bctrends/trendsbc-05.html#P380_20617. This example is of interest because the line plot was inconclusive, the SKSD test conclusion was decreasing trend, and the conclusion of a quadratic non-linear trend test was positive quadratic trend. The three lines of evidence (SK test, quadratic test and box and whisker plots) presented are a good example of looking at trends from different angles and why one should not depend only on one line of evidence. More detail:

Overall Regnier and Ryan (1999) presented 40 some plots with these three lines of evidence, and one can choose different examples from <http://www.env.gov.bc.ca/wat/wq/trendstuff/bctrends/index.html#TopOfPage>. In their summary discussion that can be seen at http://www.env.gov.bc.ca/wat/wq/trendstuff/bctrends/trendsbc-02.html#P346_17250, Regnier and Ryan (1999) used these multiple lines of evidence to conclude that non-parametric and regression tests have confirmed that total aluminum data were indeed displaying significant decreasing trends. Non-parametric tests indicated that total iron exhibited a linear decreasing trend, while regression analysis indicated a negative

quadratic (*i.e.* n-shaped trend). Total phosphorus was shown to have a marginally significant linear decreasing trend, and a positive quadratic trend. The boxplot and graph for iron indicate that iron values were elevated between 1988 and 1990, but have decreased since then, accounting for the quadratic trend. On the other hand, other plots showed that phosphorus decreased during the period from 1989 to 1993, but has increased since then.

More complex Autoregressive Moving Average Models (ARMA) and Autoregressive Integrated Moving Average Models (ARIMA) time series methods are often associated with predictive models but are also sometimes used as part of environmental trend analyses. These are mostly too complex for our beginner target audience. However, consider the following (**relatively simple**) example of first doing simple EDA plots, then autocorrelation plots, then deciding what models might fit best. The following is from NIST (2006, Section 6.4.4.6.1 at <http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4461.htm>) based on climatic data associated with southern oscillation patterns. There NIST gives an example of a continuing sequence of EDA plots one might perform when assessing long term monitoring data. First, during very early EDA (what we call Level One) NIST looked at the data in a time series line plot. The run sequence plot indicated stationarity. A follow-up seasonal subseries line plot indicated that there is no significant seasonality. Next (in what we call Level Two EDA) NIST generated both autocorrelation and partial autocorrelation correlogram plots of the raw data to begin to assess data independence by looking at autocorrelation and other patterns. In the NIST example large lag standard error ACF correlogram, one sees an indication that considerable autocorrelation is present, with a further indication of a mixture of exponentially decaying and damped sinusoidal components. In other words, though seasonal cycles were not strong, there were indications in the ACF correlogram of a longer term cycles. This provided a hint that an autoregressive model, with order greater than one, may be appropriate for these data. Finally, NIST drew a PACF correlogram based on the simple approximation equation to help identify the kind of model that best describes the data. The partial autocorrelation correlogram plot suggested that an AR(2) model might be appropriate. However, NIST further concluded that model validation should be performed before accepting this as a final model.

References Cited:

Since this document is designed to be used primarily as a digital document even in very remote locations (for example National Parks in Alaska far from libraries) the most comprehensive and free Internet summaries are prominently used as references and have been bolded below. In other cases, no alternative as good as a hard copy book (or not as easily available on the Internet) journal article was found, so no bolding was done or only a URL to find the citation was bolded.

Anderson, D., K. Burnham, W. Gould, and S. Cherry. 2001. Concerns about finding effects that are actually spurious. Wildlife Society Bulletin 29(1): 311-316. Entire Article Accessed March 2011 from:

<http://www.onepoint.ca/concerns%20about%20finding%20effects%20that%20are%20actually%20spurious.pdf>

Anderson, D. and K. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods *Journal of Wildlife Management* 66(3): 912–918. Entire paper accessed March 2011 from:

http://www.auburn.edu/~grandjb/wildpop/readings/Design%20and%20Model%20Selection/anderson_burnham_2002.pdf.

Anselin, L., S. Sridharan, S. Gholston. 2007 Using exploratory spatial data analysis to leverage social indicator databases: the discovery of interesting patterns. *Social Indicators Research*. 82(2):287–309. Accessed August 2011 from

http://www.chs.med.ed.ac.uk/ruhbc/evaluation/key_capability_store/social_indicators_FINAL.pdf

http://www.chs.med.ed.ac.uk/ruhbc/evaluation/key_capability_store/social_indicators_FINAL.pdf

APHA. 1998. Standard Methods Book, 20th Edition. **Information on the Standards Methods Books series** accessed December, 2011 from

<http://www.standardmethods.org/ViewArticle.cfm?articleID=52>

Box, G. E. and Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day. San Francisco. Note: this book is not on the Internet, but this book as well as earlier and later versions are quoted extensively in other references on the Internet, including but not limited to NIST (2006), Helsel and Hirsch, 2002) and the documentation for various statistics software packages. Reviews and summary of the book accessed December, 2011 from

http://www.amazon.com/Time-Analysis-Forecasting-Control-Revised/product-reviews/0816211043/ref=dp_top_cm_cr_acr_txt?ie=UTF8&showViewpoints=1

Chandler, R. and M. Scott, 2011. *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*. Wiley. ISBN: 978-0-470-01543-8, 388 pages, **book summary** accessed July 2011 from <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470015438.html>.

Chatfield, C. 1982. *The analysis of time series: an introduction*. 2nd Edition first published in 1980 and reprinted in 1982. Chapman & Hall, New York, NY. ISBN 0-412-22460-7. **Preview of 2003 update of this book** accessed December, 2011 from

http://books.google.com/books?id=qKzyAbdaDFAC&pg=PA24&lpg=PA24&dq=correlogram+independent+sample&source=bl&ots=svy-7bWAMj&sig=WUf70hN3zJ5NMA-Z2zOFmbUzbGw&hl=en&ei=vvRISrWNPMGFtgewpKmnDw&sa=X&oi=book_result&ct=result&resnum=6.

Cleveland, W. 1981. LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression. W. S. Cleveland. *The American Statistician*, 35:54, 1981, **Summary Notes on this paper** accessed November, 2011 from <http://stat.bell-labs.com/wsc/smoothsoft.html>

Cleveland, W, and E. Grosse. 1991. Computational Methods for Local Regression. Statistics and Computing, 1:47-62. **Summary notes on this paper** accessed November, 2011 from <http://stat.bell-labs.com/wsc/smoothsoft.html>

Coghlan, A., 2011a. Using R for Time Series Analysis. A Little Book of R for Time Series Release 0.1, Entire booklet accessed December, 2011 from https://github.com/avrilcoghlan/LittleBookofRTimeSeries/raw/master/_build/latex/TimeSeries.pdf.

Coghlan, A., 2011b. Using R for Multivariate Analysis. A Little Book of R For Time Series Release 0.1, Entire booklet accessed December, 2011 from https://github.com/avrilcoghlan/LittleBookofRTimeSeries/raw/master/_build/latex/TimeSeries.pdf

Cowpertwait, P. 2006. Introductory Time Series with R. Entire Summary Article Accessed December, 2011 from <http://www.unalmed.edu.co/~ndgiraldo/Archivos%20Lectura/Archivos%20curso%20Series%20EIO/notas%20series%20en%20r%20couperwait.pdf>.

Data Description Inc. 2011. Statistics and Data Analysis Software. Accessed November 2011 from http://www.datadesk.com/products/data_analysis/software.shtml.

Dawdy, D. and N. Matalas, 1964, Statistical and probability analysis of hydrologic data, Part III: Analysis of variance, covariance and time series, in Ven Te Chow, ed., Handbook of applied hydrology, a compendium of water-resources technology: New York, McGraw-Hill Book Company, p. 8.68-8.90.

Ellison, A. 1993. Exploratory data analysis and graphic display. Pages 14-45 in S. M. Scheiner and J. Gurevitch, editors. Design and Analysis of Ecological Experiments. Chapman & Hall, New York. [Full Text](http://harvardforest.fas.harvard.edu/personnel/web/aellison/publications/1993/ellison_1993.pdf) accessed June 2011 from http://harvardforest.fas.harvard.edu/personnel/web/aellison/publications/1993/ellison_1993.pdf

EPA. 2003 Mid-Atlantic Lessons Learned at <http://www.epa.gov/bioiweb1/pdf/EPA-903-R-03-003DevelopingBiologicalIndicators-LessonsLearnedfromMid-AtlanticStreams.pdf>

EPA. 2006. DQA Statistical Tools at <http://www.epa.gov/quality/qs-docs/g9s-final.pdf>

EPA. 2008. Scout 2008 Version 1.0 User Guide Part I. EPA/600/R-08/038, Accessed December 2011 from [http://www.epa.gov/esd/databases/scout/Scout%202008%20v.1.00.01/Scout%20Resources/Scout%202008%20Guides/Scout%202008%20v.%201.0%20User%20Guide%20\(Second%20Edition\)/Scout%202008%20v.%201.0%20User%20Guide%20\(2nd%20ed.\)%20in%20pdf%20format/Scout%202008%20Version%201.0%20Part%201.pdf](http://www.epa.gov/esd/databases/scout/Scout%202008%20v.1.00.01/Scout%20Resources/Scout%202008%20Guides/Scout%202008%20v.%201.0%20User%20Guide%20(Second%20Edition)/Scout%202008%20v.%201.0%20User%20Guide%20(2nd%20ed.)%20in%20pdf%20format/Scout%202008%20Version%201.0%20Part%201.pdf).

EPA. 2009. Statistical Groundwater Unified Guidance at
<http://www.epa.gov/wastes/hazard/correctiveaction/resources/guidance/sitechar/gwstats/unified-guid.pdf>

EPA. 2010. Technical Support Center for Monitoring and Site Characterization Statistical Software ProUCL 4.0 for Environmental Applications for Data Sets with and without Nondetect Observations. Accessed December, 2011 from
<http://www.epa.gov/nerlesd1/tsc/software.htm>.

Fortin, M-J and M. Dale 2005. Spatial analysis: a guide for ecologists. Cambridge University Press, **Preview of selected pages accessed December, 2011 from Google Books at**
http://books.google.com/books?id=bTT7tEiTFLoC&source=gbs_navlinks_s

Garson, 2009. Time Series Analysis. In: Quantitative Research in Public Administration. North Carolina State University, cited with permission of the author, Accessed December, 2011 from <http://faculty.chass.ncsu.edu/garson/PA765/time.htm>

Gilbert, R. 1987. Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold Co., **Preview pages accessed December, 2011 from**
http://books.google.com/books?id=IEo1rvDGUEkC&pg=PA257&dq=Gilbert,+R.,+1987.+Statistical+methods+for+environmental+pollution+monitoring.&hl=en&ei=4yPZTrbxJsmtgQfqlfD1Dg&sa=X&oi=book_result&ct=result&resnum=1&ved=0CDgQ6AEwAA#v=onepage&q&f=false

Gordon, N., T. McMahon, B. Finlayson, C. J. Gippel, and R. J. Nathan. 2004. Stream hydrology, an introduction for ecologists. 2nd edition. John Wiley & Sons, Ltd. West Sussex, England. **Book description accessed December 2011 from**
<http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470843586.html>.

Griffith, L. et al. 2001. Data Analysis Considerations in Producing ‘Comparable’ Information for Water Quality Management Purposes, Accessed December, 2011 from
<http://acwi.gov/monitoring/pubs/tr/nwqmc0101.pdf>

Hahn, G. and W. Meeker, 1991. Statistical Intervals: A Guide for Practitioners, New York: John Wiley & Sons, Inc.

Hall, M. and C. Wazniak. 2006. Status and trend analysis in the Maryland water quality Monitoring program. Maryland Department of Natural Resources. Accessed November, 2011 from http://www.inlandbays.org/wp-content/uploads/2010/12/hall_mcbtrendanaly.pdf

Harcum, J., J. Loftis and R. Ward. 1992. Selecting trend tests for water quality series with serial correlation and missing values. Water Resources Bulletin 28(3): 469-478.

Hargreaves, B. and. T. McWilliams, 2009 Polynomial Trendline function flaws in Microsoft Excel. Computational Statistics and Data Analysis 54: 1190-1196. Accessed December, 2011 from

http://xa.yimg.com/kq/groups/12634412/1319898326/name/ComputStatDataAnal2010v54-1190_Polynomial+Trendline+function+flaws+in+Microsoft+Excel.pdf.

Hartwig, F. and B. Dearing. 1979. Exploratory Data Analysis, Sage Publications, Beverly Hills, CA, Book preview of selected pages accessed November, 2011 from <http://books.google.com/>.

Helsel, D. 2009. An Introduction to Effects of Serial Correlation. Practical Stats January 2009 Newsletter. Accessed December 2011 from http://practicalstats.com/news/news/aesnews_files/09Jan_SerialCorr.pdf

Helsel, D. 2011. The Effect of Serial Correlation on the Mann-Kendall Test for Trend. Practical Stats December 2011 Newsletter. In 2012, the newsletter will be available at <http://practicalstats.com/news/news/>

Helsel, D. and A. Hirsch 2002. Statistical Methods in Water Resources at <http://pubs.usgs.gov/twri/twri4a3/pdf/twri4a3.pdf>.

Hirsch, R., and J. Slack. 1984. A nonparametric trend test for seasonal data with serial dependence: Water Resources Research 20: 727-732. Abstract accessed November 2011 from <http://www.agu.org/pubs/crossref/1984/WR020i006p00727.shtml>.

Irwin, R.J. 2008. Draft Part B lite QA/QC Review Checklist for Aquatic Vital Sign Monitoring Protocols and SOPs, National Park Service, Water Resources Division. Fort Collins, Colorado, Accessed January 2010 from http://www.nature.nps.gov/water/Vital_Signs_Guidance/Guidance_Documents/PartBLite.pdf

Irwin, R. 2010. Serial Correlation and Trend Option Basics. Platform Presentation at the Denver, Colorado National Water Quality Monitoring Council National Meeting. Power Point Slide Summary Accessed December, 2011 from http://acwi.gov/monitoring/conference/2010/C3/C3_Irwin.pdf

ISO. 2008. Guide to the Expression of Uncertainty in Measurement (GUM), International Organization for Standardization, Geneva, Switzerland). Accessed July 2011 from http://www.bipm.org/utis/common/documents/jcgm/JCGM_100_2008_E.pdf, also available off root directory at <http://www.bipm.org/en/publications/guides/gum.html>
Jassby, A. and J. Cloern. 2010. wq: Exploring water quality monitoring data in R. Accessed December, 2011 from <http://cran.r-project.org/web/packages/wq/vignettes/wq-package.pdf>

Jowett, I. 2009. Time Trends Software v 2.1. Analysis of water quality trends. Free Download accessed June 2009 from NIWA at <http://www.niwa.co.nz/our-science/freshwater/tools#analysis>. Plain language announcement of Time Trends software availability accessed June 2009 from <http://www.niwa.co.nz/news-and-publications/publications/all/wru/2008-27/water>.

Kirchner, D. 2001. Data Analysis Toolkit #11. Accessed February 2011 from http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_11.pdf.

Kirchner, J, X. Feng, C. Neal, and A. Robson. 2004. The fine structure of water-quality dynamics: the (high-frequency) wave of the future. *Hydrological Processes* 18:1353-1359. Accessed December, 2011 from http://seismo.berkeley.edu/~kirchner/reprints/2004_67_high_frequency_sampling.pdf.

Koch, R.W. and G.M. Smillie. 1986. Bias in hydrological prediction using log-transformed regression models. *Water Resources Bulletin* 22 (5):717-723).

Kittel, T. 2009. The Development and Analysis of Climate Datasets for National Park Science and Management: A Guide to Methods for Making Climate Records Useful and Tools to Explore Critical Questions. Final Report. Prepared for National Park Service Inventory and Monitoring Program. Entire document accessed November, 2011 from NPS website at https://irma.nps.gov/App/Reference/DownloadDigitalFile?code=426579&file=KittelT_2009_Climate_Data_Analysis_Report.pdf (if it doesn't open quickly, try copying and pasting that URL into your browser).

Legendre, P. and L. Legendre. 1998. Numerical ecology. 2nd English Edition, Elsevier, ISBN 0444892508. Limited preview pages accessed December, 2011 from http://books.google.com/books?id=KBoHuoNRO5MC&dq=legendre+numerical+ecology&source=gbs_summary_s&cad=0

Lenth, R. 2011. Java Applets for Power and Sample Size [Computer software]. Accessed May, 2011 from <http://www.stat.uiowa.edu/~rlenth/Power/>

Maindonald, J. 2008. Using R for Data Analysis and Graphics Introduction, Code and Commentary. Accessed December 2011 from <http://cran.r-project.org/doc/contrib/usingR.pdf>.

Manly, B. and D. Mackenzie. 2003. CUSUM environmental monitoring in time and space. *Environmental and Ecological Statistics* 10 (2):231-247. Article description accessed December, 2011 from <https://commerce.metapress.com/content/q86728063040g67k/resource-secured/?target=fulltext.pdf&sid=c3oqsm3e13mnh0vsontul345&sh=www.springerlink.com>

Matalas, N. 1963, Autocorrelation of rainfall and streamflow minimums: U.S. Geological Survey, Prof. Paper 434-B.

Matalas, N. 1967. Time series analysis. *Wat. Resour. Res.* 3(3): 817-829.

Matalas, N. 1998. Note on the Assumption of Hydrologic Stationarity, Universities Council of Water Resources. Water Resources Update Universities Council on Water Resources Issue No. 112, Accessed April 2010 from http://www.ucowr.siu.edu/updates/pdf/V112_A11.pdf.

Matalas, N., and A. Sankarasubramanian. 2003. Effect of persistence on trend detection via regression. Water Resources Research 39 (12), Accessed December, 2011 from North Carolina State University Website at <http://www.ce.ncsu.edu/research/hydroclimatology/people/sankar/pubs/2003WR002292.pdf>.

Mathworks. 2010. LOWESS and LOESS: Local Regression Smoothing. Section in curve fitting tool box. Documentation for MathWorks Products (Release 13 with Service Pack 2 for Matlab 6.5 statistical software). Accessed March 2010 from <http://www.mathworks.com/help/toolbox/curvefit/smooth.html>.

Meko, D. 2009. Correlogram. Online class notes for University of Arizona Applied Time Series Analysis class GEOS 585a, cited by permission of the teacher (D. Meko). Accessed July 2009 from www.ltr.arizona.edu/~dmeko/notes_3.pdf

McBride, G. 2005. [Using Statistical Methods for Water Quality Management: Issues](http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471470163.html), Options and Solutions. Wiley, NY, 313 pp. Book description accessed November, 2011 from <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471470163.html>

McDonald, T., Manly, B., Nielson, R. 2009. Review of Environmental Monitoring Methods: Trend Detection, Accessed December, 2011 from <http://www.west-inc.com/reports/Review%20of%20Trend%20Detection2009.pdf>.

Milly, P. et al. 2008. Climate change. Stationarity is dead: whither water management? Science. 319(5863):573-4. Entire Paper accessed December 2011 from http://www.azwaterinstitute.org/media/CCADAPT/milly_et_al.pdf.

Morrison, J. 2011. Autoregressive Integrated Moving Average Models (ARIMA). Econometric Plain Language Summary accessed December 2011 from <http://www.forecastingsolutions.com/arima.html>.

Morrison, L, 2008. The Use of Control Charts to Interpret Environmental Monitoring Data. Natural Areas Journal 28(1):66-73, entire article accessed December 2011 from http://science.nature.nps.gov/im/units/htln/library/Monitoring/Monitoring_CtrlCharts_2008_a.pdf

NIST (National Institute of Standards and Technology). 2002. DATAPLOT Software. Accessed December 2011 from <http://www.itl.nist.gov/div898/software/dataplot/summary.htm>. DataPlot tutorial and FAQs accessed December 2011 from <http://www.itl.nist.gov/div898/software/dataplot/homepage.htm>.

NIST. 2006. Handbook of Statistical Methods, Chapter 1. Exploratory Data Analysis at <http://www.itl.nist.gov/div898/handbook/eda/eda.htm>

Palmer, M. 2011. Hypothesis-Driven and Exploratory Data Analysis. Accessed September, 2011 from <http://ordination.okstate.edu/motivate.htm>. Addendum to Oklahoma State University Home Page summary on Ordination Methods for Ecologists, accessed September, 2011 from <http://ordination.okstate.edu/>

Phillips, J. 1999. How to think about statistics, Sixth Edition. W.H. Freeman, NY, 202 pages.

Polhemus, N. 2005. How To: Apply SPC Techniques to Correlated Variables Using STATGRAPHICS Centurion, Accessed November, 2011 from <http://www.statlets.com/How%20To%20Apply%20SPC%20Techniques%20to%20Correlated%20Variables.pdf>

Pyper, B. and R. Peterman. 1998. Comparison of methods to account for autocorrelation in correlation analyses of fish data. Canadian Journal of Fisheries and Aquatic Sciences, 55:(9) 2127-2140, 10.1139/f98-104/, Abstract accessed December, 2011 from <http://www.nrcresearchpress.com/doi/abs/10.1139/f98-104>.

Reckhow, K., K. Kepford, and W. Warren-Hicks. 1993. Methods for the Analysis of Lake Water Quality Trends. EPA 841-R-93-003. Washington, DC. Available from EPA National Environmental Publications Internet Site (NEPIS). Accessed December 2011 from <http://www.epa.gov/nscep/> (to get to the document from there, search for 841R93003).

Regnier, R. and A. Ryan. 1999. Water Quality Trends at Selected Monitoring Sites in British Columbia, Chapter on Statistical Methods. Prepared for Environment Canada and BC Ministry of Environment, Lands and Parks. Accessed November 2011 from <http://www.env.gov.bc.ca/wat/wq/trendstuff/bctrends/trendsbcb-01.html>). Also accessed off homepage at <http://www.env.gov.bc.ca/wat/wq/trendstuff/bctrends/index.html#TopOfPage>

Sanger, D., M. Arendt, Y. Chen, E. Wenner, A. Holland, D. Edwards and J. Caffrey. 2002. A synthesis of Water Quality Data: National Estuarine Research Reserve System-wide Monitoring Program (1995-2000). National Estuarine Research Reserve Technical Report Series 2002: 3. South Carolina Department of Natural Resources, Marine Resources Division Contribution No. 500, 135 p., full report accessed December, 2011 from <http://nerrrs.noaa.gov/Doc/PDF/Research/SWMPSynthesis1995.pdf>.

SAS. 2011. The AUTOREG Procedure, accessed December, 2011 from http://support.sas.com/rnd/app/ets/proc/ets_autoreg.html

Shah, S., S. Munawar, S. Khan, M. Ahmed, I. Hussain and K. Eskridge. 2009 Nonparametric methods in combined heteroscedastic experiments for assessing stability of wheat genotypes in Pakistan. Pak. J. Bot., 41(2): 711-730. Entire Document Accessed December, 2011 from [http://www.pakbs.org/pjbot/PDFs/41\(2\)/PJB41\(2\)711.pdf](http://www.pakbs.org/pjbot/PDFs/41(2)/PJB41(2)711.pdf)

Singh, A., N. Armbya, and A.K. Singh. 2010, ProUCL Version 4.00.05 Technical Guide (Draft). EPA/600/R-07/041. Entire Document Accessed December, 2011 from [http://www.epa.gov/nerlesd1/tsc/ProUCL_v4.00.05/ProUCL_v4.00.05_tech_guide\(draft\).pdf](http://www.epa.gov/nerlesd1/tsc/ProUCL_v4.00.05/ProUCL_v4.00.05_tech_guide(draft).pdf) (future updates will be posed at http://www.epa.gov/nerlesd1/tsc/TSC_form.htm)

STATGRAPHICS Software. 2005. Time Series – Descriptive Methods. STATGRAPHICS Centurion XV Software PDF summary of methods used, StatPoint Technologies, Inc.

Statgraphics. 2009a. Exploratory Data Analysis in Statgraphics Centurion Software. Accessed October, 2011 from <http://www.statgraphics.com/eda.htm>. Note: this Internet summary gives some basic descriptions of some of Tukey (1977) suggested tools for EDA.

Statgraphics. 2009b. Time Series Analysis and Forecasting at http://www.statgraphics.com/time_series_analysis.htm

StatSoft 2008. Electronic Textbook for Statistica at <http://www.statsoft.com/textbook/>

Taylor, B. and C. Kuyatt. 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results NIST Publication TN 1297 (accessed December, 2011 from <http://www.nist.gov/pml/pubs/tn1297/index.cfm>)

Thoma, D., R Irwin and P. Penoyer. 2011 Documenting measurement sensitivity and bias of field-measured parameters in water quality monitoring programs. Environ Monit. Assess., DOI 10.1007/s10661-011-2347-5. Published Online, Full Report Accessed November, 2011 from <http://www.springerlink.com/content/h7255516w2552178/fulltext.pdf>

Tukey, J. 1977. Exploratory Data Analysis. Addison-Wesley. ISBN 0-201-07616-0. Note: this is the original summary work on EDA but not available on the internet.

van Mantgem, P, and D. Schwilk. 2009. Negligible influence of spatial autocorrelation in the assessment of fire effects in a mixed conifer forest. Fire Ecology 5 (2): 116-125, Accessed December, 2011 from <http://fireecology.net/docs/Journal/pdf/Volume05/Issue02/116.pdf>.

Wainer. H. 2007. The most dangerous equation. American Scientist 95: 249-256. Accessed December 2011 from <http://nsm.uh.edu/~dgraur/niv/TheMostDangerousEquation.pdf>

Wessa, P. 2008. ACF and PACF Correlograms in R. Free Statistics Software (Calculator) accessed December, 2011 from http://www.wessa.net/rwasp_autocorrelation.wasp.

Yue, S., and C. Wang, 2002. Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test, Water Resour. Res., 38(6), 1068, doi:10.1029/2001WR000861.

Zar J. 1984. Biostatistical analysis 2nd Ed. Prentice-Hall Inc, Engelwood Cliffs, NJ.

Zuur, A. E. N. Ieno, and G. Smith 2007. Analysing Ecological Data. Part of the Springer Series for Statistics for Biology and Health. 680 p. Publisher (Springer) **book description accessed March 2011 from www.springer.com/0-387-45967-7**.