Statistical Methods for the Analysis of Lake Water Quality Trends

Technical Supplement to

The Lake and Reservoir Restoration Guidance Manual

1993

This document should be cited as:

Reckhow, K.H., K. Kepford, and W. Warren Hicks. 1993. Methods for the Analysis of Lake Water Quality Trends. EPA 841-R-93-003.

This technical supplement was prepared by the Terrene Institute and Duke University, School of the Environment under EPA Cooperative Agreement No. CX-814969 from the Assessment and Watershed Protection Division and No. CX-820957 from Region V. Points of view expressed in this technical supplement do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency nor any of the contributors to its publication. Mention of trade names and commercial products does not constitute endorsement of their use.

PREFACE

This manual and the accompanying software in the SAS system present nonparametric statistical methods for trend assessment in water quality, with an emphasis on lakes. The purpose of the manual and software is to furnish lake program managers with guidance on the application and interpretation of methods for the detection of trends in lake water quality.

To provide a foundation, the manual begins with identification of basic concepts and approaches in applied statistics that are important in trend detection. This is followed by a discussion of hypotheses testing and common assumptions for parametric tests and nonparametric tests for water quality trend detection in lakes. The procedures and tests are presented in detailed examples for both single-lake and regional analyses.

The guidance manual concludes with a list of pertinent references, software, and appendices which provide a description of the trend detection software and additional background on descriptive statistics.

This manual is intended to be a living document that will be updated and improved as technology and circumstances change. As such, we request that you send all suggestions to the Clean Lakes Program, U.S. Environmental Protection Agency, 401 M Street, S.W., Washington, D.C. 20460.

CONTENTS

Section	Page
ABSTRACT	iii
ACKNOWLEDGMENTS	viii
1. Introduction	1
2. Basic Statistics and Statistical Concepts	7
2.1. Descriptive Statistics	7
2.2. Robustness, Resistance, and Influence	8
2.3. Hypothesis Testing	8
2.3.1. Introduction	8
2.3.2. Common Assumptions for Statistical Hypothesis Tests	9
2.4. Statistical Methods for Trend Detection	12
2.4.1. Summarizing Trend Data	12
2.4.2. Graphical Methods	13
2.4.3. Parametric Methods and Tests	14
2.4.4. Distribution-Free Methods and Tests	14
3. Individual Lake Analysis	17
3.1. Introduction	17
3.2. Examples - Background	18
3.3. Summary Statistics	22
3.4. Graphical Analyses	22
3.4.1. Histogram	25
3.4.2. Time Series Graph	25
3.4.3. Box Plot	25
3.5. Normality	29
3.6. Seasonality	29
3.7. Independence	31
3.8. Trend Detection in Total Phosphorus	38
3.9. Total Nitrogen	40
3.9.1. Normality	41
3.9.2. Seasonality	41
3.9.3. Independence	41
3.9.4. Trends in Total Nitrogen	41
3.10. Conclusions from the Phosphorus and Nitrogen Examples	50
3.11. Regional and Statewide Lake Analysis	50
3.11.1. Introduction	50
3.11.2. Tests of Significance	51
References	53

Section Page

Appendix A: Basic Descriptive Statistics A-1
Measures of Central Tendency A-1
Measures of Dispersion A-3
Graphical Analyses A-5

Appendix B: Introduction to SAS Macros B-1
Data Preparation B-2
Naming of Macro Variables B-2
Saving Files B-3
Graphics B-3

Appendix C: SAS Tables C-1

Appendix D: List of SAS Programs and Files on Disk D-1
Basics.sas D-2
Boxplt.sas D-5
Corr.sas D-16
Kens.sas D-20
Adjust.sas D-23
Corradj.sas D-27

LIST OF TABLES

Table	Page
2.1. Possible Outcomes from Hypothesis Testing	9
3.1. Falls Lake Data Set	20
3.2. Univariate TP Statistics	23
3.3. Falls Lake Correlogram Results (TP)	33
3.4. Falls Lake Kendall's Tau (TP)	36
3.5. Falls Lake Correlogram Results for Adjusted TP	39
3.6. Univariate TN Statistics	42
3.7. Falls Lake Kendall's Tau (TN)	49
3.8. Lake Information and Meta-Analytic Formula	51

LIST OF FIGURES

Figure — Page

1.1. Patterns of Variability 3
1.2. Hypothetical Example with Trend and Background Variability 4
1.3. Hypothetical Example with Trend, Seasonality, and Background Variability . 6

3.1. Flow Chart for Macro Programs 19
3.2. Univariate Graphs for Falls Lake TP Data 24
3.3. Histogram for Falls Lake TP Data 26
3.4. Time Series Plot for Falls Lake TP Data 27
3.5. Seasonal Box Plots for Falls Lake TP Data 28
3.6. Yearly Box Plots for Falls Lake TP Data 30
3.7. Correlogram for Falls Lake TP data 32
3.8. Data Analyses Used to Identify Appropriate Test for Trend Detection 34
3.9. Correlogram for Falls Lake Adjusted TP Data 38
3.10. Univariate Graphs for Falls Lake TN Data 43
3.11. Histogram for Falls Lake TN Data 44
3.12. Seasonal Box Plots for Falls Lake TN Data 45
3.13. Correlogram for Falls Lake TN data 46
3.14. Time Series Plot for Falls Lake TN Data 47
3.15. Correlogram for Falls Lake Adjusted TN Data 48

ACKNOWLEDGMENTS

The technical guidance presented in this manual has improved as a consequence of the involvement of several people. USEPA provided overall direction and valuable comments throughout the preparation of the document. J.R. Slack of the US Department of the Interior wrote the FORTRAN programs that are used within the SAS procedures presented here. Finally, many students in the Duke School of the Environment worked with early versions of the software and technical guidance and provided many helpful comments that improved both the software and the guidance manual.

Chapter 1

Introduction

Water quality varies in time and space as a function of many macroscopic and microscopic processes. On a large scale, changes in land use and pollutant discharge over time can cause permanent changes or trends in water quality in receiving waterbodies. On a yearly basis, seasonal changes in solar radiation, temperature, and precipitation can cause cyclical patterns in water quality that repeat each year. On a microscopic scale, many minor factors can influence water quality. For example, a variety of factors (e.g., wind, temperature, and shoreline irregularities) may collectively cause turbulent or molecular diffusion in water bodies that results in an apparent random behavior in water quality in time and space.

Detection of a trend in water quality over time is dependent on: (1) the acquisition of water quality data from a properly-designed monitoring program, (2) the application of appropriate statistical methods of trend detection, and (3) a good understanding of relevant water quality relationships. Both parametric and nonparametric (distribution-free) statistical methods have been proposed and applied for water quality trend detection purposes. With either type of procedure, the modeler seeks to separate a signal (the trend) from the noise (the "unexplained" component) in the water quality data.

The assessment of possible trends in lake water quality can be an important scientific task in support of lake water quality management. The presence or absence of trends over time in key water quality variables is a good indication of the degree to which water quality is responding to changes (land use and pollutant discharge) in the watershed. This information, in turn, provides a basis for predictive models of the pollutant loading - lake response relationship; these models can then be used to forecast future lake response to future watershed changes.

Formal statistical trend analysis also provides a rational, scientific basis for addressing concerns that may arise due to natural variations in water quality. For example, citizens who participate in water quality recreation may be distressed about undesirable "changes" in lake water quality that may be due entirely to natural variations. An ongoing water quality trend detection program could provide estimates of the likelihood that the observed "changes" reflect natural variability or real trends over time. This helps in citizen education, and in turn, may suggest

alternative management actions that may be directed at either reversing trends in water quality or reducing in-lake variability.

To help motivate the need for the application of the statistical methods presented in this manual, a hypothetical water quality data set is created and analyzed. To do this, first consider what factors cause **measured** water quality to vary or change over time. A reasonably comprehensive list of these factors is:

- ·trends
- ·seasonal cycles
- ·daily cycles
- ·variations in hydrology (e.g., streamflow, lake level)
- ·natural (unexplained) variability
- ·measurement error

In brief, "trends" refers to permanent changes in the level (e.g., mean value) of a water quality variable, "seasonal cycles" and "daily cycles" refer to oscillating patterns caused primarily by periodic changes in solar radiation, "variations in hydrology" refers to (for example) the often observed inverse relationship between volume of streamflow and concentration of a water quality variable, "natural variability" includes all factors (e.g., microscopic processes) not explicitly identified, and "measurement error" refers to the fact that there is always some error in the field and laboratory methods of analysis.

Using these definitions, a ten-year data series for monthly measurements of total phosphorus concentration in a lake is created. At the onset of sampling, the mean concentration is 20ug/l. A 20% linearly increasing trend is imposed over the ten year period, so that the mean concentration after ten years is 24ug/l. In addition, a seasonal cycle (with annual frequency) of amplitude 10ug/l is included as a sine wave. Finally, natural variability and measurement error are incorporated as a "noise" term at three different levels characterized by standard deviation of 1, 3, and 5 ug/l, respectively.

The trend and seasonal cycle are shown in Figure 1.1. Notice that the trend is visible on the graph even when combined with the sine wave. Thus, when the graphical evidence is as clear as presented in Figure 1.1, there may be little need for rigorous statistical analysis to confirm the existence of a trend (although the statistical analysis may still be useful to provide the best estimate of the magnitude of the trend).

Figure 1.2 presents a series of three graphs that combine the linear trend with the noise term (natural variability and measurement error) at successively higher levels of noise (characterized by the noise standard deviation). When the noise standard deviation is only 1.0 (top graph), the trend is still visible. However, as the standard deviation of the noise increases, the linear trend becomes visually

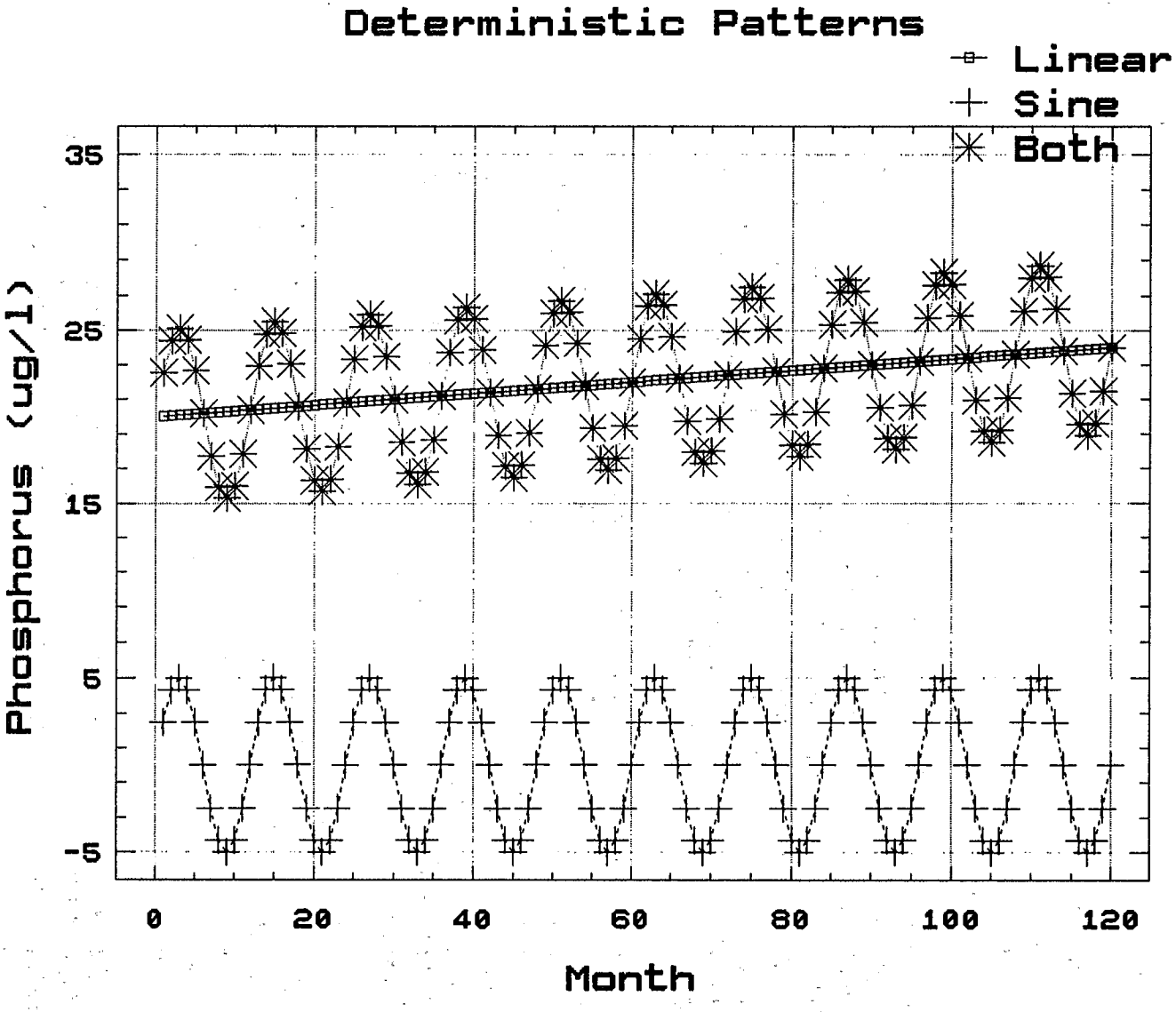


Figure 1.1. Patterns of variability.

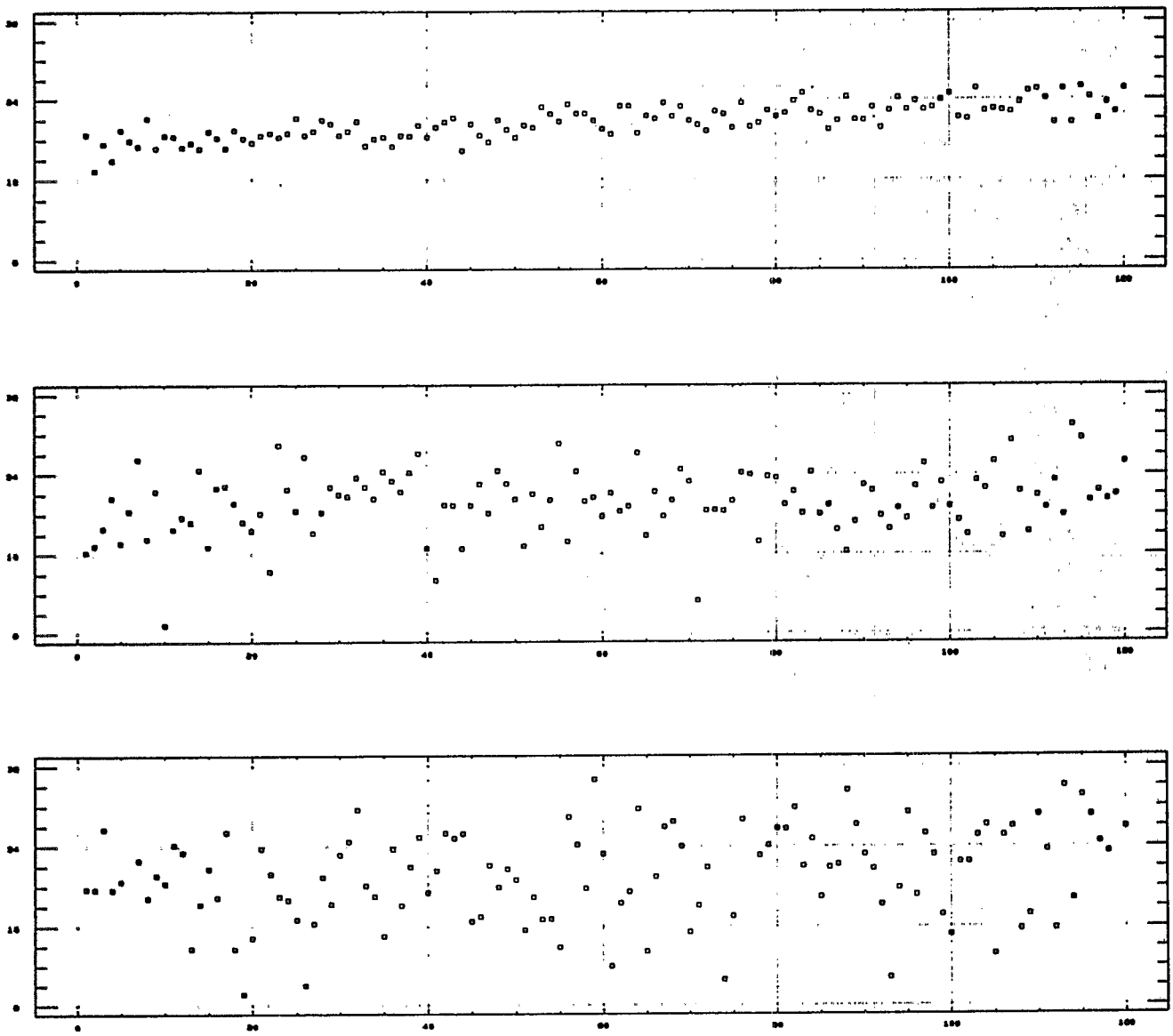


Figure 1.2. Hypothetical example with trend and background variability, shown from top graph to bottom with increasing levels of noise.

obscured by the natural variability and measurement error. In these cases, good statistical methods are needed to separate the signal (trend) from the noise.

In Figure 1.3, the seasonal cycle is added to the noise and trend. When the noise term is small (top graph), each separate component (particularly the sine wave) is visible. However, as the noise increases (bottom two graphs), the separate components become less evident visually. A combination of good limnological judgment (to assess the seasonal cycle) and statistical methods is necessary to successfully interpret a water quality time series like that in the bottom graph of Figure 1.3. The methods presented below are appropriate for this task.

The purpose of this manual is to furnish lake program managers with guidance on the application and interpretation of methods of trend detection in lake water quality. To provide a foundation, the manual begins with identification of basic concepts and approaches in applied statistics that are important in trend detection. This is followed by a discussion of hypothesis testing and common assumptions for parametric tests and nonparametric tests for water quality trend detection in lakes. These procedures and tests are presented for both single-lake and regional analyses. The guidance manual concludes with a list of pertinent references, software, and an appendix which provides a description of the trend detection software and additional background on descriptive statistics.

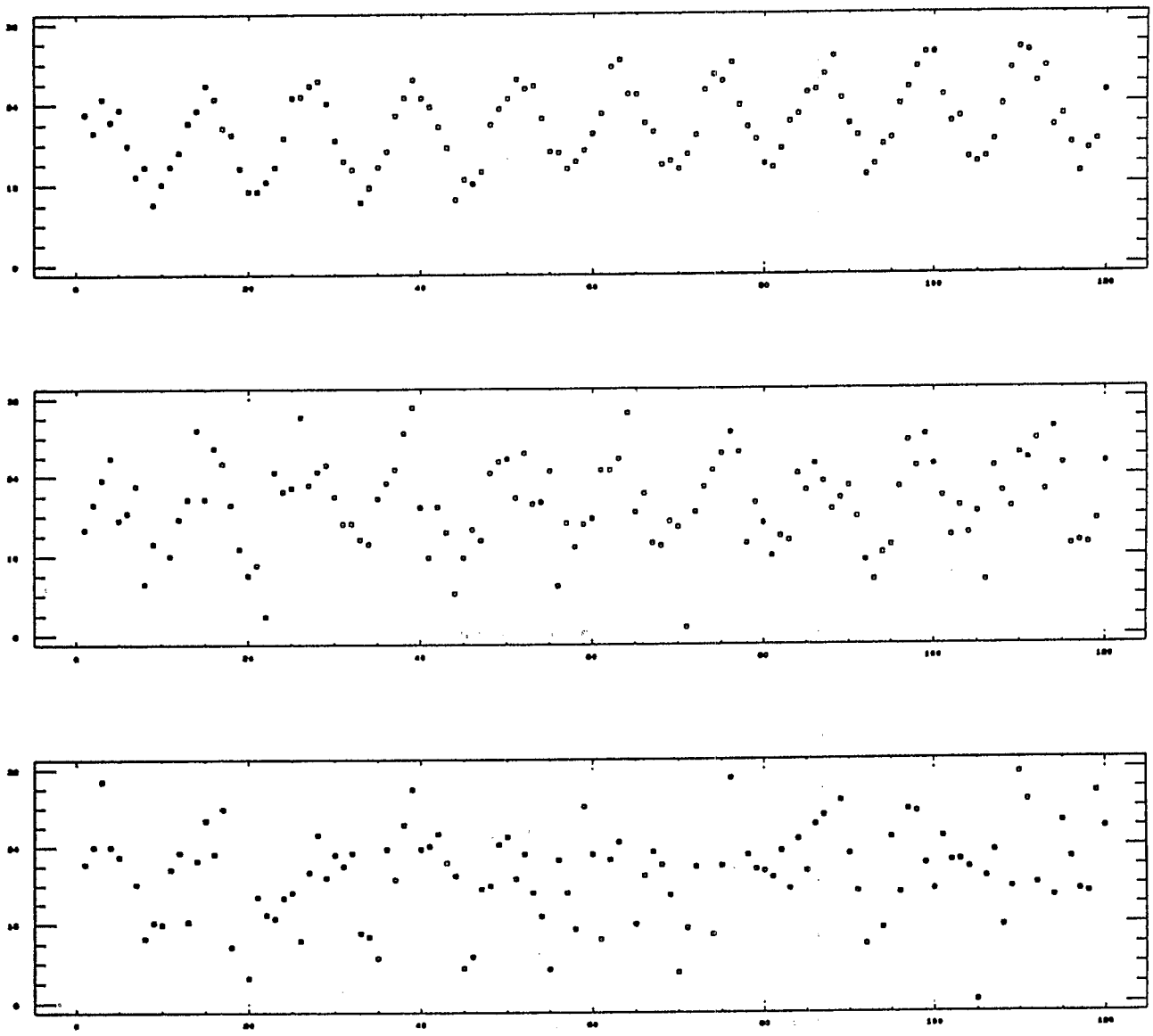


Figure 1.3. Hypothetical example with trend, seasonality, and background variability, shown from top graph to bottom with increasing levels of noise.

Chapter 2

Basic Statistics and Statistical Concepts

2.1. Descriptive Statistics

When a set of data is quite small, one may choose to use all of the data in an analysis or to present the entire data set in a report. For large data sets, the scientist recognizes that to most effectively transfer information he must summarize the data set with a few well-chosen statistics. A choice is made to trade some of the information contained in the entire data set for the convenience of a few descriptive statistics. This choice is usually a good one, provided the descriptive statistics that are selected correctly represent the original data.

Some descriptive statistics are so commonly used that we forget that they actually represent only one option among many candidate statistics. For example, the mean and the standard deviation (or variance) are statistics used to estimate the center of a data set and the spread on those data. When these statistics are to be used, the scientist should decide beforehand that they are the best choices to describe the aforementioned characteristics of the data set. Often they are (notably for symmetrically-distributed data following an approximate normal distribution), so their use is frequently justified. However, it is noted below and in Appendix A that there are many situations with lake water quality data where alternative descriptive statistics are preferred. These alternatives are robust/resistant statistics.

In the selection of descriptive statistics, it is important that the scientist have a clear understanding of the purpose that the statistic serves. In many limnological studies descriptive statistics are selected because the convenience of a few summary numbers outweighs the loss on information that results when the entire data set is described by the statistics. It is therefore essential that as much information as possible be summarized in the descriptive statistics because the alternative may be a misrepresentation of the original data.

Certain specific features of the data set are characterized using descriptive statistics. For example, the center, or central tendency of a set of data, is probably the most important measure. Among the candidate statistics for central tendency are the mean, median, mode, and geometric mean. Once the center of a data set is described, the next important feature for the data distribution is the spread,

dispersion, or scale. Among the candidate estimators of this feature of a data set are the range, standard deviation, and interquartile range. These two characteristics of a data set, central tendency and dispersion, are the most common descriptive statistics. Other characteristics, such as skewness and kurtosis, are occasionally important as well. At this point, if specific information and examples on descriptive statistics is desired, the reader should turn to Appendix A.

2.2. Robustness, Resistance, and Influence

In the statistics literature, robustness refers to insensitivity to assumption violations, resistance refers to insensitivity to outliers, and influence concerns the effect of observations on summary measures (statistics) of the data. In parametric statistical analysis, we make an assumption concerning an underlying population model (often normal). We hope that estimators (e.g., sample mean and variance) selected to summarize the data are robust if the probability model is incorrect, are resistant to influential data points or outliers, yet are efficient (low standard error) under any situation. If outliers and lack of resistance are concerns, we may choose a distribution-free method or nonparametric test for analysis. In the future, robust statistical methods may be the best choice for analysis of water quality data. At present, we tend to recommend nonparametric methods and tests unless there is little doubt that a parametric model is correct.

2.3. Hypothesis Testing

2.3.1. Introduction

In conventional statistical analysis, hypothesis testing for a trend is usually based on a point null hypothesis. Typically, the point null hypothesis is that there is no trend; it is often stated in this way as a "straw man" (Wonnacott and Wonnacott 1977) that the scientist expects to reject on the basis of the data evidence. To test this hypothesis, data are obtained to provide a sample estimate of the effect (e.g., change in surface pH in Adirondack lakes), the data are used to provide a sample estimate of a test statistic, and a table for the test statistic is consulted to estimate how unusual the observed value of the test statistic is if the null hypothesis is true. If the observed value of the test statistic is unusual, the null hypothesis is rejected.

In a typical application of parametric hypothesis testing, an hypothesis, H_0, called the null hypothesis, is proposed and then evaluated using a standard statistical procedure like the t-test. Competing with this null hypothesis for acceptance is the alternative hypothesis, H_1. Under this simple scheme, there are four possible outcomes of the testing procedure associated with the truth (true or false) and the test results (accept or reject) for each hypothesis; see Table 2.1.

Table 2.1 Possible Outcomes From Hypothesis Testing

State of the World	Decision	
	Accept H_0	Reject H_0
H_0 is True	Correct decision. Probability = 1 - α; corresponds to the *confidence level.*	Type I error. Probability = α; also called the *significance level.*
H_0 is False (H_1 is True)	Type II error. Probability = β.	Correct decision. Probability = 1 - β; also called *power.*

The point null hypothesis is a precise hypothesis that may be symbolically expressed as:

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

where θ is a parameter of interest. An example of a point null hypothesis is, in words, "there is no change in mean surface water total phosphorus concentration after imposition of a phosphate detergent ban." Symbolically, this may be expressed as:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

where μ_1 is the pre-ban true mean and μ_2 is the post-ban true mean. The test of this null hypothesis proceeds with the calculation of the sample means, x_1 and x_2. In most cases, the sample means will differ as a consequence of natural variability and/or measurement error, so a decision must be made concerning how large this difference must be before it is considered too large to be due to variability and/or error alone. In classical statistics, this decision is often based on standard practice (e.g., accept a type I error of 0.05), or on informal consideration of the consequences on an incorrect conclusion.

2.3.2. Common Assumptions for Statistical Hypothesis Tests

Virtually all statistical procedures and tests require that one or more assumptions hold. These assumptions concern either the underlying population being sampled or the distribution for a test statistic. Since lack of compliance with an assumption can

have a substantial effect on a statistical test, the common assumptions of normality, equality of variances, and independence are discussed in this section. Among the topics presented are the extent to which an assumption may be violated without serious consequences, and approaches that might be recommended to address possible assumption violations.

Normality

One of the common assumptions of many parametric statistical tests is that samples are drawn from a normal distribution. Alternatively, a normal population may not be required, but instead the statistic of interest (e.g., a mean) is assumed to be described by a normal sampling distribution (i.e., the mean is normally distributed). In either case, the key distinction between parametric and nonparametric (or distribution-free) statistical tests is that a probability model (often normal) is assumed.

Empirical evidence (e.g., Box et al. 1978) indicates that the significance level but not the power is robust (i.e., not greatly affected) to mild violations of a normality assumption for statistical tests concerned with the mean; this should also apply to tests concerned with trend detection. This suggests that a test result indicating "statistical significance" is reliable, but a "nonsignificant" result may be due to lack of robustness to non-normality. Normality of a sample can be checked using normal probability plots, chi square tests, or Kolmogorov-Smirnov tests; unfortunately, many water quality studies often are not designed to produce enough samples to make these tests definitive.

Normality of the sampling distribution for a test statistic is important because it provides a probability model for interval estimation and hypothesis tests that makes use of the test statistic. In many cases, distributional properties of the test statistic could be assessed using Monte Carlo simulation. Alternatively, given the limited robustness to non-normality and the uncertainty in the sampling distributions of selected water quality statistics (e.g., what is the true underlying distribution for a test statistic for model errors?), it may be wise to routinely transform to achieve approximate normality (or symmetry) in a sample, if normality is required. Since non-negative concentration data cannot truly be normal, and since there is empirical evidence to suggest that environmental contaminant data often may be described with a lognormal distribution, the logarithmic transformation is a good first choice. Thus, in the absence of contrary evidence, it is generally recommended that water quality data be log-transformed prior to analysis. This recommendation is often compatible with the conventional approach to model deterministic patterns of variability in the time series data (e.g., a streamflow effect); this is illustrated in Chapter 3.

Equality of Variance

A second common assumption is that, when two or more distributions are involved in a test (i.e., to assess the difference in concentration at two sites), the variances are to be constant across distributions. Many tests are also robust to mild violations of this assumption. Since it is quite unlikely that there will be a need to compare trends in lake water quality variables with widely different variances, this assumption is not addressed further. The interested reader may consult Snedecor and Cochran (1967) for discussion and examples.

Independence

The assumption that is likely to be of greatest concern is that of independence. Most statistical tests (parametric and nonparametric) require a random sample, or a sample composed of independent observations. Dependency between or among observations in a data set means that each observation contains some of the same information already conveyed in other observations. Thus, there is less new, independent information in a dependent data set than in an independent data set of the same sample size. Unfortunately, statistical procedures are often not robust to violation of the independence assumption, so adjustments are generally recommended to address anticipated problems.

Dependence in a sample can result from trend, cyclical patterns, and autocorrelation in the disturbances. One way to mathematically describe a water quality time series is:

$$y_i = \beta_1(time) + \beta_2(\sin\{2\pi(time)\}) + \varepsilon$$

In this expression, we have a linear time trend, a simple seasonal sinusoid, and a disturbance term (ε) that characterizes all remaining unexplained variation in the water quality data. In most types of analyses, the assumption of independence refers to independence in the disturbances; this is the case for time trend hypothesis testing. Thus, autocorrelation or dependence in the data series for the water quality variable (y_i) may exist, but may be due to a deterministic feature of the data (e.g., a time trend or seasonal pattern). This type of autocorrelation poses no difficulty and is addressed by modeling the deterministic feature of the data and subtracting the modeled component from the original series. Of particular concern in testing for trend is autocorrelation that remains (i.e., is in the disturbances) after all deterministic features are removed. When this situation arises, an adjustment to the trend test is necessary; this issue is discussed below.

In the common situation of positive autocorrelation in the disturbances (i.e., each disturbance is positively correlated with nearby disturbances in the series, perhaps due to persistence in behavior over time), confidence interval estimates will be too

narrow and are thus more apt to lead to rejection of the null hypothesis. For simplicity, the common assumption of a lag-one autoregressive structure is often adopted (i.e., each disturbance is correlated with only the immediately preceding disturbance in the series). This assumption is probably reasonable in many situations and might be difficult to reject with the typical short water quality data series of 25-50 observations.

Autocorrelation in the disturbances is the most common and potentially troublesome of the causes of assumption violations. The degree of autocorrelation is a function of the frequency of sampling; this means that a data set based on an irregular sampling frequency cannot be characterized by a single, fixed value for autocorrelation. For water quality time series, stream data obtained more frequently than monthly may be expected to be autocorrelated (after trends and seasonal cycles are removed). Stream water quality data based on less frequent sampling are less likely to exhibit sample autocorrelation estimates of significance.

Autocorrelation in lake water quality data (in the absence of trend and seasonal cycle) may be found at even longer frequencies than in streams and may be expected in data collected on a sampling schedule that is shorter than the hydraulic detention time. This occurs because a lake generally does not act as a "flow-through" system; in-lake mixing may often result in a persistence in behavior over many cycles of the water residence time.

2.4. Statistical Methods for Trend Detection

2.4.1. Summarizing Trend Data

Common statistical estimators are discussed above and in the Appendix; the reader should refer to these sections for explanation of terms. In trend analyses, we may have no observations, one observation, or perhaps a few observations per time interval. If data are missing, there are fill-in methods that may be used for: (1) simple interpolation, (2) estimation based on an assumed probability model (see Gilliom and Helsel, 1986), or (3) estimation based on an assumed autoregressive, moving average model. However, since: (1) interpolation adds no new information, and (2) the two estimation methods require an assumption concerning the underlying parametric model, no special adjustments for missing values are recommended. In effect, relatively few missing values are irrelevant, while a high percentage of missing values is apt to mean that there is too little information for any conclusions in trend testing.

If there is more than one observation per time period, then a summary statistic is needed. The likely options are: (1) select the data point closest to the center of the time interval, or (2) select the median, trimmed mean, or mean of the

observations. Selection of the single data point closest to the center of the time period is the simplest option, but it has the disadvantage of losing the information from the observations not used. If the number of observations per time period is essentially the same within each time period, then it is recommended that a median or trimmed mean be used[1]. However, if the number varies substantially among time periods, then heteroscedasticity (non-constant variance) may be a problem since the location statistics for the time periods will be based on different sample sizes. Van Belle and Hughes (1984) note that the resultant heteroscedasticity does not affect the distribution of the trend test statistic under the null hypothesis, but the effect on test power is uncertain. Thus, a safe approach is to use the median or trimmed mean if the number of data points per time period does not differ greatly, and to use the data point closest to the center if sample sizes differ substantially. Finally, if the number of data points per time period is n_t which is always greater than one, then summarize each time period with the median (or trimmed mean) of the n_t data points closest to the center of the period.

2.4.2. Graphical Methods

Once the time series data have been prepared for analysis, they should be examined graphically using some or all of the methods described in the appendix. A bivariate plot of concentration versus time gives a visual perspective of trend. Since water quality concentration data are often skewed-right, and large outliers are more troublesome than are small outliers, it may be wise to log-transform the concentration data before plotting. In addition, the smoothing spline in SASGRAPH may help the eye see patterns in the data.

Bivariate scatter plots are also useful for examination of deterministic patterns other than those associated with time (e.g., temporal trends and seasonality). For example, there may be a deterministic relationship between water inflow and concentration in river-run lakes, or perhaps dam operating policy in an impoundment has a systematic effect on water quality. Identification of the effect of these forcing functions may be enhanced with graphics.

One particularly helpful graph is the box and whisker plot. For example, a time trend may be examined with a set of annual box and whisker plots: one box for each year, with concentration on the vertical axis and year on the horizontal axis. This graph displays the time sequence of annual medians, quartiles, and extremes, which is a more thorough expression of trend than is a simple graph of median versus time alone. Box plots may also be used to visually capture seasonal patterns: one box for each season, with concentration on the vertical axis and season on the horizontal axis. As with annual box plots, the sequence of seasonal

---

[1] The median may be preferred because it is invariant under transformation (or nearly invariant when there is an even number of observations); e.g., the ordering, and hence the middle value, do not change under a log-transform.

medians, quartiles, and extremes may be extremely helpful in diagnosing seasonal patterns.

2.4.3. Parametric Methods and Tests

Parametric approaches in trend detection involve a model for the trend and a probability model for the errors. The model for the trend is typically a linear, curvilinear, or step function, while the model for the errors is typically a normal probability distribution with independent, identically-distributed errors. If the trend is believed to be continuous (linear or curvilinear), ordinary least squares regression may be applied to fit a continuous trend model, and the test of trend would be based on the statistical significance of the regression parameters. If the trend is believed to be abrupt (step function), a t-statistic may be used to evaluate a step trend (Lettenmaier 1976; Montgomery and Loftis 1987). If seasonal patterns and autocorrelation are present in a time series data set (in addition to a possible trend), then autoregressive, integrated, moving average models (ARIMA, or Box-Jenkins, models) may be the appropriate parametric modeling choice (Pankratz 1983).

The parametric approach is appropriate if the trend model is a reasonable characterization of reality and if the model for the errors holds. The advantage to the parametric approach is that, if the models hold, the statistical tests for trend should be more powerful that distribution-free alternatives. Thus, the assumption that trend and probability models are correct is the basis on which the superior performance of parametric methods rest. If the assumptions concerning these models are incorrect, then the results of the parametric tests may be invalid and distribution-free procedures may be more appropriate.

Given the features of water quality data identified in the previous section, parametric trend modeling often begins with seasonal adjustment or a model (perhaps sinusoidal) for the seasonal pattern. In addition, other deterministic features of the data, such as a predictable relationship between concentration and streamflow, should be modeled. These (and any other) deterministic causes of water quality variability need to be explicitly modeled. In doing so, the non-trend variability in the data can then be removed, or subtracted, from the raw data, which reduces the background variability. This means that the "noise" component is smaller, so that a "signal" (trend) can be more easily detected.

2.4.4. Distribution-Free Methods and Tests

If there is uncertainty concerning the applicability of the trend model or the model for the errors, or if it is known that one or both of these models does not hold, then distribution-free (or nonparametric) methods should be considered.

Distribution-free methods, as the name suggests, do not require an assumption concerning the underlying probability model for the data generation process. However, an assumption of independence is usually made; thus, autocorrelation can be a serious problem, just as it is a problem for parametric methods and robust methods.

Kendall's Tau or the seasonal Kendall's Tau test (Hirsch et al. 1982, Hirsch and Slack 1984, Gilbert 1987) are often good choices for distribution-free tests. The Kendall's Tau test is used to determine if a time series is moving upward, downward, or remaining relatively level over time. This is accomplished by computing a statistic, based on all possible data pairs, that represents the net direction of movement of the series. To do this, the data are first ordered according to time: $x_1, x_2, x_3,\ldots,x_t,\ldots,x_n$, where t goes from 1 to n. All possible pairs of differences $x_i - x_j$ are calculated, where i>j (observation j precedes observation i in time). This difference will either be positive ($x_i>x_j$), negative ($x_i<x_j$), or zero ($x_i=x_j$) for each of the pairs. The number of positive differences minus the number of negative differences is calculated; this becomes the test statistic (the Mann-Kendall statistic).

If a water quality data series is increasing (decreasing) over time, then $x_i>x_j$ ($x_i<x_j$) for most pairs and the test statistic will be a large positive (negative) number. If the trend in the water quality data series over time is negligible, then the number of positive pairs and the number of negative pairs will be essentially equal, and the test statistic will be small in absolute value. For small sample sizes (n<40) the Mann-Kendall statistic is tabulated in most nonparametric statistics texts; for large sample sizes (typical of most applications for water quality trends) a normal approximation may be used as shown in the examples in Chapter 3.

The seasonal Kendall's Tau test yields the same analysis on a seasonal basis; with monthly data, we may define a "season" as one month in length. If there is distinct seasonal behavior, then the seasonal Kendall's Tau test is a good choice. Berryman et al. (1988) or Gilbert (1987) provide useful guidance on the selection and application of tests.

The Kendall's Tau statistic provides a nonparametric assessment of the presence or absence of a trend. For a nonparametric estimate of the **magnitude** of trend, the Sen or seasonal Kendall slope estimator (Gilbert 1987) are good choices. These estimators are based on the median slope from the set of slope estimates for the lines connecting all possible pairs of data.

Finally, it must be noted that the nonparametric tests and statistics are appropriate if the parametric assumptions cannot be justified; otherwise, the parametric procedures are more powerful. Since the parametric assumptions are often questionable with water quality data, and since the nonparametric approaches are almost as effective as the parametric methods when the assumptions are correct, it is our belief that nonparametric procedures should be routinely used for

trend detection (and parametric procedures used only when justified). This recommendation is based on: (1) concern for the effects of non-normality, (2) concern for the effects of occasional outliers in water quality data, (3) the realization that nonparametric methods are becoming "standard practice" in water quality trend detection studies.

Chapter 3

Trend Detection in Lakes - Examples and Discussion

3.1. Introduction

The discussion of basic statistical and graphical methods in Chapter 2 serves an important purpose in a comprehensive approach to trend analysis. To be specific, it is strongly recommended that certain graphs of the data be examined, and that specific statistics be calculated, before the trend detection test is run. In most cases, this preliminary analysis provides useful information and possible adjustments to the data that result in improvements in the trend detection test. Some examples are:

1) A bivariate times series graph may indicate presence or absence of seasonal variation. This helps determine the need for seasonal adjustment and for the seasonal version of the Kendall Tau test.

2) A bivariate graph of monthly (or weekly) stream inflow versus monthly (or weekly) water quality concentration may indicate a flow-effect, particularly in lakes with short hydraulic detention times. This helps determine the need for a flow-concentration model to reduce background variability.

3) A bivariate time series graph, and/or a histogram for the water quality variable, will indicate the presence or absence of extreme observations (outliers). This helps determine the need for a transformation and/or for a nonparametric test of trend.

4) Seasonal (yearly) boxplots for the water quality variable will show the range, median, upper and lower quartiles, and confidence interval for the median, for each of the seasons (years) plotted. This can provide a visual indication of the presence or absence of trend or seasonality.

These and other issues are illustrated in examples below.

In a parametric test of trend, deterministic features of the water quality time series are often accounted for with separate terms for season, streamflow, and trend in a regression model. In a nonparametric test, some deterministic features of the data (e.g., a flow effect) are modeled with a simple parametric regression model, and some deterministic features (e.g., seasonality) are adjusted for in a

nonparametric manner. In either case, the analyst would usually like to assume that the "unexplained" remainder of the water quality data exhibit random (white noise) behavior. However, autocorrelation, or persistence, in the water quality time series may still be present. As stated in Section 2.3.2, autocorrelation indicates that each observation in a time series is not independent of other observations. In the most common case of lag-one autocorrelation, each observation is correlated with the previous observation. This means that some of the information that is conveyed in the current observation has already been conveyed in the previous observation. Thus, with autocorrelation, we do not have as much information that we believe we have on the basis of sample size.

It is important to consider the nature of the model and the possible causes of autocorrelation when examining a data series for autocorrelation. For example, a data series with a strong linear trend or seasonal cycle is likely to yield large value(s) for autocorrelation at one or more lags. These are apt to reflect the deterministic trend, or cycle, and in fact, calculation of autocorrelation is a useful diagnostic device for selecting a time series model. However, for the purpose of trend detection in water quality analyses, autocorrelation is of interest in the data series **after** all deterministic patterns are removed. When autocorrelation still remains at this point in the analysis, then the procedure employed for trend analysis must explicitly account for the autocorrelation.

3.2. Examples - Background

The first example involves trends in total phosphorus and total nitrogen in Falls Lake, North Carolina. Falls Lake is a 10,700 acre lake located in the north central piedmont region of North Carolina. The lake provides flood control, recreation, downstream water quality control, fish and wildlife conservation, and water supply for the Raleigh area. The North Carolina Division of Environmental Management (DEM) collected data from Falls Lake on total phosphorus and total nitrogen on an approximately monthly basis over a five year period from 4/26/83 until 10/14/87.

The flow chart illustrated in Figure 3.1 shows the outline of the procedures followed in this case study, and should serve as a basic model for other trend detection studies. The macros described in this diagram are all found in Appendix D. The chart begins with entry of data into a Statistical Analysis System (SAS) data set. Various statistical procedures and macro programs are then run, depending on the outcome of each step. The final step in the flow chart provides information on trend for variables with (or without) seasonality, while ignoring autocorrelation.

The procedures outlined in Figure 3.1 were first run on the total phosphorus data set. Sections 3.2 through 3.7 guide the reader through these procedures, using the

total phosphorus data set. Following that, section 3.8 presents the results for total nitrogen.

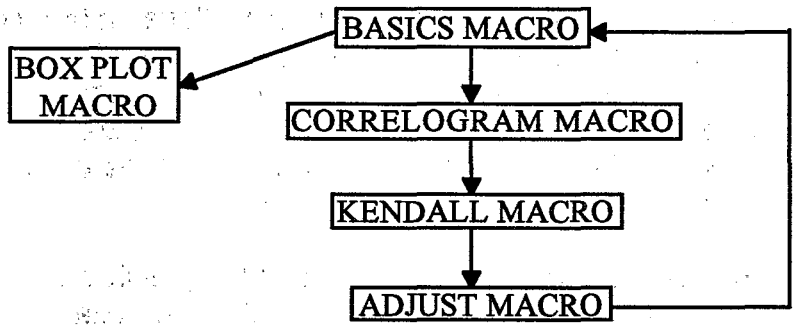


Figure 3.1 Flow chart for macro programs.

To begin, raw data were first entered into a SAS data set containing variables for time (SAS date variable (date)), total phosphorus (TP), and total nitrogen (TN). The date variable was then converted to calendar form and variables for the day, month, and year were created for later use. Table 3.1 shows the completed data set.

As suggested earlier, the data sets were analyzed for basic statistical information, and histograms and bivariate time series graphs were constructed using the macro "Basics" found in Appendix D. The macro was designed to calculate measures of central tendency and dispersion, such as those discussed in Chapter 2 and in Appendix A. For those unfamiliar with SAS, Appendix B contains a brief introduction on how to run a macro on a new data set. With graphics capability (SASGRAPH), the Basics program will also provide a histogram and bivariate time series plot of the data set; otherwise a simple printer plot will be drawn.

Information from the macro Basics was used to evaluate some of the underlying assumptions needed to perform further statistical analyses and tests of significance on the data. Specifically, the probability distribution approximated by the data (normal, lognormal,...), the presence or absence of seasonal cycles within the data, and the dependence of each observation on previous observations (autocorrelation). The importance of normality and independence is discussed in Section 2.3.2., and seasonality is discussed in Sections 2.4.3. and 2.4.4. Other patterns in the data, such as a deterministic flow/concentration relationship, should also be considered (depending on lake detention time and reaction rates). In this case study, corresponding information on inflow to the lake was not available. For the total phosphorus data set used in this case study, the evaluations of normality, seasonality, and independence, are discussed in Sections 3.5 through 3.7.

OBS	OBS	DATE	TP	TN	NEWDATE	MONTH	DAY	YEAR
1	1	8426	.	.	01/26/83	01	26	83
2	2	8456	.	.	02/25/83	02	25	83
3	3	8486	.	.	03/27/83	03	27	83
4	4	8516	0.07	0.58	04/26/83	04	26	83
5	5	8539	0.04	0.36	05/19/83	05	19	83
6	6	8579	0.05	0.36	06/28/83	06	28	83
7	7	8609	0.04	0.31	07/28/83	07	28	83
8	8	8643	0.04	0.51	08/31/83	08	31	83
9	19	8670	0.03	0.42	09/27/83	09	27	83
10	10	8699	0.03	0.71	10/26/83	10	26	83
11	11	8733	0.03	0.87	11/29/83	11	29	83
12	12	8748	.	.	12/14/83	12	14	83
13	13	8774	0.07	0.85	01/09/84	01	09	84
14	14	8805	0.09	0.76	02/09/84	02	09	84
15	15	8839	0.10	0.72	03/14/84	03	14	84
16	16	8873	0.07	0.65	04/17/84	04	17	84
17	17	8901	0.04	0.45	05/15/84	05	15	84
18	18	8931	0.03	0.41	06/14/84	06	14	84
19	19	8966	0.02	0.41	07/19/84	07	19	84
20	20	8994	0.05	0.41	08/16/84	08	16	84
21	21	9026	0.04	0.51	09/17/84	09	17	84
22	22	9062	0.04	0.62	10/23/84	10	23	84
23	23	9085	0.04	0.73	11/15/84	11	15	84
24	24	9118	0.03	0.63	12/18/84	12	18	84
25	25	9161	0.03	0.85	01/30/85	01	30	85
26	26	9183	0.09	0.93	02/21/85	02	21	85
27	27	9210	0.05	0.71	03/20/85	03	20	85
28	28	9238	0.03	0.50	04/17/85	04	17	85
29	29	9252	0.03	0.33	05/01/85	05	01	85
30	30	9294	0.02	0.41	06/12/85	06	12	85
31	31	9322	0.04	0.41	07/10/85	07	10	85
32	32	9356	0.03	0.51	08/13/85	08	13	85
33	33	9398	0.02	0.41	09/24/85	09	24	85
34	34	9427	0.02	0.73	10/23/85	10	23	85
35	35	9449	0.03	0.94	11/14/85	11	14	85
36	36	9482	0.05	0.74	12/17/85	12	17	85
37	37	9512	0.04	0.76	01/16/86	01	16	86
38	38	9547	0.06	0.87	02/20/86	02	20	86
39	39	9567	0.06	0.80	03/12/86	03	12	86
40	40	9614	0.04	0.40	04/28/86	04	28	86
41	41	9643	0.04	0.31	05/27/86	05	27	86
42	42	9671	0.03	0.41	06/24/86	06	24	86
43	43	9693	0.02	0.41	07/16/86	07	16	86
44	44	9714	0.03	0.41	08/06/86	08	06	86
45	45	9769	0.01	0.41	09/30/86	09	30	86
46	46	9797	0.02	0.61	10/28/86	10	28	86
47	47	9818	0.03	0.94	11/18/86	11	18	86
48	48	9845	0.02	0.88	12/15/86	12	15	86
49	49	9876	.	.	01/15/87	01	15	87
50	50	9895	0.08	0.85	02/03/87	02	03	87
51	51	9945	0.08	0.55	03/25/87	03	25	87
52	52	9958	0.04	0.44	04/07/87	04	07	87
53	53	9994	0.04	0.41	05/13/87	05	13	87
54	54	10016	0.03	0.41	06/04/87	06	04	87

Table 3.1: Completed Data Set

55	55	10043	0.04	0.21	07/01/87	07	01	87
56	56	10085	0.02	0.41	08/12/87	08	12	87
57	57	10119	0.03	0.52	09/15/87	09	15	87
58	58	10148	0.02	0.81	10/14/87	10	14	87
59	59	10178	.	.	11/13/87	11	13	87
60	60	10208	.	.	12/13/87	12	13	87

3.3. Summary Statistics

The moments printed under the univariate procedure (Table 3.2) are explained in SAS User's Guide: Basics (1989). They contain many of the measures of central tendency and dispersion discussed in Appendix A. In addition to the moments, quantiles and extreme values are also calculated. The univariate procedure also draws a stem and leaf diagram, a box and whiskers plot, and a normal probability graph. Figure 3.2 contains these graphs for total phosphorus. All three diagrams provide a visual indication of the distribution of the data set, and their construction is documented in the SAS User's Guide (1989).

The skew and kurtosis calculations are documented in most statistics texts, as well as in the SAS procedures guide. The test for normality (W statistic) is based upon the null hypothesis that the data values are a random sample from a normal distribution. The test calculates the Shapiro-Wilk statistic, W, which must be greater than zero and less than or equal to one, with small values of W leading to rejection of the null hypothesis. The value for W of .859662 is small enough (indicated by the PROB < W of 0.0001) to require us to reject the null hypothesis of a normal distribution for the total phosphorus data set. See Gilbert (1987), for further discussion of the W statistic.

The stem and leaf diagram in Figure 3.2 gives an indication that the data set is slightly skewed, because of the concentration of data in the lower portion and the spread of the upper portion of the diagram. The box and whiskers plot shows less evidence of skew, with the mean and median falling on the same line, and the upper whisker just slightly longer than the lower one. However, the two circles above the box do indicate that there are values greater than 1.5 times the interquartile range from the median, which would indicate a skewed data set (see Appendix A for a discussion of interquartile range).

The normal probability plot found in Figure 3.2 is a graph of the probability density function. In this figure, the empirical data (the observations) are plotted against a standard normal density function with the same (sample) mean and standard deviation. The asterisks (*) symbolize the observations; if the data follow a normal distribution, then the asterisks will fall in a straight line along the same path as the normal function (+ symbols). As Figure 3.2 shows, the data do not follow a normal distribution.

3.4. Graphical Analyses

In addition to the diagrams already mentioned, a histogram and a bivariate time series graph of the data (including an estimated trend line) are constructed within the macro BASICS. In the absence of graphics capability, a simple printer plot of the data over time will be drawn as the bivariate graph. With graphics capability,

UNIVARIATE PROCEDURE

ariable=TP

Moments

N	53	Sum Wgts	53
Mean	0.040943	Sum	2.17
Std Dev	0.020406	Variance	0.000416
Skewness	1.222268	Kurtosis	1.013003
USS	0.1105	CSS	0.021653
CV	49.83928	Std Mean	0.002803
T:Mean=0	14.60717	Prob>|T|	0.0001
Sgn Rank	715.5	Prob>|S|	0.0001
Num ^= 0	53		
W:Normal	0.859662	Prob<W	0.0001

Quantiles(Def=5)

100% Max	0.1	99%	0.1
75% Q3	0.05	95%	0.09
50% Med	0.04	90%	0.07
25% Q1	0.03	10%	0.02
0% Min	0.01	5%	0.02
		1%	0.01
Range	0.09		
Q3-Q1	0.02		
Mode	0.03		

Extremes

Lowest	Obs	Highest	Obs
0.01(	45)	0.08(	50)
0.02(	58)	0.08(	51)
0.02(	56)	0.09(	14)
0.02(	48)	0.09(	26)
0.02(	46)	0.1(	15)

Missing Value	
Count	7
% Count/Nobs	11.67

Table 3 .2: Univariate Moment Statistics

UNIVARIATE PROCEDURE

Variable=TP

Stem	Leaf	#	Boxplot
10	0	1	0
9			
9	00	2	0
8			
8	00	2	|
7			|
7	000	3	|
6			|
6	00	2	|
5			|
5	0000	4	+-----+
4			| |
4	00000000000000	14	*--+--*
3			| |
3	000000000000000	15	+-----+
2			|
2	000000000	9	|
1			|
1	0	1	|

----+----+----+----+

Multiply Stem.Leaf by 10**-2

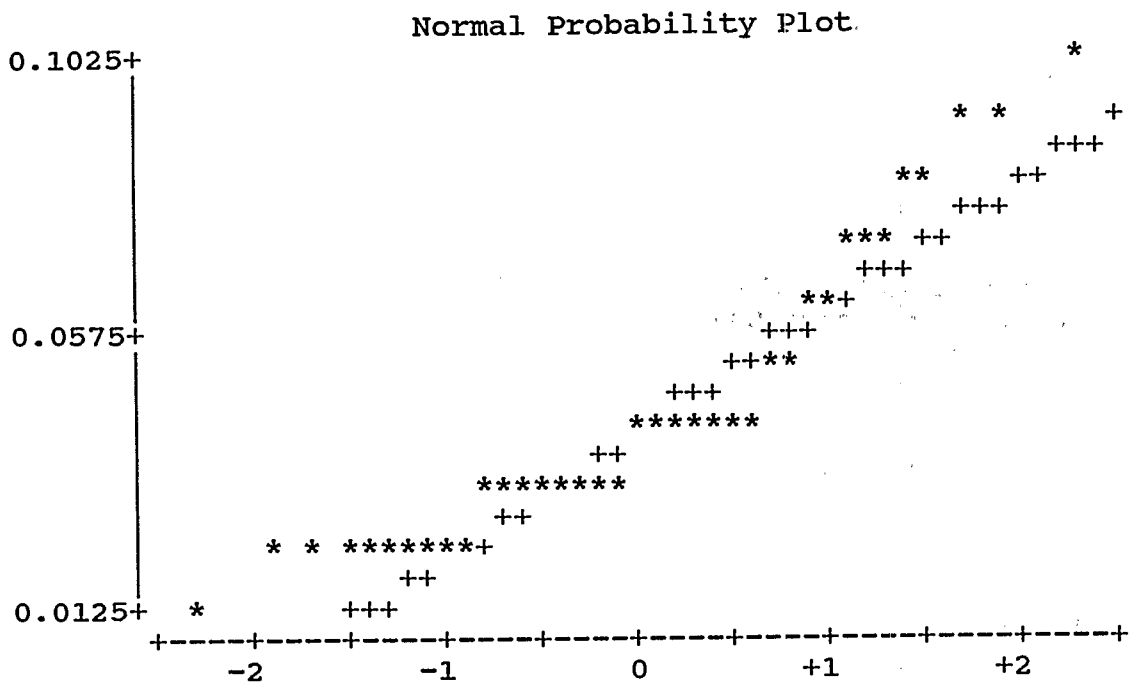


Figure 3.2

additional macros can be run to provide the user with boxplots of the chosen variable. These macros are presented in Appendix D (and in the accompanying disk) under the "Boxplt.sas" program listing. The macros are: (1) NOBS, (2) ORDER, (3) BOXVARS, and (4) BOXPLOT. Information and examples using these macros can be found below and in Appendix D.

3.4.1. Histogram

The histogram in Figure 3.3 was constructed according to the procedure outlined in Appendix A. The range and other information needed to construct the histogram were obtained from an initial run of the univariate procedure. These results were then entered into the macro BASICS, to allow for bars of uniform width within the given range. The initial run did not specify midpoints of bar widths, which can be added as options (see SAS User's Guide, 1989).

When constructed correctly (for a sample containing bars of equal width, as outlined in Appendix A), the histogram can also be used to determine the normality of a data set. The histogram found in Figure 3.3 provides an informative picture of the distribution of total phosphorus data.

3.4.2. Time Series Graph

Bivariate time series graphs are discussed in the Appendix. For the case study, Figure 3.4 is a bivariate plot of total phosphorus concentration over time. The cyclic pattern it displays is an indication of seasonality in the data. The peak values occur in spring, and the lowest in the fall of each year. The graph also includes a predicted trend line based on a simple (ordinary least squares) regression of concentration over time. It can be used to give an indication of the presence of trend, but should not be relied upon because it does not take into account the presence of seasonality or autocorrelation.

3.4.3. Box Plot

The construction and use of box plots is discussed in Appendix A. Figure 3.5 presents seasonal box plots for total phosphorus from Falls Lake, with the X-axis as the seasonal variable (MONTH), and the Y-axis the water quality variable being studied. In this figure, a notched box plot is drawn separately for each month. See the box and whiskers plot discussion in Appendix A for a more complete description of the construction of a box plot.

When compared vertically, the notches for all of the boxes illustrated in Figure 3.5 do not overlap. This means that the medians for some seasons (months) are

FREQUENCY OF TP

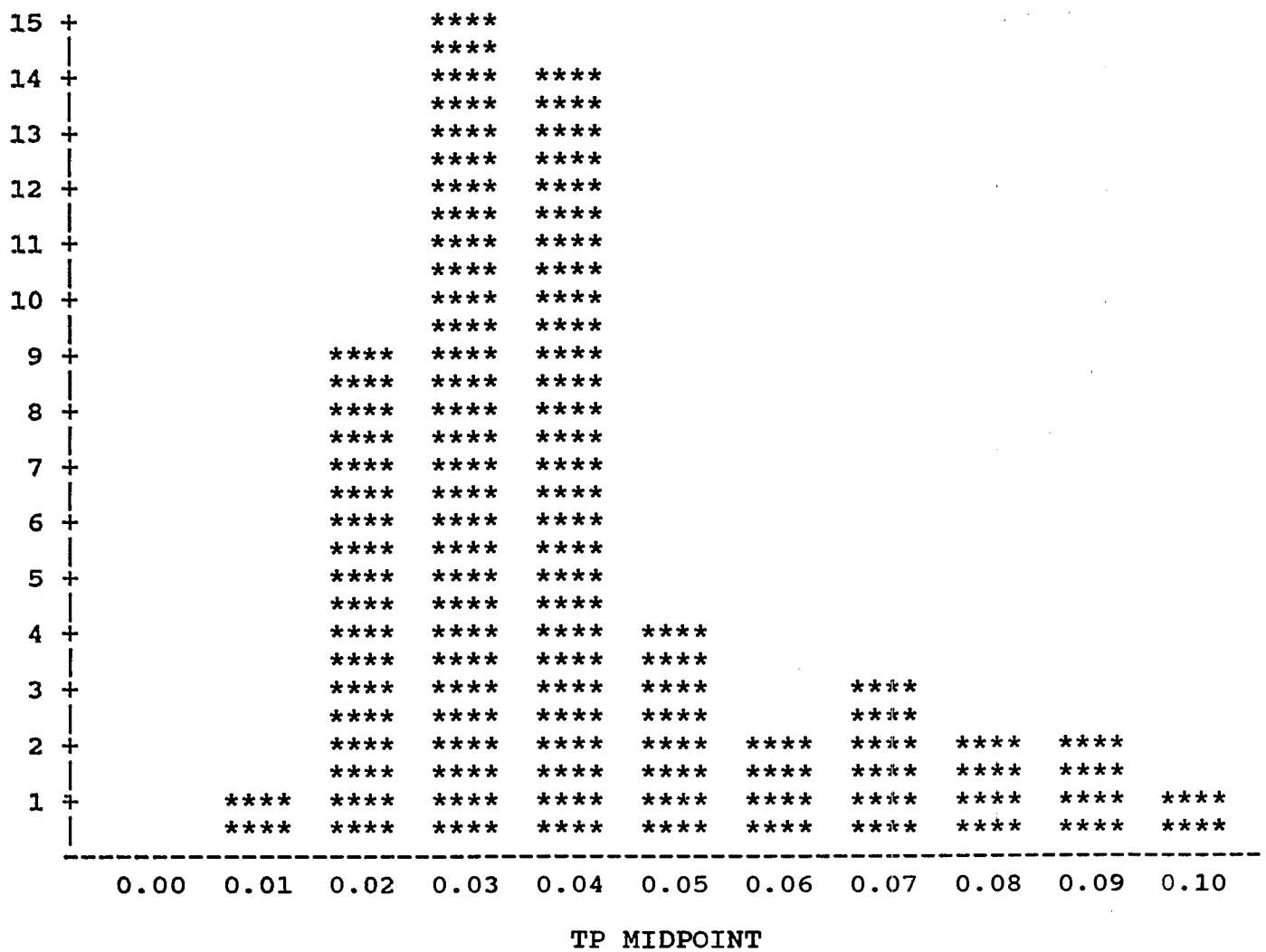


Figure 3.3

Figure 3.4. FALLS TEST DATA
Plot of Observed and Linear Regression Model Predictions Against Time

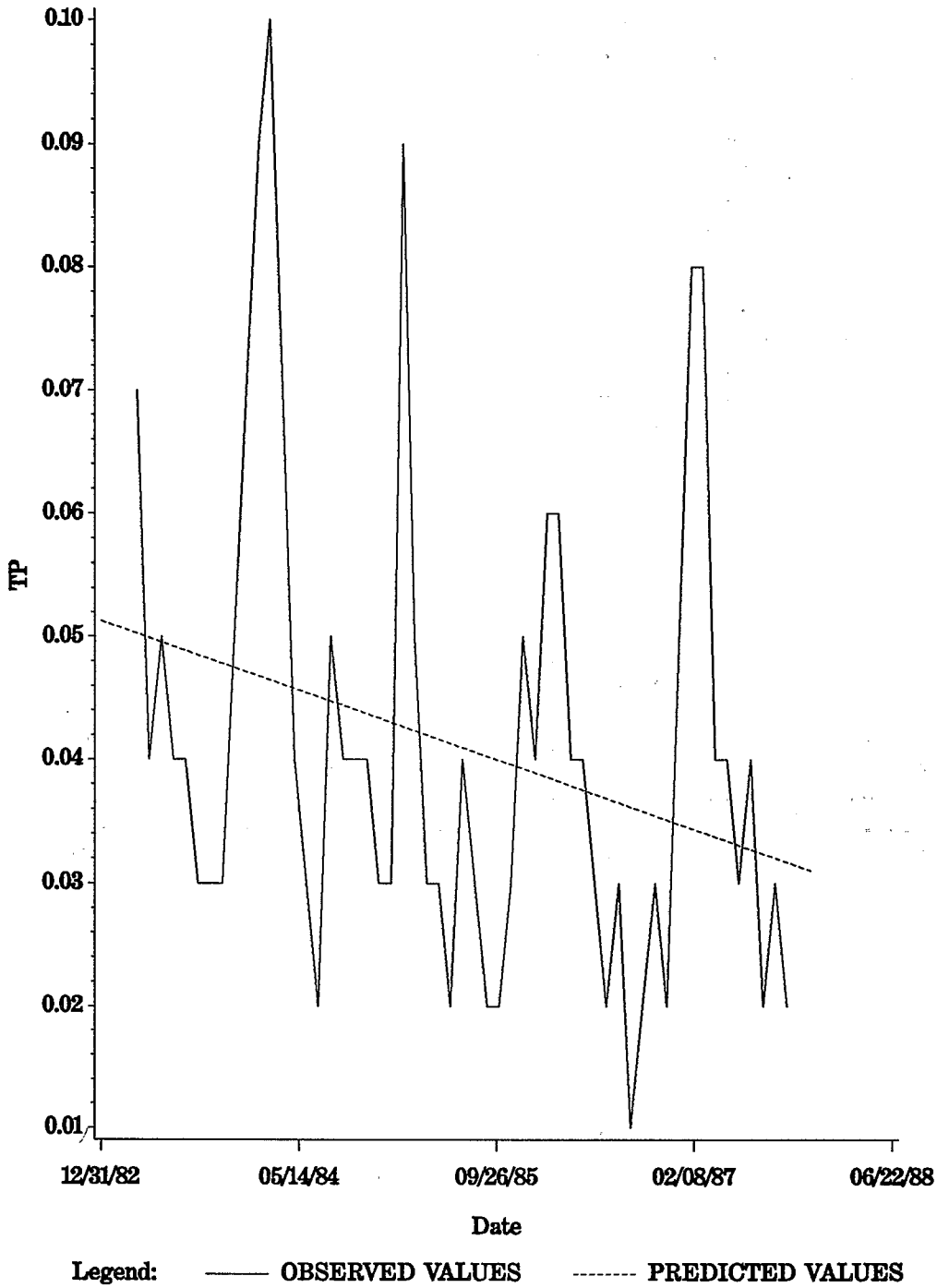
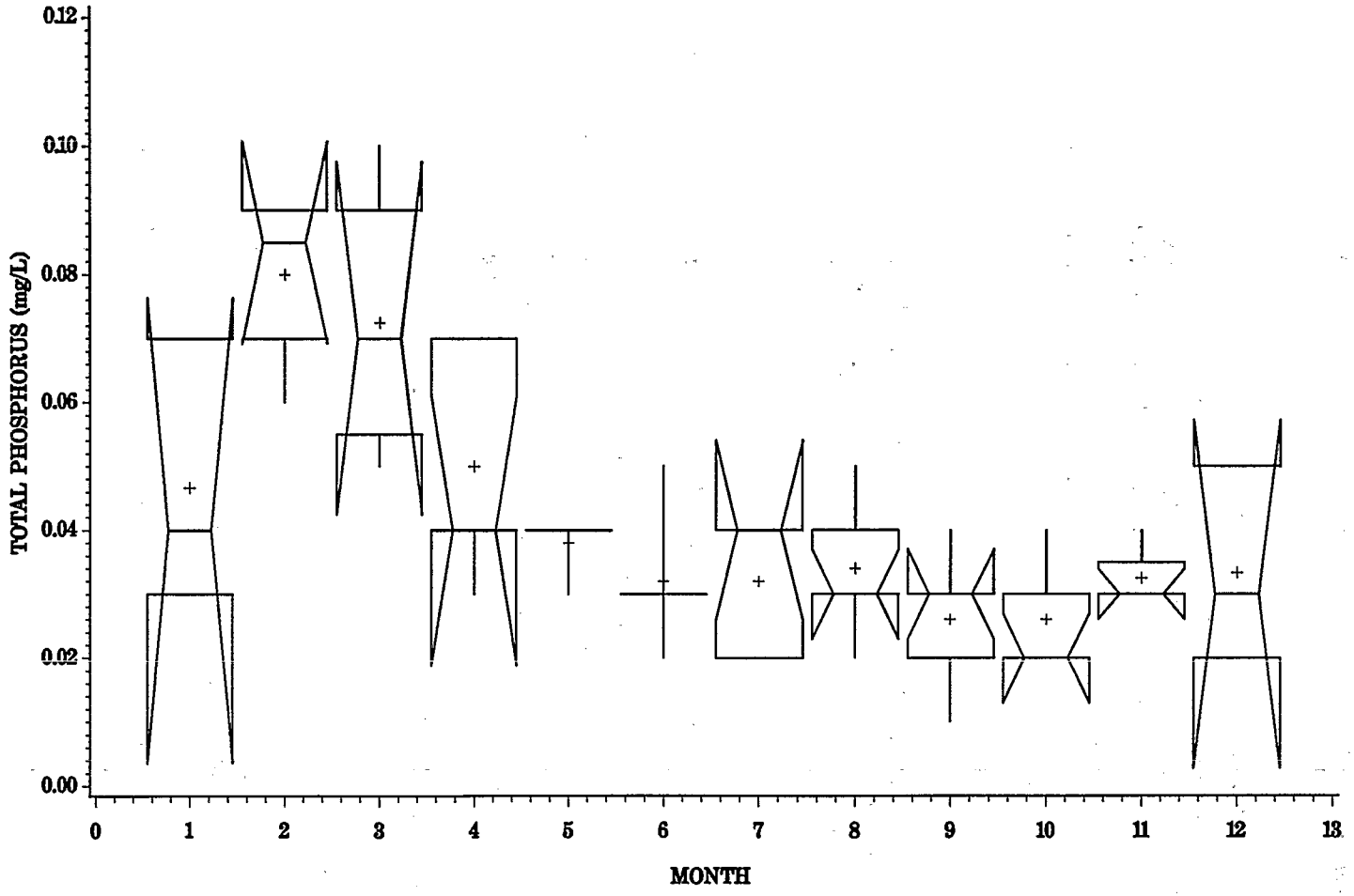


Figure 3.5 Falls Lake Data
Seasonal Boxplots for Total Phosphorus



significantly different at an approximate alpha level of 0.05, which is an indication of seasonality. Also note that the height of some boxes is larger than that for others. This indicates that those months have a greater variability in total phosphorus values.

Figure 3.6 illustrates the yearly box plots for the same data set. The notches of these boxes do overlap vertically. This means that the yearly median values for total phosphorus do not show statistically significant differences at an alpha level of 0.05. Moreover, the yearly box plots do not show a trend in the median value for total phosphorus over the years. Of course, trend in the median value is only one trend of interest; the box plots also display trend in extremes (minimum and maximum) and trends in variability.

3.5. Normality

The importance of normality is discussed in Section 2.3.2. As noted in that section, a normal distribution is required for many hypothesis testing procedures and for parametric tests of trend. Normality is also required for most tests of autocorrelation; thus a log transformation of the data may be appropriate when testing for the presence of autocorrelation.

The visual image supplied by the histogram, combined with the information from the univariate procedure, led to the conclusion that the total phosphorus data do not follow a normal distribution. Thus, the decision to apply a nonparametric (distribution free) trend detection method in this case study was based on the lack of normality, combined with the presence of missing values. It should be noted that in some cases, water quality concentration will be reported at or below a specified detection level, reported as a missing value, or will yield a skewed histogram. These data sets should either be transformed, or analyzed using a distribution free method (such as the one used in this case study).

3.6. Seasonality

Figure 3.4 is a bivariate time series graph for total phosphorus. The graph indicates the possible presence of seasonality in the data set by the cyclic pattern of phosphorus concentration over time. Since the graph indicated the possibility of seasonality, the data set was run through the macro "CORR" to construct a correlogram and print the autocorrelation values.

Autocorrelation (or correlation over time) can be thought of as an indicator of persistence in behavior, or how similar one data point (e.g., observation or residual; see Section 2.3.2) is to other data points taken at nearby time periods. Autocorrelation is expressed in terms of time lags; for example, lag1 autocorrelation refers to correlation between data points one sampling period apart. Similarly, lag12

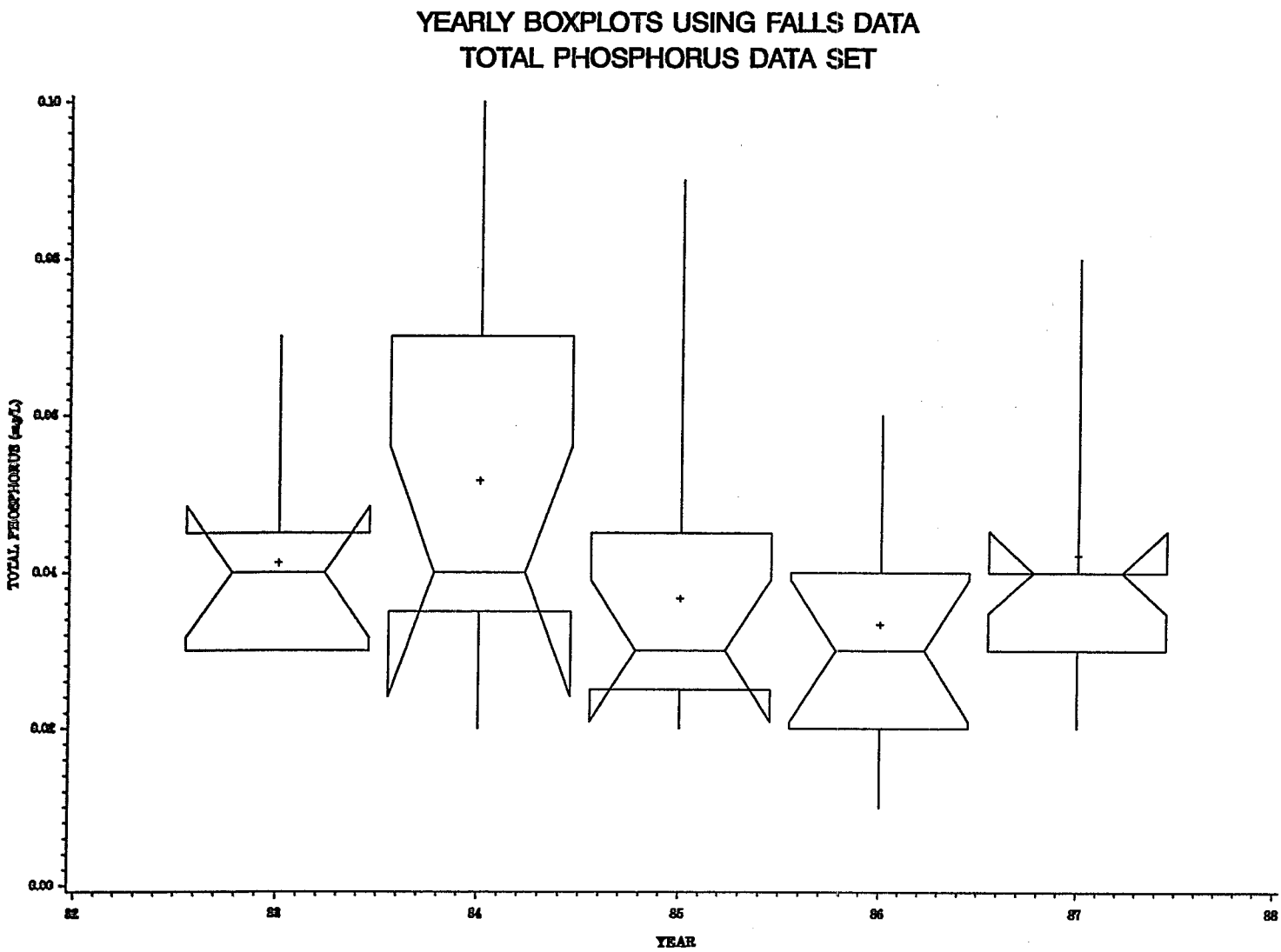

YEARLY BOXPLOTS USING FALLS DATA
TOTAL PHOSPHORUS DATA SET
TOTAL PHOSPHORUS (mg/L)
0.10
0.06
0.04
0.02
0.00
82
83
84
85
86
87
88
YEAR

Figure 3.6

autocorrelation refers to correlation between data points twelve sampling periods apart. For water quality data, positive lag1 autocorrelation is most common, and indicates a persistence in behavior between data points adjacent in time. Likewise, a positive lag12 autocorrelation for monthly water quality data is indicative of cyclical behavior that repeats every twelve months (seasonality). A negative autocorrelation at six months lag is also indicative of a twelve-month (seasonal) cycle, as it suggests an opposite response (e.g., high chlorophyll in summer and low chlorophyll in winter) six months apart.

A correlogram is a graphical illustration of the autocorrelation values versus the lag. The correlogram for total phosphorus, presented in Figure 3.7, clearly displays a twelve month cyclical pattern. The value of .56642 for the lag12 autocorrelation is close to the upper limit of significance (0.05 level) of .57692 (which means that it is almost two standard deviations away from a zero autocorrelation value). This correlogram was constructed according to the equations found in Pankratz (1983), using the raw data set. Recall from the discussion in Chapter 2 that, for tests of trend, autocorrelation becomes a concern only after all deterministic patterns (including seasonality) are removed. Thus autocorrelation at this point in the example should cause no alarm, and in fact aids in the diagnosis of seasonality. This issue is treated further in Section 3.7.

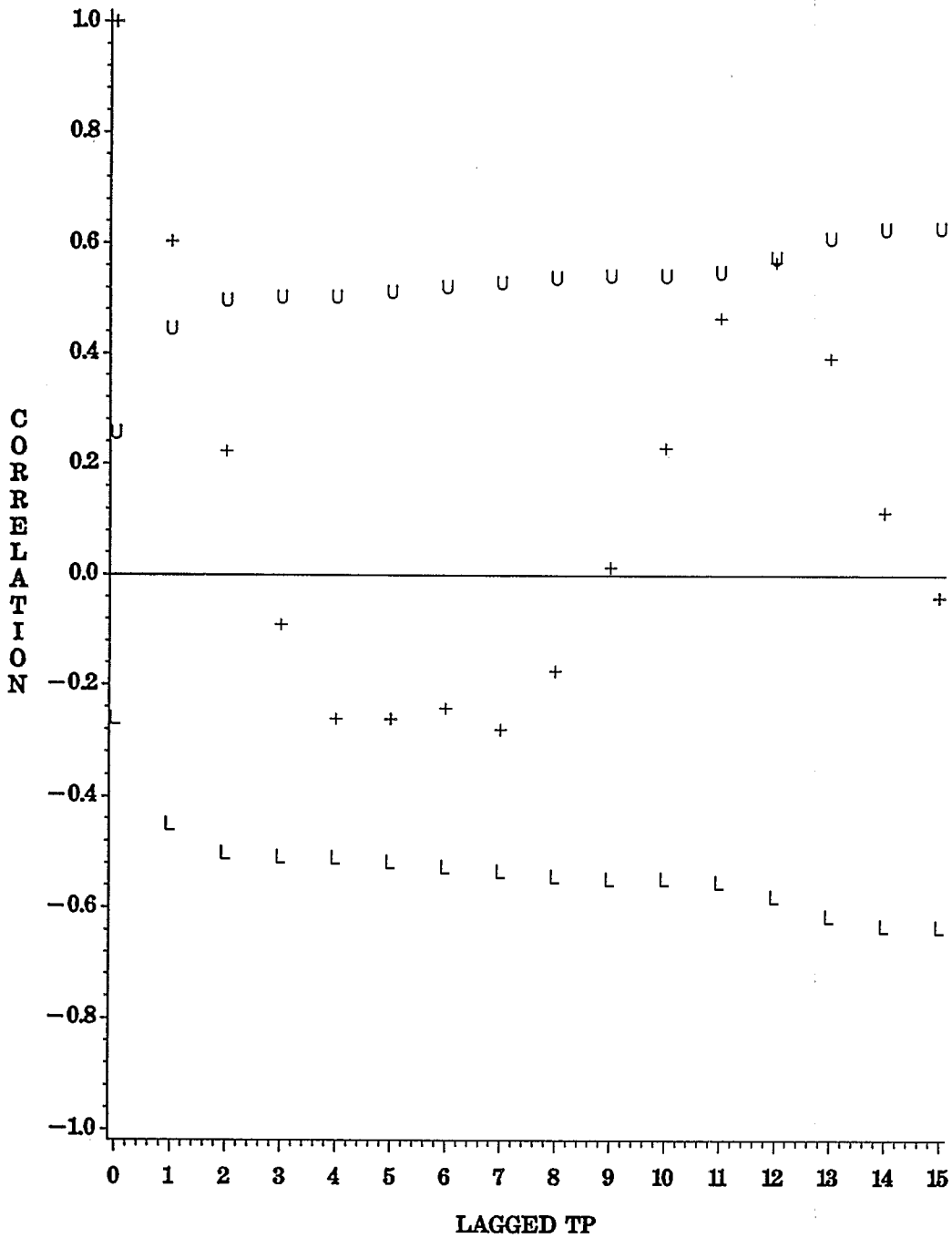
The negative correlations with values six months apart, and strong positive correlations with values twelve months apart, shown in Table 3.3 and Figure 3.7, imply seasonality in the data set. These indications combined with the time series graph (Figure 3.4) demonstrate a clear picture of the presence of seasonality in total phosphorus. Thus, any trend detection method used must account for or remove the effects of seasonality in order to get an accurate measure of trend in the data.

3.7. Independence

As noted in the discussion in Section 2.3.2, violation of the independence assumption for tests of trend refers not to the raw water quality data, but to the residuals left after the removal of all identified deterministic patterns, including seasonality and trend. Thus, to properly select and apply the test for trend, the analyst must first model and remove the very trend that he\she is trying to estimate. The objective of this task is to assess the independence assumption, so that the appropriate trend test may be chosen.

Figure 3.8 presents the sequence of data analyses leading to the selection and application of the test for trend. Starting from the top of the figure, the analyst is guided through a set of questions and statistical analyses intended to remove all known deterministic features from the original water quality data series. Once this is completed, the data\residuals are tested for autocorrelation. If autocorrelation is rejected, then the standard Kendall test is applied for trend; if autocorrelation is not

Figure 3.7. FALLS LAKE DATA
CORRELOGRAM WITH UPPER AND LOWER 95% CONFIDENCE LIMITS



OBS	LAGGED TP	CORRELATION	STE	UPPER LIMIT	LOWER LIMIT
1	0	1.00000	0.12910	0.25820	-0.25820
2	1	0.60342	0.22361	0.44721	-0.44721
3	2	0.22298	0.24927	0.49855	-0.49855
4	3	-0.09027	0.25258	0.50515	-0.50515
5	4	-0.26161	0.25311	0.50623	-0.50623
6	5	-0.26187	0.25758	0.51516	-0.51516
7	6	-0.24185	0.26198	0.52396	-0.52396
8	7	-0.28056	0.26567	0.53135	-0.53135
9	8	-0.17439	0.27057	0.54113	-0.54113
10	9	0.01384	0.27243	0.54487	-0.54487
11	10	0.22991	0.27245	0.54489	-0.54489
12	11	0.46544	0.27566	0.55132	-0.55132
13	12	0.56642	0.28846	0.57692	-0.57692
14	13	0.39239	0.30644	0.61288	-0.61288
15	14	0.11456	0.31470	0.62940	-0.62940
16	15	-0.03949	0.31540	0.63079	-0.63079

Table 3.3

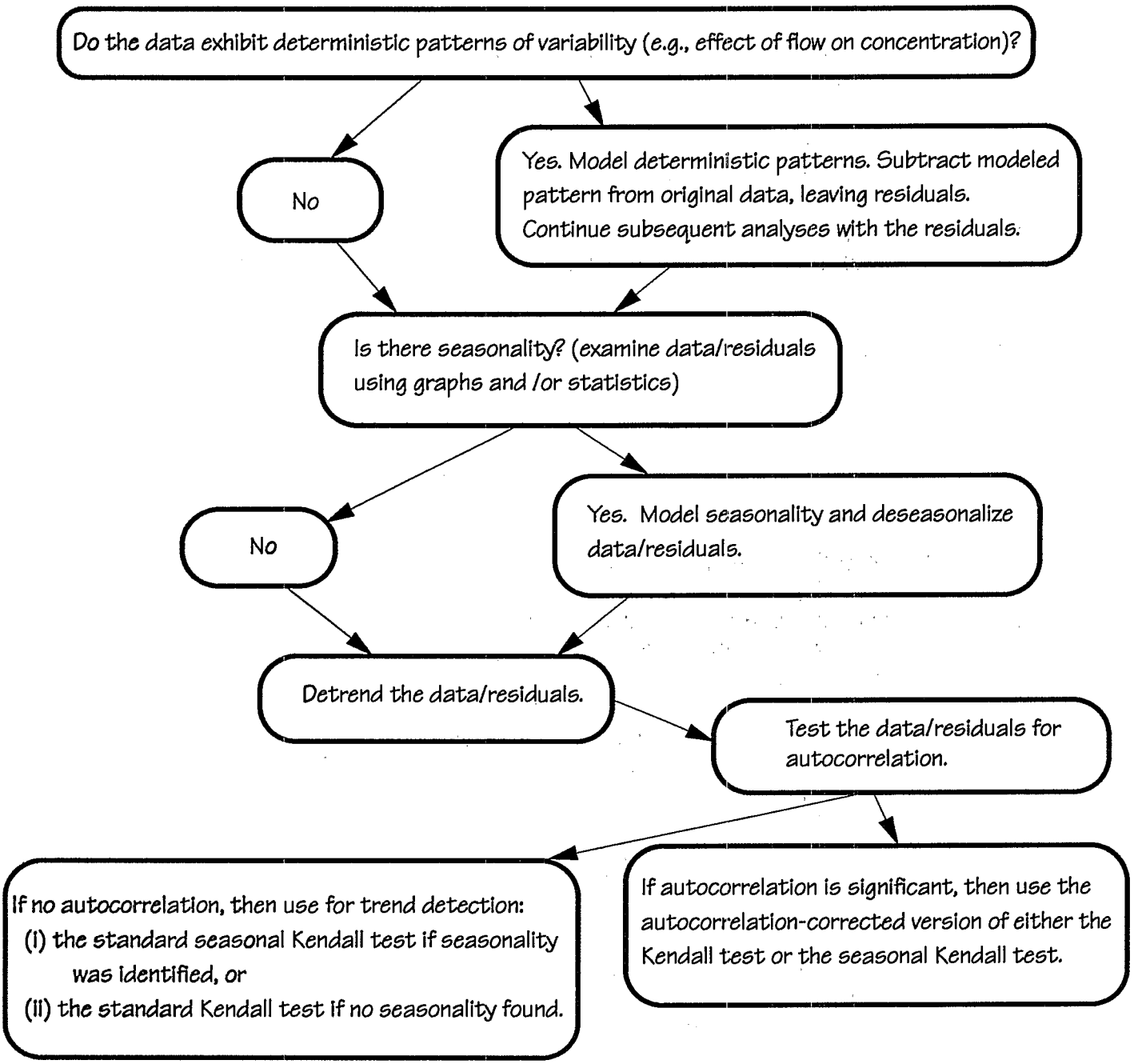


Figure 3.8. Data analyses used to identify the appropriate test for trend detection.

rejected, then the autocorrelation-corrected Kendall test is used. See the discussion of independence in Section 2.3.2 for the explanation of this strategy for analysis.

Large (relative to the 0.05 significance level) positive values for lag1 and lag2 autocorrelations (after the removal of seasonality and any trend from the data) are the most common indicators of serial correlation. Since these lag values represent observations one or two months away from the observation in question, they may be related to lake detention time. In examining the correlogram for autocorrelation, rules-of-thumb adopted here are to:

- look at the general shape of the correlogram for deterministic patterns (e.g., is seasonality still evident?) that might have been missed in earlier analyses,
- consider that autocorrelation will be positive and will occur only at lag1, lag2, or lag12, if at all,
- use the 0.05 significance level for autocorrelation as the cutoff for presence/absence of serial correlation.

These rules are suggested to help avoid misinterpreting the correlogram. Many autocorrelations are presented simultaneously in a correlogram, so some are apt to appear significant (0.05 level) purely by chance. Faulty inferences can be minimized by taking advantage of expectations concerning water quality data. For example, autocorrelation is expected to be positive (indicating persistence associated with a common unexplained phenomena) and to be highest at low lags (indicating month-to-month persistence) and/or highest at annual lags (indicating year-to-year persistence). Other "significant" autocorrelations will generally be assumed to be by chance and thus ignored.

To remove seasonality and trend in order to check for the presence of autocorrelation, the data set was first run through the macro "KENS", which makes use of the "Kendall" Fortran program. This macro determines the seasonal Kendall Tau test statistic, the significance of that statistic (with and without a correction for the covariance caused by autocorrelation), and the seasonal Sen slope estimate for the trend. Table 3.4 contains these values for the total phosphorus data set. The formulas used in the calculation of these statistics are documented in Hirsch and Slack (1984).

The seasonal Sen slope estimate was then used with the seasonal median to deseasonalize and detrend the data, in the macro "ADJUST". It should be noted that the median value (as opposed to the mean value) was used to provide resistance to outliers. The macro program detrended the data by subtracting the trend line estimated using the seasonal Sen value for slope (see Appendices B and D, ADJUST macro, for the formula used). The output from the macro ADJUST was then run through the macro CORR (the "corradj.sas" program) again, to construct a correlogram from the deseasonalized, detrended data. This correlogram was then

OBS	TAU STATISTIC	P-VALUE WITHOUT SERIAL CORRELATION	P-VALUE WITH SERIAL CORRELATION	SLOPE STATISTIC
1	-0.29787	0.013974	0.23909	-.0033333

Table 3.4

used to test for autocorrelation (see Figure 3.9). Recall that this is the procedure outlined in Figure 3.8.

The values for the adjusted (deseasonalized, detrended) lag1 and lag2 correlations for total phosphorus are 0.10560 and 0.16765, respectively (see Table 3.5). Both of these are well within the significance limit (0.05 level), indicating there was no serial correlation within the data set based on the stated rules-of-thumb. This means that tests of significance using the seasonal Kendall Tau test statistic will be run without a correction for serial correlation.

The presence of serial correlation in data (or residuals) results in violation of the assumption of independence required for most statistical trend detection tests. Thus, if autocorrelation is found, the test must be adjusted or the autocorrelation eliminated. One possibility for elimination is to aggregate data, or reduce the frequency of sampling, from monthly to bimonthly or quarterly. In most cases this will eliminate serial correlation. However, the corresponding reduction in sample number means that data must be collected over a longer time period in order to account for the loss in statistical power.

The test used in the macro KENS has a correction for the covariance caused by serial correlation (calculated according to Hirsch and Slack 1984). The macro KENS is designed to calculate this correction, and report how it influences the significance of the seasonal Kendall Tau test statistic in terms of a p-value. Thus, if the data (or residuals) being analyzed exhibit serial correlation, the p-value of interest is that reported with serial correlation; if there is no serial correlation in the data (or residuals) choose the p-value without the correction.

Table 3.4 presents the output from the KENS macro, and as was noted above, the P-value of interest for total phosphorus is the one without correction for serial correlation. The particular values reported for this case study are discussed in the next section.

3.8. Trend Detection in Total Phosphorus

The method of trend detection used in the macro KENS found in Appendix B is a variation of the Mann-Kendall test discussed earlier in Section 2.4.4. The program performs a seasonal variation of the test, with an optional correction for serial correlation. The test can be used when the following conditions hold:

(1) the data set contains over 40 observations. (If the data set does not contain over 40 observations, a simple Mann-Kendall test should be run) (see Gilbert 1987, for the calculations).

(2) the data set exhibits seasonality.

FALLS TEST DATA
CORRELOGRAM WITH UPPER AND LOWER 95% CONFIDENCE LIMITS
19:51 Thursday, September 20, 199

Plot of VAR*X. Symbol used is '*'.
Plot of UP*X. Symbol used is 'U'.
Plot of LOW*X. Symbol used is 'L'.

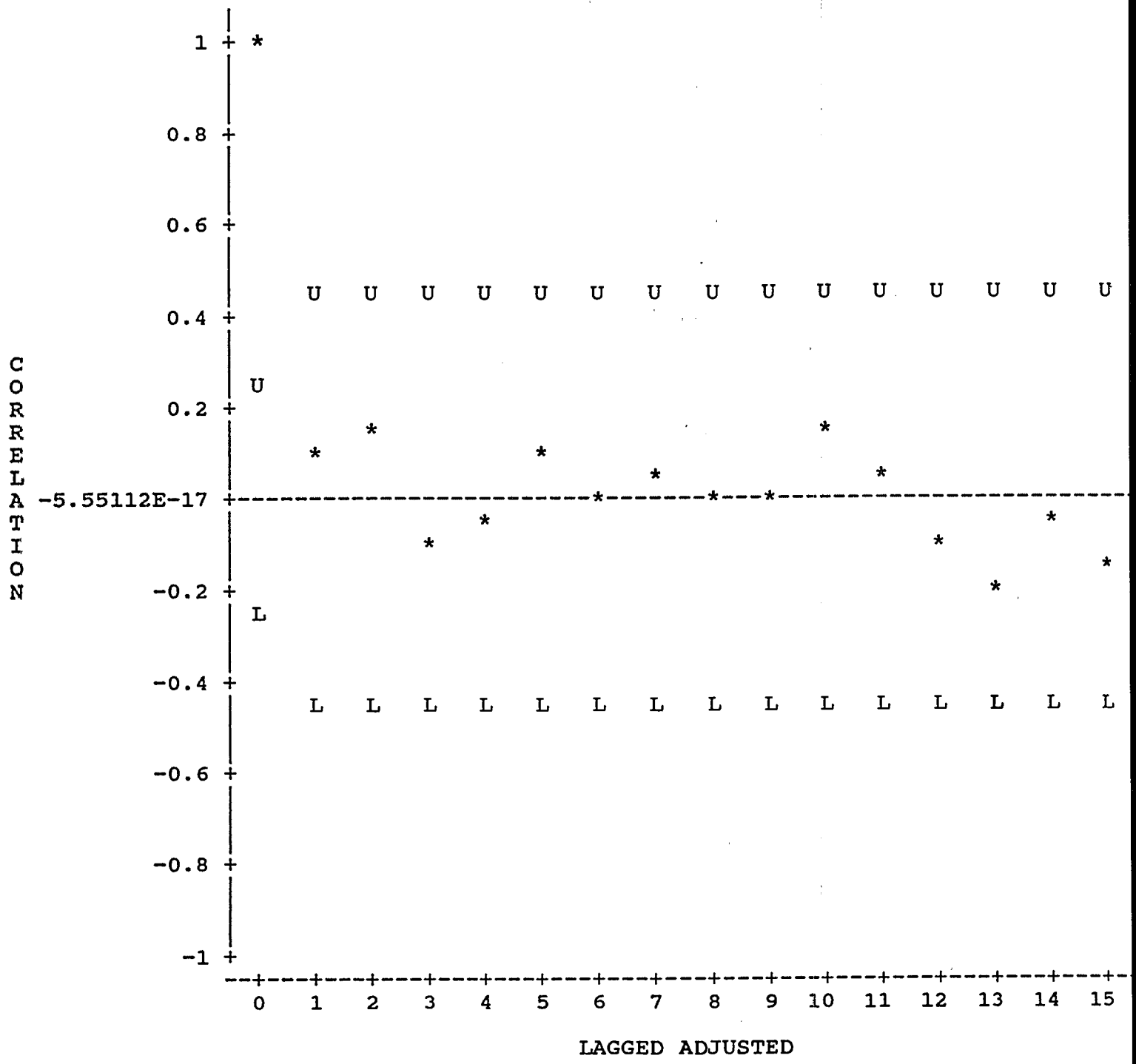


Figure 3.9

FALLS TEST DATA
PRINT OF DATA USED IN CORRELOGRAM
19:51 Thursday, September 20, 1990

OBS	LAGGED ADJUSTED	CORRELATION	STE	UPPER LIMIT	LOWER LIMIT
1	0	1.00000	0.12910	0.25820	-0.25820
2	1	0.10560	0.22361	0.44721	-0.44721
3	2	0.16765	0.22444	0.44887	-0.44887
4	3	-0.11928	0.22651	0.45303	-0.45303
5	4	-0.07445	0.22756	0.45512	-0.45512
6	5	0.09431	0.22796	0.45593	-0.45593
7	6	-0.01524	0.22861	0.45723	-0.45723
8	7	0.02570	0.22863	0.45726	-0.45726
9	8	-0.00180	0.22868	0.45736	-0.45736
10	9	-0.00711	0.22868	0.45736	-0.45736
11	10	0.16451	0.22868	0.45737	-0.45737
12	11	0.02589	0.23065	0.46129	-0.46129
13	12	-0.10352	0.23069	0.46139	-0.46139
14	13	-0.20313	0.23147	0.46294	-0.46294
15	14	-0.02947	0.23442	0.46884	-0.46884
16	15	-0.16716	0.23448	0.46896	-0.46896

Table 3.5

The Seasonal Kendall test is nonparametric, so it allows for the presence of missing values and does not require a normal distribution. The total phosphorus data set contained over 40 observations, did not appear to be normally distributed, exhibited seasonality, and the residuals did not indicate lag1 autocorrelation. Thus, we decided to use the information from the macro KENS without correction for serial correlation.

As mentioned in Section 3.7, results from the Seasonal Kendall Test on total phosphorus calculated using the macro KENS are shown in Table 3.4. The formulas for deriving these statistics can be found in Gilbert (1987), Hirsch et al. (1982), or Hirsch and Slack (1984). In this case study we chose a significance (alpha) level of 0.05.

The negative value of -0.29787 for the Tau test statistic indicates there is a negative trend in total phosphorus. As stated in Hirsch et al. (1982), the distribution of this statistic should be normal if the null hypothesis is true. The P-value of .013974 indicates that the trend is significant at an alpha (significance) level of 0.05. This value represents the probability of Z values for the test statistic at least as extreme as the one actually calculated from the observed values, if the null hypothesis of no trend was true.

The Z statistic provides a measure indicating the position of the Tau test statistic on a normal probability distribution table. It is based on the null hypothesis of no trend in the data, which would give the statistic a normal probability distribution. Hence, large (absolute) values of Z, and consequently small values of P, lead one to reject the null hypothesis of no trend. It should be noted here that the output from the macro reports only the P-value associated with the calculated Z statistic.

The KENS macro program also calculates the seasonal Kendall-Sen slope estimator for the data set. This value is the seasonal equivalent to Sen's nonparametric estimate of slope. It is the median of all possible slopes generated between all possible pairs of data points. The value for slope in this case is -0.0033333 units/year.

Thus, the conclusion for total phosphorus in Falls Lake is one of a slight (slope= 0.0033333 mg/l-year) decreasing trend. This trend is significant at the 0.05 level, and is distinct from the seasonal cycle in the data.

3.9. Total Nitrogen

The data set for total nitrogen from Falls Lake was analyzed in the same manner as was the total phosphorus data set (described in Sections 3.1 through 3.8). However, the results for total nitrogen were slightly different.

3.9.1. Normality

Table 3.6 contains the SAS PROC univariate tables from the macro Basics, and Figure 3.10 presents the corresponding plots. The stem and leaf diagram, box and whiskers plot, and normal probability graph all indicate that the data do not follow a normal distribution. This is reinforced by a W statistic for normality of 0.90432, with an accompanying probability of 0.0002 (see Table 3.6). This indicates a .02 percent probability of finding a W statistic as small or smaller than the one observed if the null hypothesis (a normal distribution) is true. Finally, the histogram constructed in Figure 3.11 confirms the lack of normality in the total nitrogen data.

3.9.2. Seasonality

The monthly box plots illustrated in Figure 3.12 show a strong pattern of seasonality. The notches do not overlap for several of the boxes, and the boxes themselves show a cyclical pattern. The lack of overlap between the notches indicates a statistically significant difference (approximate 0.05 level) between some pairs of median monthly values for total nitrogen.

The correlogram constructed in Figure 3.13 shows significant (0.05 level) correlation at several of the lag values. A cyclic pattern is evident, with the most significant negative correlation occurring at lag6 and significant positive correlation at lag12. The bivariate time series graph in Figure 3.14 also shows strong seasonal cycles in the data.

3.9.3. Independence

After deseasonalizing and detrending the data as explained in Section 3.6, a correlogram was constructed using the values from the adjusted data set (Figure 3.15). For this new data set, there were no significant (0.05 level) values for autocorrelation. This means that the tests of significance on the trend test statistic are valid without a correction for serial correlation.

3.9.4. Trends in Total Nitrogen

Table 3.7 contains the output from the Seasonal Kendall test for trend in total nitrogen. A synopsis of the statistical information calculated for the total nitrogen data set showed that the data set:

(1) contained missing values
(2) did not follow a normal distribution

UNIVARIATE PROCEDURE

Variable=TN

Moments

N	53	Sum Wgts	53
Mean	0.576604	Sum	30.56
Std Dev	0.200758	Variance	0.040304
Skewness	0.340677	Kurtosis	-1.1916
USS	19.7168	CSS	2.095789
CV	34.81726	Std Mean	0.027576
T:Mean=0	20.90949	Prob>|T|	0.0001
Sgn Rank	715.5	Prob>|S|	0.0001
Num ^= 0	53		
W:Normal	0.90432	Prob<W	0.0002

Quantiles(Def=5)

100% Max	0.94	99%	0.94
75% Q3	0.74	95%	0.93
50% Med	0.51	90%	0.87
25% Q1	0.41	10%	0.36
0% Min	0.21	5%	0.31
		1%	0.21
Range	0.73		
Q3-Q1	0.33		
Mode	0.41		

Extremes

Lowest	Obs	Highest	Obs
0.21(	55)	0.87(	38)
0.31(	41)	0.88(	48)
0.31(	7)	0.93(	26)
0.33(	29)	0.94(	35)
0.36(	6)	0.94(	47)

Missing Value	
Count	7
% Count/Nobs	11.67

Table 3.6

UNIVARIATE PROCEDURE

ariable=TN

```
Stem Leaf                    #        Boxplot
   9 344                     3           |
   8 555778                  6           |
   8 01                      2           |
   7 66                      2           |
   7 112334                  6        +-----+
   6 5                       1        |     |
   6 123                     3        |     |
   5 58                      2        |  +  |
   5 01112                   5        *-----*
   4 5                       1        |     |
   4 0111111111111124       16        +-----+
   3 66                      2           |
   3 113                     3           |
   2                                     |
   2 1                       1           |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-1
```

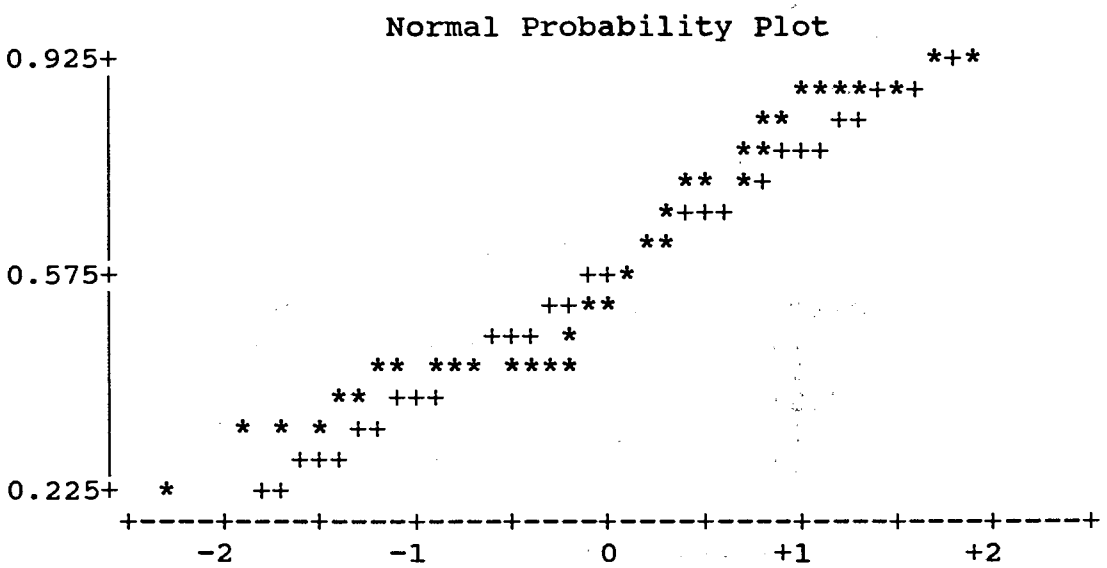


Figure 3.10

FREQUENCY OF TN

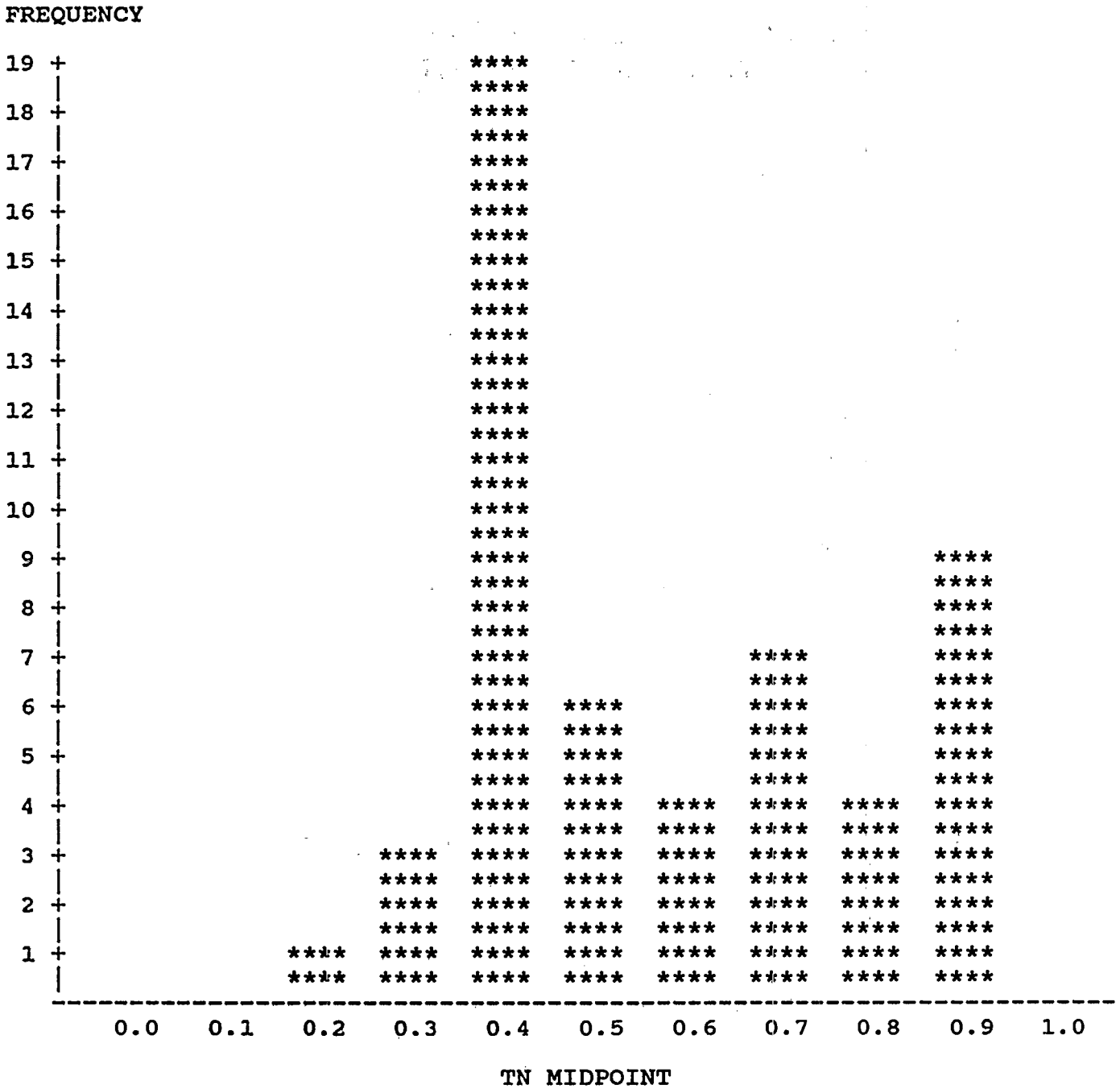


Figure 3.11

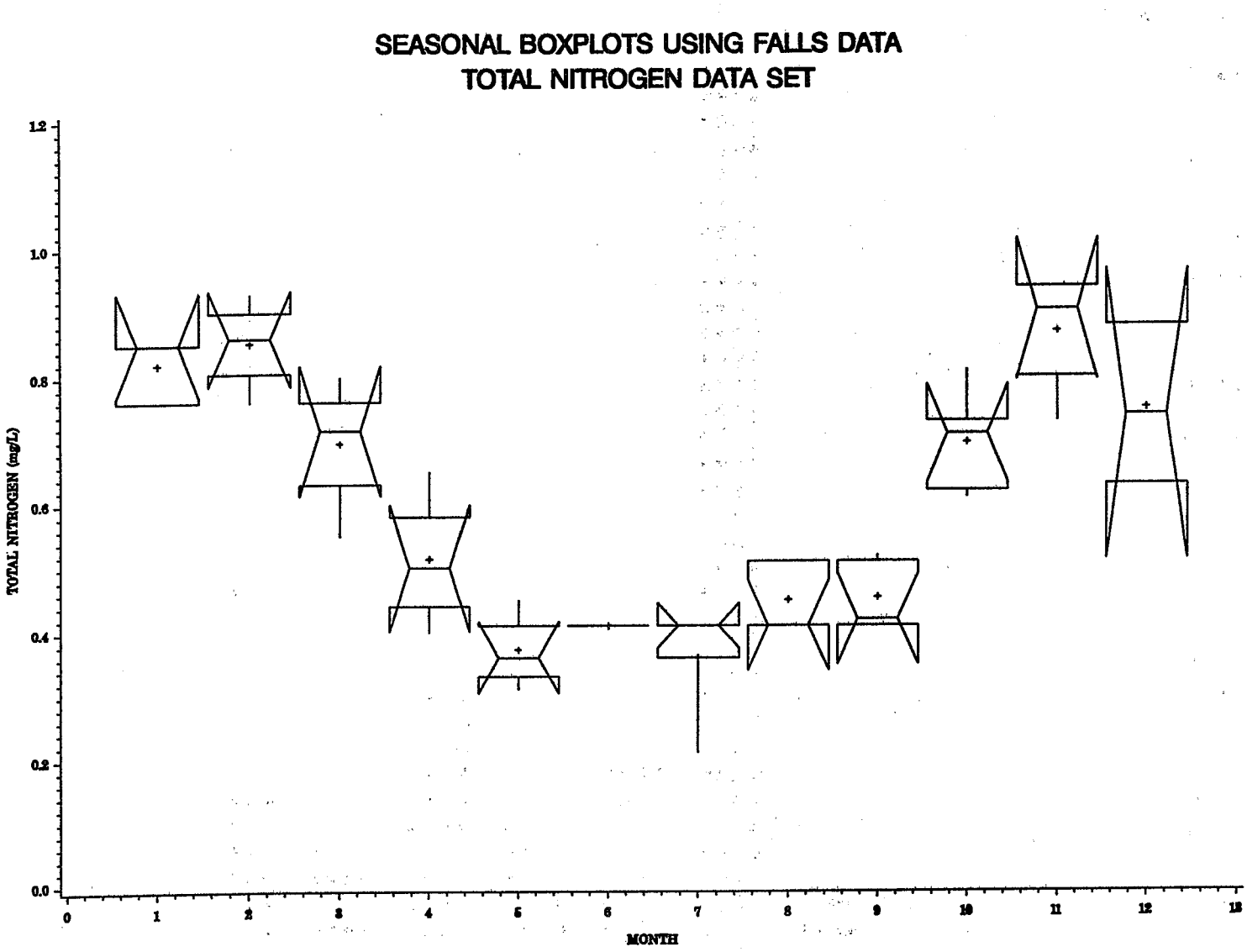

SEASONAL BOXPLOTS USING FALLS DATA
TOTAL NITROGEN DATA SET
TOTAL NITROGEN (mg/L)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
MONTH

Figure 3.12

FALLS TEST DATA
CORRELOGRAM WITH UPPER AND LOWER 95% CONFIDENCE LIMITS
20:** Thursday, September 20, 1990

Plot of VAR*X. Symbol used is '*'.
Plot of UP*X. Symbol used is 'U'.
Plot of LOW*X. Symbol used is 'L'.

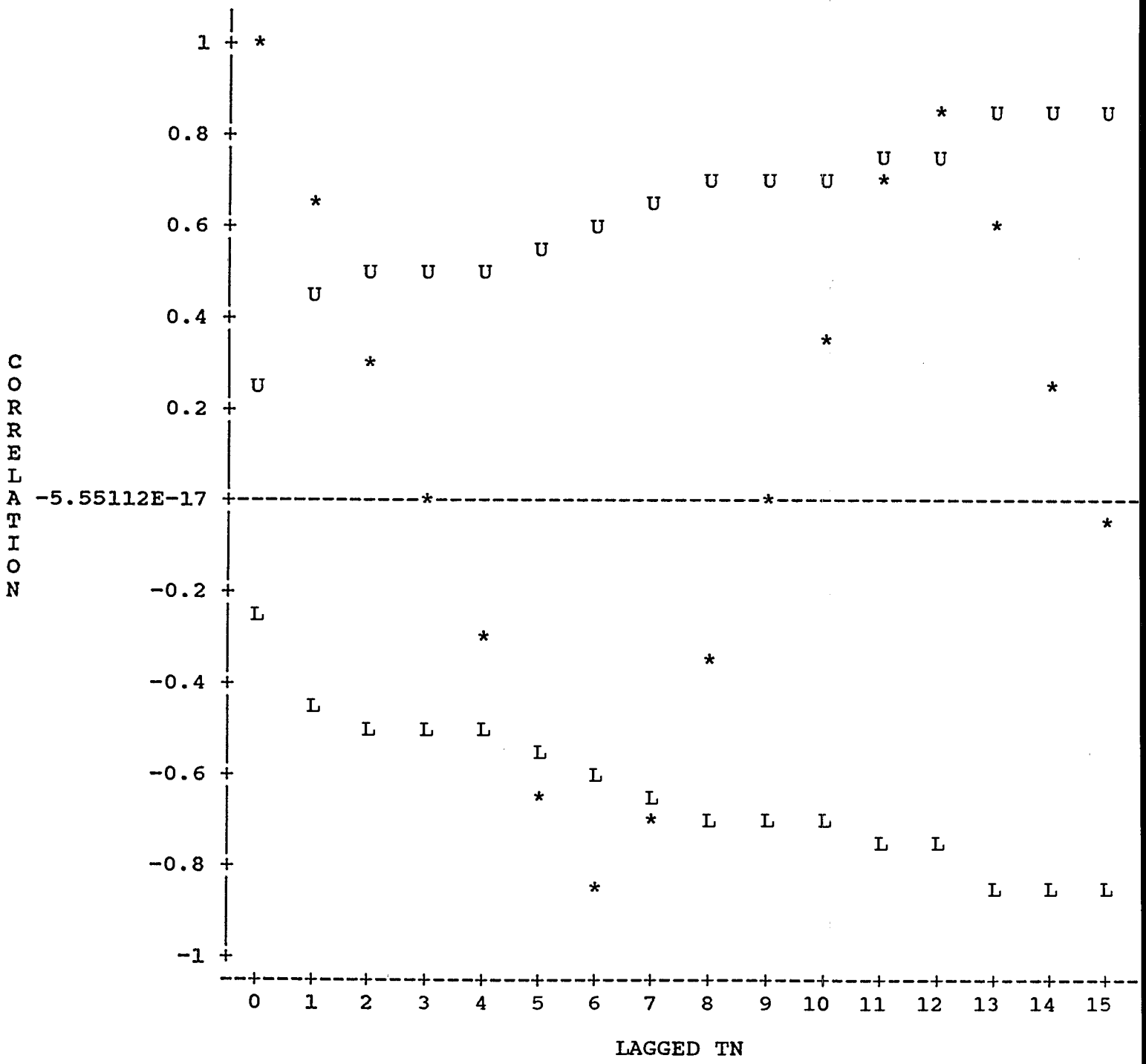


Figure 3.13

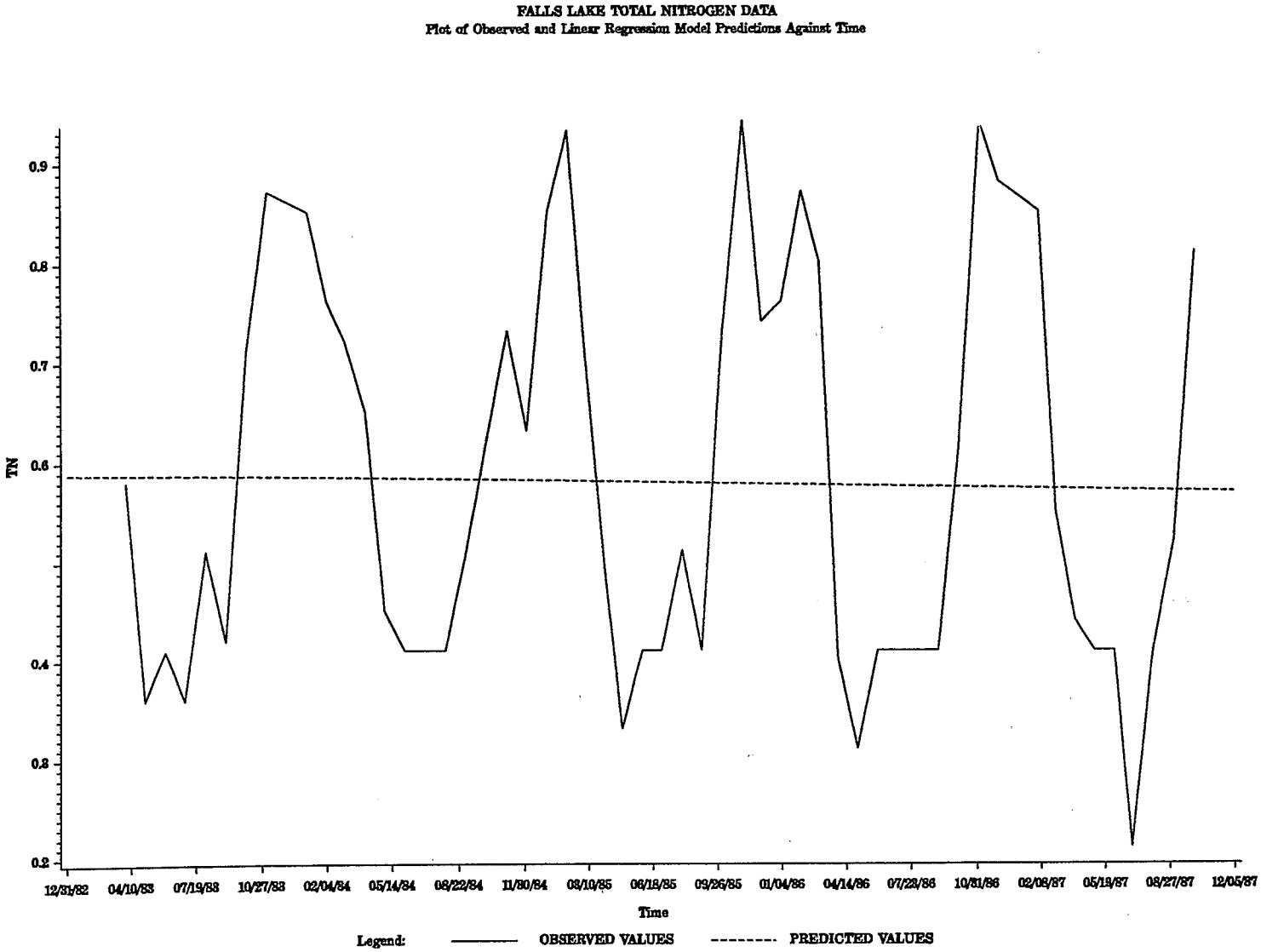

FALLS LAKE TOTAL NITROGEN DATA
Plot of Observed and Linear Regression Model Predictions Against Time
TN
0.9
0.8
0.7
0.6
0.4
0.2
12/31/82
04/10/83
07/19/83
10/27/83
02/04/84
05/14/84
08/22/84
11/30/84
03/10/85
06/18/85
09/26/85
01/04/86
04/14/86
07/23/86
10/31/86
02/08/87
05/19/87
08/27/87
12/05/87
Time
Legend:
OBSERVED VALUES
PREDICTED VALUES

Figure 3.14

FALLS TEST DATA
CORRELOGRAM WITH UPPER AND LOWER 95% CONFIDENCE LIMITS
20:** Thursday, September 20, 1990

Plot of VAR*X. Symbol used is '*'.
Plot of UP*X. Symbol used is 'U'.
Plot of LOW*X. Symbol used is 'L'.

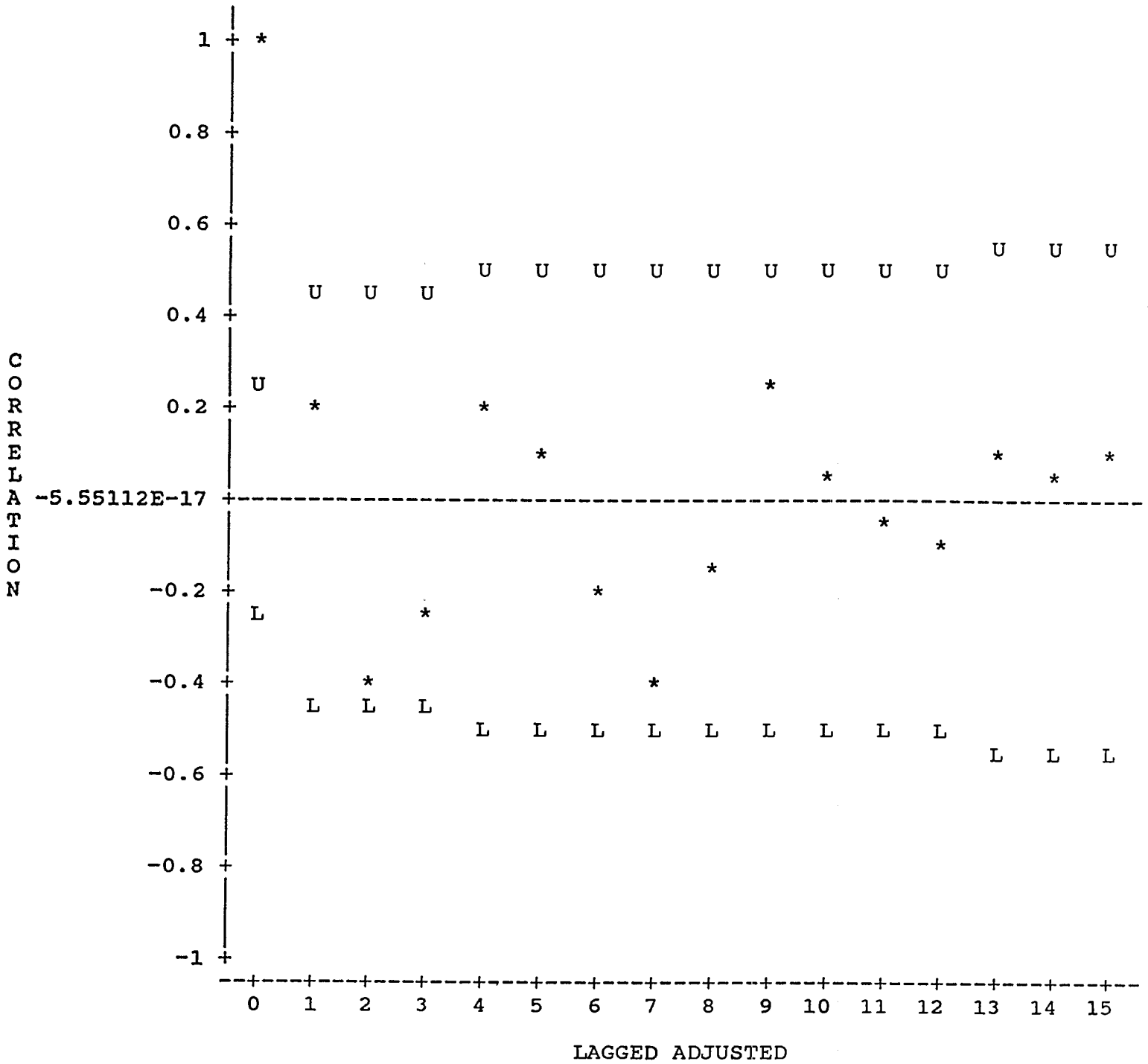


Figure 3.15

OBS	TAU STATISTIC	P-VALUE WITHOUT SERIAL CORRELATION	P-VALUE WITH SERIAL CORRELATION	SLOPE STATISTIC
1	-0.042553	0.79246	0.69530	0

Table 3.7

(3) contained seasonal cycles, and
(4) did not show serial correlation.

Therefore, the P-value without serial correlation for the Tau statistic of -0.042553 is the chosen indicator of significance in the trend. The reported P-value for the Tau statistic in this case was 0.69530, which is quite unlikely to be associated with a trend. In effect, this P-value means that the probability of getting a Z value as extreme or more extreme as that observed is .695, given that the null hypothesis of no trend is true.

3.10. Conclusions from the Phosphorus and Nitrogen Examples

While the total phosphorus and total nitrogen data sets did show similarities in distribution, seasonality, and independence, they differed in trend. Total phosphorus displayed a slight (but statistically significant at the 0.05 level) downward trend over the years studied. Total nitrogen did not show any statistically significant trend.

3.11. Regional and Statewide Lake Analysis

3.11.1. Introduction

Tests of statistical significance (hypothesis tests) are usually run so that they are to be interpreted **individually**. This means, for example, that "the trend in pH in Lake A is significantly different from zero at the 0.05 level" is a permissible statement. However, the statements should not be made collectively or simultaneously without appropriate adjustments. That is, we cannot say that "Lakes A, B, and C all have upward trends in pH that are simultaneously significant at the 0.05 level," unless the individual significance level is adjusted downward (e.g., 0.05/3), or the test is explicitly designed for multiple comparisons.

An alternative to either individual or simultaneous statements of statistical significance are "collective" statements. These statements may be expressed as "the trend in pH in the sampled population of lakes is significantly different from zero at the 0.05 level." Collective statements of statistical inference may be made using meta-analysis (Hedges and Olkin 1985), which is, literally, the statistical analysis of statistics. In this case, we can perform a meta-analysis on the seasonal Kendall's Tau statistics or the seasonal Kendall slope estimates for all of the lakes in the sample to draw collective conclusions concerning the sample.

3.11.2. Tests of Significance

Once the trend analysis presented above is run for each lake of interest, meta-analysis can be applied to these results to make a collective statement concerning regional trends. The lake statistics used in this example of meta-analysis are hypothetical; they were created to illustrate the relatively simple calculations necessary to make regional inferences concerning trends in lake water quality. The statistics represent estimated p-values that could result from the same seasonal Kendall trend detection test illustrated above.

This example uses the trend detection results from 10 hypothetical lakes to make a collective statement about trend for all of the lakes. This particular example of meta-analysis uses a method of adding Z scores (standard normal deviates) to combine probabilities; several other statistical methods may also be used as shown in Rosenthal (1984) and in Hedges and Olkin. Information on the hypothetical lakes, along with the formula used for the meta-analysis, is presented in Table 3.8.

The results of the meta-analysis in Table 3.8 indicate that even though three of the sample lakes do not show statistically significant (0.05 level) trends, there is collectively a statistically significant trend for the lakes as a whole. The highly significant Z (Z= 6.49) for the meta-analysis is due to the influence of all of the trends being in the same direction (i.e., all have a positive Z score), most of which are statistically significant (at alpha = 0.05).

Table 3.8 Lake Information And Meta-Analytic Formula

Lake	Number of Observations	p-value	Z-score
1	60	+ 0.013974	+2.45
2	120	+ 0.002460	+3.02
3	100	+ 0.012367	+2.50
4	72	+ 0.106785	+1.61
5	84	+ 0.965324	+0.04
6	60	+ 0.014672	+2.43
7	72	+ 0.003671	+2.90
8	120	+ 0.005968	+2.75
9	60	+ 0.026861	+2.21
10	84	+ 0.546129	+0.60
sum			+20.51
mean			+2.05
median			+2.44

The meta-analytic method of adding Z's (standard normal deviates) uses the following simple formula to calculate the test statistic:

$$Z = \frac{Z}{\sqrt{N}} = \frac{20.51}{\sqrt{10}} = 6.49$$

The resultant Z-statistic (6.49) is compared to a table of standard normal deviates to assess its significance.

References

Berryman, D., B. Bobee, D. Cluis, and J. Haemmerli. 1988. Nonparametric tests for trend detection in water quality time series. *Water Resources Bulletin.* 24:545-556.

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

Hedges, L.V., and I. Olkin. 1985. *Statistical Methods for Meta-Analysis.* Academic Press. Orlando, FL.

Hirsch, R.M., J.R. Slack, and R.A. Smith. 1982. Techniques of trend analysis for monthly water quality data. *Water Resources Research.* 18:107-121.

Hirsch, R.M., and J.R. Slack. 1984. A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research.* 20:727-732.

Lettenmaier, D. 1976. Detection of trends in water quality data from records with dependent observations. *Water Resources Research.* 12:1037-1046.

McGill, R., J.W. Tukey, and W.A. Larsen. 1978. Variations of box plots. *Am. Stat.* 32:12-16.

Montgomery, R. H., and J.C. Loftis. 1987. Applicability of the t-test for detecting trends in water quality variables. *Water Resources Bulletin.* 23:653-662.

Pankratz, A. 1983. *Forecasting with Univariate Box-Jenkins Models.* Wiley. NY.

Reckhow, K.H., and S.C. Chapra. 1983. *Engineering Approaches for Lake Management - Volume I: Data Analysis and Empirical Modeling.* Butterworths. Boston, MA.

Rosenthal, R. 1984. *Meta-Analysis Procedures for Social Research.* Sage Pub. Beverly Hills, CA.

SAS Institute, Inc. 1989. *SAS Users Guides for Personal Computers.* Version 6. SAS Institute. Cary, NC.

Wonnacott, T.H., and R.J. Wonnacott. *1977. Introductory Statistics.* 3rd. ed. Wiley. NY.

Appendix A: Basic Descriptive Statistics

Measures of Central Tendency

Probably the single most useful statistic summarizing a data set is an indication of the center of the sample. "Center" suggests the vague notion of the middle of a cluster of data points of perhaps the region of greatest concentration of data. Since samples of data exhibit a variety of distributions when plotted as histograms, it is not possible to unambiguously define the center, and as a result there are several statistical estimators that serve as candidates for determining central tendency or location. Each candidate, as noted below, may be considered to have its own advantages and disadvantages for the task at hand.

Mean (arithmetic)

The arithmetic mean, or simply, the mean, is the most frequently used of the central tendency estimators. It is so commonly used that the scientist often loses sight of the true reason for calculating descriptive statistics. The result is that the mean is sometimes calculated as the central tendency statistic in situations where another estimator would be better.

The arithmetic mean (x) is the sum of the observations (x_i) divided by the number of observations (n):

$$x = \frac{\sum x_i}{n} \qquad \text{(A1)}$$

Each observation contributes its magnitude to the sum of the observations and hence to the mean. For symmetric distributions (like the normal or Gaussian distribution), the mean calculated from a sample of data (the sample mean) often comes quite close to the center, or peak, of the histogram for that sample. However, limnological data are often not symmetrically distributed. The extremely high or extremely low observations characteristic of skewed data distributions "pull" the mean in the direction of the skew; this means that a few extremely high observations can pull the mean away from the bulk of the observations and toward the few high data points. In those situations, a resistant estimator, like the median or the mode, might be preferred.

Median

When a set of data is ordered from lowest to highest value, the median is identified as the middle value. The median is therefore known as an "order statistic"

since it is based on an ordering or ranking of observations. When the total number of observations is an even number, leading to two middle values, the median is then the average of the two middle values.

The "order" of the median observation is:

$$\text{Median Observation} = (n + 1)/2 \qquad \text{(A2)}$$

Since the effect on the median of all but the middle-ranking observations is simply to hold a place in the ranking, outlying observations do not pull the median toward the extremes. The median is **resistant** to the influence of any single observation, and thus it is a good statistic to use when the histogram is skewed or unusually shaped.

Trimmed Mean

The trimmed mean is the mean value from a subsample of the original sample. The subsample is formed by symmetrically trimming a small percentage of the data points from either end of the ordered observations. For example, a 10% trimmed mean is calculated from the subsample remaining after the highest and lowest 10% of the observations are removed from the data set. At the extreme, the median is the trimmed mean with all but the middle observation removed.

The trimmed mean is a good (efficient) choice for central tendency when censoring occurs or when a few outlying observations are found in the data. Here, censoring refers to data points reported as "below detection limits." In that case, if 15% of the data points are below detection limits, then a 15% trimmed mean estimator (involving 15% trimming from each end) should have lower bias than the arithmetic mean estimator based on all uncensored observations.

Mode

The mode is the value in the sample that is most frequently observed. For water quality concentration data on a continuous scale of milligrams per liter, it is possible that no value is repeated more than once. In that case, the mode may not be a useful estimator. Alternatively, if a histogram is used to represent a data set, the mode is defined as that range of values associated with the tallest bar on the histogram. The mode is considered a good estimator for central tendency because the most frequently observed value is usually near the center of the distribution. An

examination of a histogram for the sample will indicate whether the mode actually does correspond with the center.

Geometric Mean

The geometric mean is the antilog of the mean of logarithmically-transformed observations. Therefore, it is a reasonable measure of central tendency for a set of data that exhibit a lognormal distribution. The lognormal data distribution is skewed in the original units of measurement, but it is normal (Gaussian) when the original measurements are log-transformed. The lognormal distribution has been suggested by several investigators as a good probability model for concentration data for environmental contaminants. Data sets described by the lognormal have a few high values that are somewhat extreme from the bulk of the observations.

The geometric mean may be calculated in two ways:

$$Geometric\ Mean = antilog\left(\frac{\sum \log(x_i)}{n}\right) \qquad (A3)$$

or:

$$Geometric\ Mean = [\Pi x_i]^{\frac{1}{n}} \qquad (A4)$$

where $\Pi x_i = x_1 \bullet x_2 \bullet x_3 \bullet \ldots \bullet x_n$.

Measures of Dispersion

Other than central tendency, measures of dispersion or spread are the most commonly cited statistics used to summarize a data set. Dispersion in a data set refers to the variability in the observations about the center of the distribution. Good measures of dispersion will be obtained from symmetric distributions. Asymmetry, or skewness, will affect the estimate of dispersion so that it overestimates spread in the shorter tail of the data distribution (while underestimating the spread in the longer tail). A transformation (e.g., log transform) should be considered in cases of asymmetry in order to create a symmetric distribution in the transformed metric. Statistics are then calculated on the basis of the transformed metric.

Standard Deviation

The most commonly used statistic for dispersion is the standard deviation. Like the mean, the standard deviation has been used so often that it sometimes is thought to be equivalent in definition to dispersion. In fact, like the mean, the standard deviation is strongly affected by extreme values. Thus, the standard deviation for a distribution of data with a long tail to the right is inflated by the values at the extreme right. It may be preferable to apply a transformation to create a symmetric distribution before calculating the standard deviation.

For a sample, the sample variance (s^2) is:

$$s^2 = \frac{(x_i - x)^2}{n-1} \qquad \text{(A5)}$$

and the sample standard deviation (s) is the square root of the variance ($\sqrt{s^2}$).

Absolute Deviation

The standard deviation is based on squared error; squaring the deviation between a data point and the sample mean increases the influence of the largest and smallest observations on the estimate of deviation. To reduce the influence of outliers on the dispersion statistic, the absolute deviation should be considered. To calculate an absolute deviation, the mean (or median) is first estimated, and then the absolute value of the difference between the mean (median) and each data point is calculated. The mean (or median) of these absolute deviations is then calculated and is called the mean (median) absolute deviation.

Interquartile Range

Since the standard deviation is unduly influenced by extreme observations in both symmetric and asymmetric distributions of data, a resistant alternative to the standard deviation (like the median is to the mean) is needed for situations in which the data are skewed but a transformation is undesirable. Fortunately a good alternative exists - the interquartile range. Like the median, the interquartile range is based on order statistics, and thus it is unaffected by the magnitude of the extreme observations in either tail. It is calculated as the difference between the observation at the 75%ile (upper quartile) and the observation at the 25%ile (lower quartile):

Lower quartile rank order = (1/2)(1 + median rank order)

Upper quartile rank order = (1/2)(1 + n + lower quartile rank)

Interquartile range = I = lower quartile value - upper quartile value

Range

An easily determined and therefore frequently cited measure of dispersion is the range. The range is simply the maximum value minus the minimum value. Since it is clearly affected by the magnitude of the observations at either extreme, the range should not be relied upon as the sole indicator of variability. Nonetheless, it is often informative to list the range along with one of the other two dispersion statistics mentioned above.

Graphical Analyses

It is good practice in statistical analysis to begin a study with a graphical display of the data. That is, before descriptive statistics are calculated from a data set, and before the data are statistically analyzed for trend, it is wise to look at selected graphical displays of the data. Many of the graphs recommended for this task are useful in identifying important patterns in the data or in identifying the need to transform the data prior to analysis. If inferences drawn from analysis of the data are to correctly represent actual behavior, then it is important that any summary statistics used to draw inferences are representative of the data set. The graphical displays help guide the choice of any necessary manipulations of the data and selection of statistics and statistical tests.

Graphs can also be useful during the course of a statistical study. For example, bivariate plots are helpful in identifying seasonal patterns or examining the relationship between inflow and concentration. Upon completion of the statistical analysis, the scientist often wisely chooses to present some of the results in graphical form. Not infrequently, conclusions are most effectively conveyed in a graphical display.

Histograms

In even the simplest of limnological studies, data on a single characteristic need to be analyzed. Likewise, in a simple trend analysis of a single water quality variable, it is often useful to examine the distribution of the data in order to assess the central values, variability, and extremes. The limnologist could calculate the mean, standard deviation, minimum, and maximum of the sample data set;

alternatively, he could calculate other statistics representing central tendency and dispersion. Prior to calculating any statistics for the sample, however, the scientist should first look at a plot of the data. For data representing a single characteristic (such as total phosphorus concentration), the histogram is often a useful graphical display.

As an example, data on total phosphorus concentration from 1982-1988 in Jordan Lake (North Carolina) are to be analyzed for trend, but first the scientist would like to summarize the entire data set with a few statistics (perhaps to present on a graph of the time series). To obtain one picture of the sample to aid in the selection of statistics, the scientist plots the histogram shown in Figure A1. To construct the histogram, the scientist must first divide the range (highest value to lowest value) into equal-sized intervals. In Figure A1, the range is approximately 0.030mg/l to 0.200mg/l and is divided into intervals of 0.010mg/l. For each interval, 0.030 to 0.040, 0.040 to 0.050, and so on, simply count the number of data points that lie in the interval and construct vertical bars with height proportional to that number. So, for example, there are two observations in the 0.070 to 0.080 range and three observations in the 0.080 to 0.090 range. Thus, the bar for the 0.080 to 0.090 interval is 1.5 times the height of the 0.070 to 0.080 bar.

What does the histogram tell us about the sample? Basically, it provides us with a visual image of the distribution of data points in the sample. In specific terms, this means that we are able to quickly see such things as location of the "center" of the sample, amount of "dispersion," extent of "symmetry," and existence of "outliers" in the sample. In Figure A1, the center appears to be between 0.030mg/l and 0.060mg/l, depending on choice of central tendency statistic (e.g., mode, median, mean). Dispersion could perhaps be characterized by stating that over 75% of the observations lie between 0.030mg/l and 0.060mg/l, although this does not indicate the obvious skew in the data. The histogram clearly displays one outlying point which should be checked as a valid data point.

The picture created by the histogram is of considerable value in the selection of descriptive statistics. Some care should be observed in the construction of the histogram, however. With changes in interval size (e.g., changing interval width from 0.010mg/l to 0.020mg/l), the histogram may assume different shapes which might affect the inferences drawn.

As noted above, the histogram provides an impression of the extent of symmetry in the sample. Symmetry in a data set is a desirable attribute for two reasons. First, it often means that one can characterize the sample as having a distribution with a shape similar to those symmetric distributions (e.g., the normal and uniform distributions) which are commonly an assumption of statistical analysis. Stating, for example, that a sample approximates the normal distribution conveys useful information to a reader. Beyond that, symmetry implies that the common

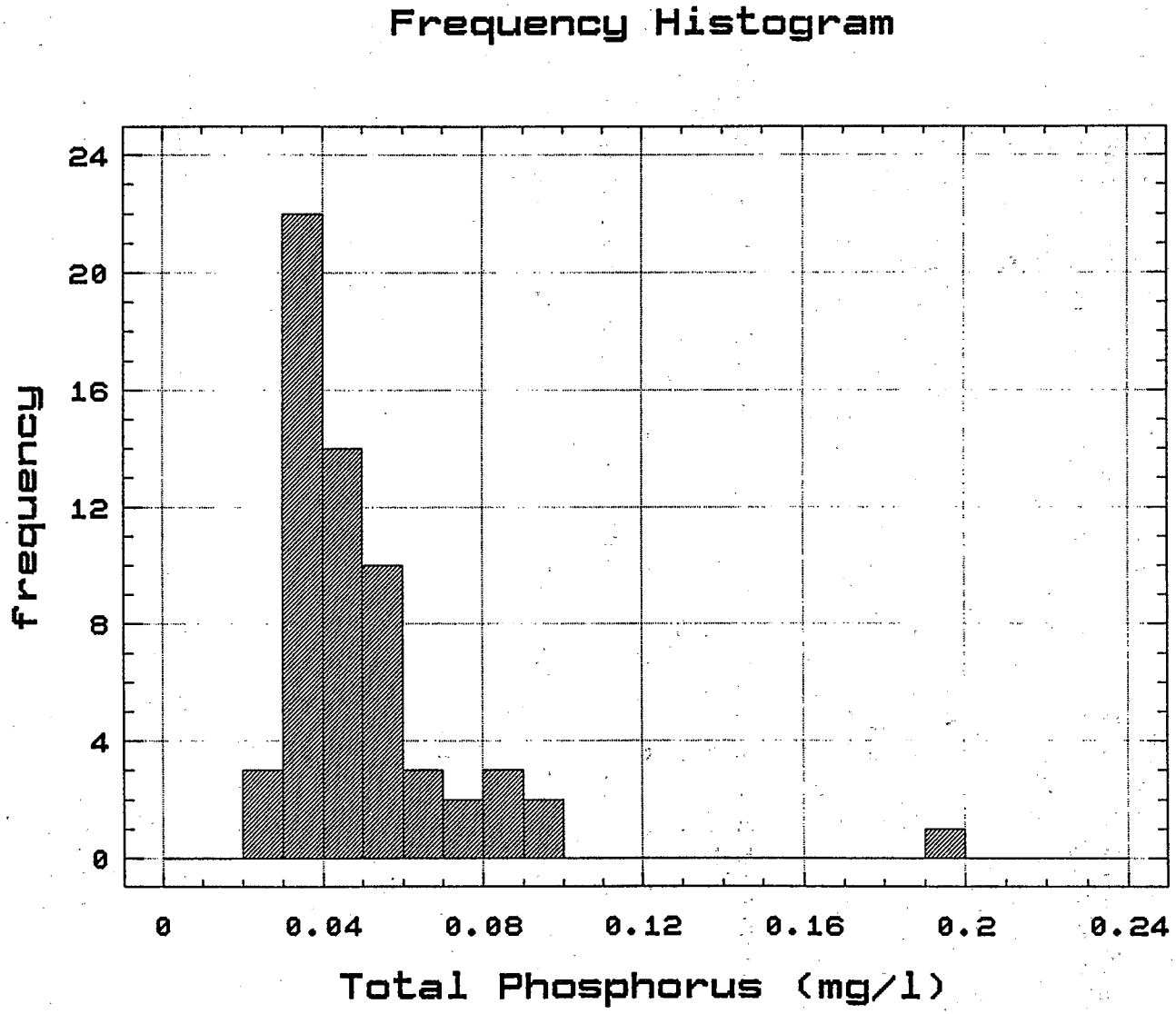

Frequency Histogram
frequency
24
20
16
12
8
4
0
0
0.04
0.08
0.12
0.16
0.2
0.24
Total Phosphorus (mg/l)

Figure A1

descriptive statistics such as the mean and standard deviation can be used to provide an adequate summary of the sample.

The foregoing discussion suggests that it might be useful to apply a transformation, if necessary, in order to create symmetry in an asymmetric data set. Fortunately, limnological data are often approximately lognormally distributed, so there is an obvious choice for transformation. The lognormal distribution is strictly positive (all observations > 0), and it is skewed right. As an example, the Jordan Lake total phosphorus data in Figure A1 approximately fit this description. To check for lognormality, the logarithmic transformation is applied to the data, and a histogram of the transformed data is plotted in Figure A2. Comparison of this histogram with a normal distribution (i.e., a bell-shaped curve) provides a rough test of lognormality; formal tests do exist (e.g., Kolmogorov-Smirnov test or chi-square test) and may be found in many statistics texts.

The difference between Figure A1 and Figure A2 illustrates how a transformation may change the shape of a histogram. While the log-transformation in Figure A2 did not achieve symmetry of the original data plotted in Figure A1, it did alter the histogram shape. To be specific, the logarithmic transformation tends to "spread out" observations that are low in value and "squeeze in" observations that are high in value. As a result of this effect, the outlier in Figure A2 is not as separate from the bulk of the observations as it is in Figure A1.

Through the study of the histograms of the sample, we should be in a better position to determine descriptive statistics for the data and to make inferences from the data.

Bivariate Plots

In time trend analysis, the basic relationship of concern is the bivariate relation between concentration of a contaminant and time. Many statistics (e.g., correlation coefficients) and many statistical methods (e.g., regression analysis) are also fundamentally concerned with relationships between pairs of variables. Without question, the single best way to examine a relationship between pairs of variables is through a bivariate graph.

For example, a bivariate graph of the time series for the Jordan Lake data discussed above is shown in Figure A3. This graph provides some indication of trend, variability, seasonality, and outliers. While this pictorial impression is clearly helpful, it must be recognized that certain patterns (e.g., seasonality) can be masked by the background variability; this is shown in some of the examples that follow in this manual. Figure A3 displays one outlier (which suggests either a transformation or nonparametric methods in subsequent statistical analyses), but no clear graphical evidence of seasonality or trend. In later sections, we see if these

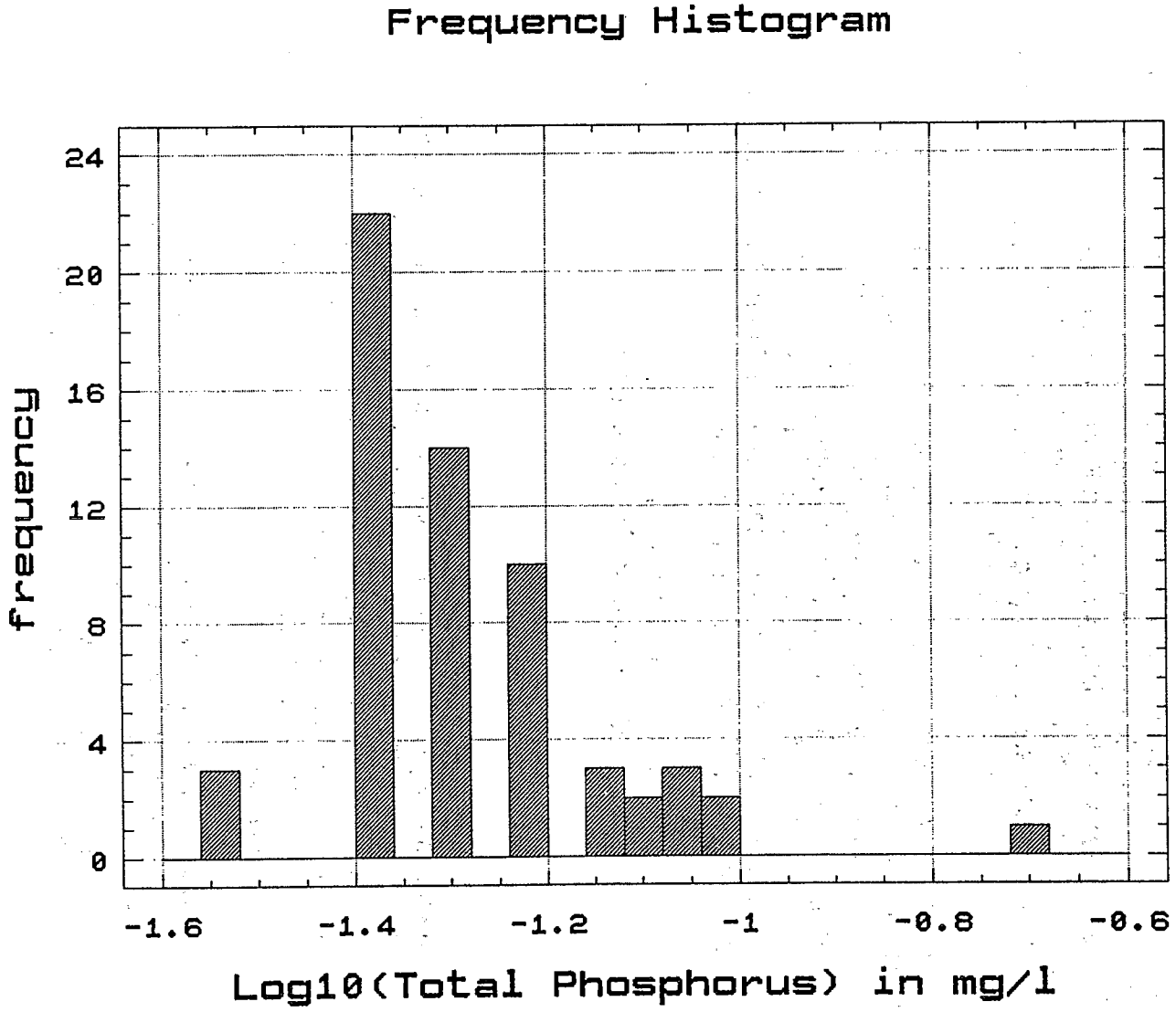

Frequency Histogram
Frequency
24
20
16
12
8
4
0
-1.6
-1.4
-1.2
-1
-0.8
-0.6
Log10(Total Phosphorus) in mg/l

Figure A2

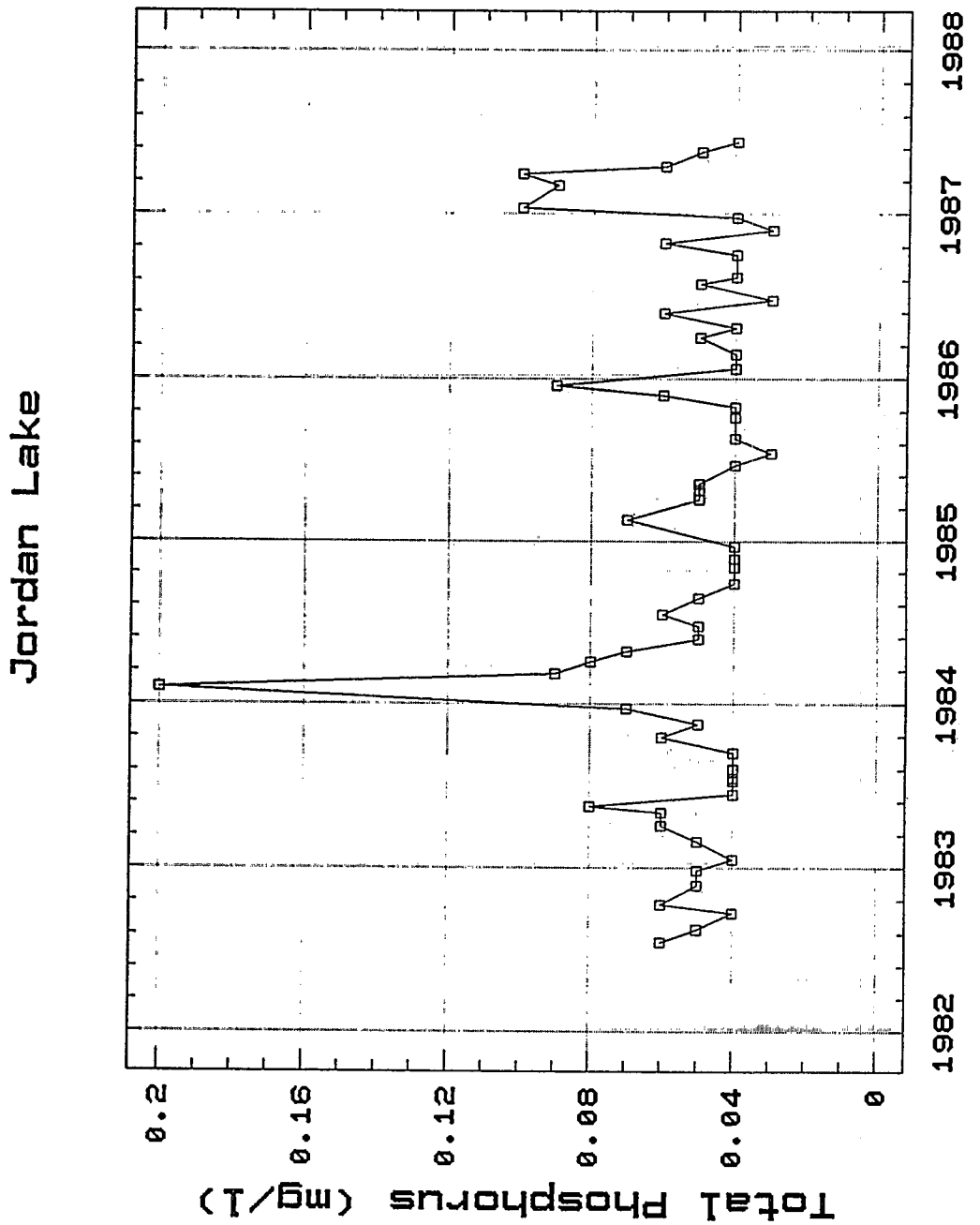

Jordan Lake
Total Phosphorus (mg/L)
0.2
0.16
0.12
0.08
0.04
0
1982
1983
1984
1985
1986
1987
1988
Date

Figure A3

conclusions are maintained when time series methods are employed to examine the data.

Box and Whisker Plots

Multiple observations, predictions, or residuals of a single water quality variable can be effectively analyzed graphically using a box and whisker plot. Box and whisker plots are based on order statistics (statistics determined based on the ordering of observations from lowest to highest value). For a data set, or for comparison of two or more data sets, the box and whisker plots display information on the sample median, dispersion, skew, relative size of the data set, and statistical significance of the median.

The SAS macros presented in this manual, SAS PROC UNIVARIATE (which produces small, un-notched plots), or the steps (from Reckhow and Chapra 1983) below may be followed to construct a box and whisker plot for a single variable:

1. Order the data from lowest to highest.

2. Plot the lowest and highest values on the graph as short horizontal lines. These represent the extreme values for each box and whisker plot.

3. Determine the upper and lower quartiles for the data set. (The quartiles are the values at the 25^{th} and 75^{th} percentiles.) These values define the positions of the upper and lower edges of the box. Using vertical lines, connect the highest value with the upper quartile and the lowest value with the lower quartile.

4. Plot the median as a dashed horizontal line within the box.

5. Select a scale so that the width of the box represents the sample size. For example, each centimeter of width could represent 25 observations.

6. determine the height of the notch (in the box at the median) based on the statistical significance of the median. Based on work by McGill et al. (1978), the height of the notch above and below the median is approximately:

$$Notch\ Limits = Median \pm (1.57 I / \sqrt{n})$$

where:

I = interquartile range = upper quartile - lower quartile

n = sample size

With this mathematical definition of the notch limits, the notch in the box provides an approximate 95% confidence interval for comparison of box medians. Therefore, when the notches for any two boxes overlap in a vertical sense, the medians are not significantly different at about the 5% level.

As an example, Figure A4 presents two box and whisker plots for a water

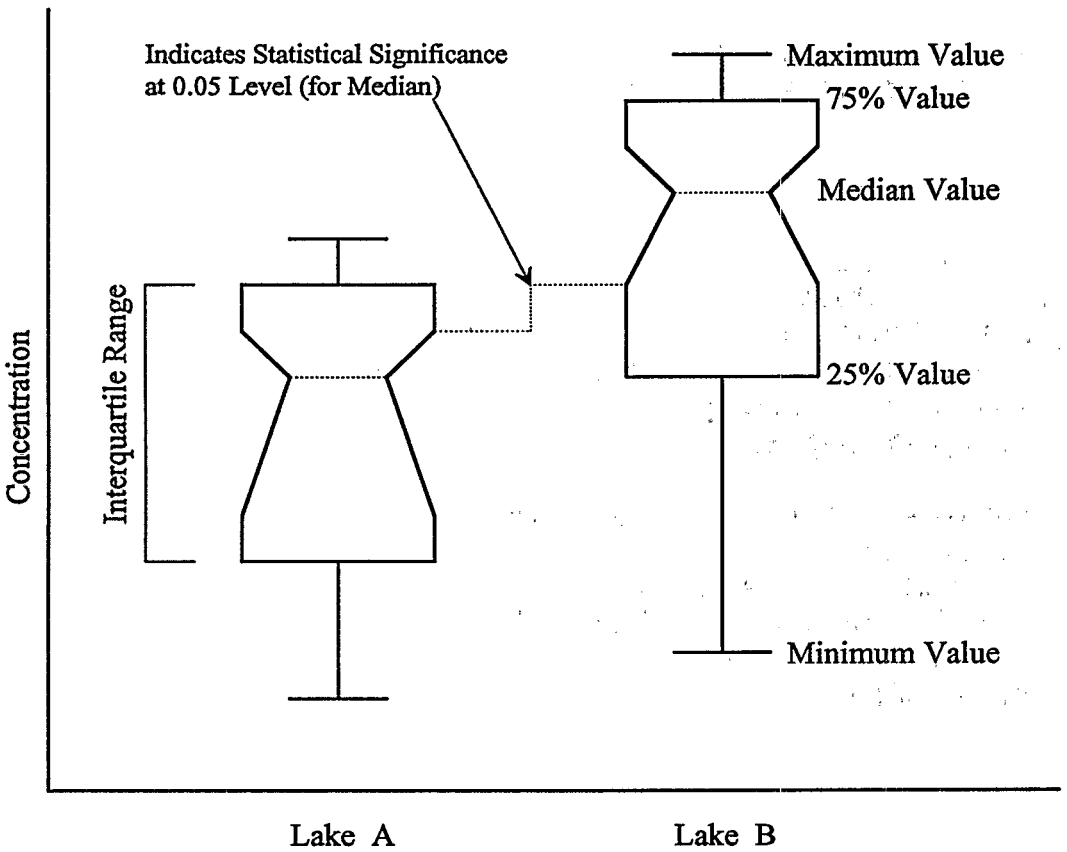


Figure A4. Box and whisker plots.

quality variable of interest measured in two lakes. The graph provides information on:

1. An estimate of the median concentration in each lake

2. A measure of dispersion in the concentration (the interquartile range)

3. The range (highest concentration - lowest concentration), and an indication of skew (based on lack of symmetry in the box shape above and below the median).

For a study of the concentrations from two or more samples (e.g., each year could represent a sample in a time series), the scientist can display:

1. A statistical test of significance in the difference between the two medians, based on vertical overlap between notches

2. A visual comparison between the two samples, based simply on observing the similarities and differences between features of two box and whisker plots.

Note that the notches in Figure A4 do not overlap in a vertical sense, indicating that the medians are significantly different at the 5% level.

Spline Smoothing

A spline smoother is a nonparametric regression estimator that may describe a locally-persistent pattern in a data set. In a graphical presentation, the spline is a smoothly-curving line that describes the general patterns in the data. It is similar to a moving average, in that it provides a local fit at each point. However, all data points in the local "window" are not weighted equally; data closest to the point of fit are assigned the highest weights.

At one extreme, a straight-line regression model is the outcome from spline smoothing; at the other extreme, the spline will zig-zag through every data point. In between the two extremes are literally an infinite number of compromises of these two fits. The SAS spline function in PROC GPLOT may be used to produce a graph of the spline smooth; with a smoothing parameter of about 0.5, the graph from GPLOT may be quite informative in displaying trend and seasonal pattern.

Appendix B: Introduction to SAS Macros

The computer functions in this manual make use of the SAS macro processing language. Macros are separate programs within the SAS system that can be used to simplify repetitive data entry tasks. The macro processing language allows the user to change input variables in a program that performs a given function, without having to rewrite the entire function. The programs can be stored separately from other SAS functions, and can be invoked at any time within a given program. Thus, the user can store programs that perform basic (or complicated) functions outside a given program, then invoke them to be used when needed.

Macros are invoked by the use of a percent sign (%), and the variables within the program are preceded by an ampersand (&). Each macro program has its own name, and can be invoked simply by writing (%) followed by the name of the macro. Each of the variables used in the macro (denoted by the "&" preceding them) is then defined in parentheses; for example, see the third page of Appendix D, where the macro "Basics" is being called. Notice that each of the variables defined in the program (found on the second page of Appendix D) must be assigned a value in order for the program to run. Comments at the beginning of each macro explain the purpose of the macro and list the variables that must be defined for that macro to run.

In order to run the macros used in this case study, the user must be familiar with the data set being analyzed. The information needed to run the macros can be found in the comment section at the beginning of each macro program. Comments are separated from program language by a row of star (*) symbols. All of the macros require a time variable, in the form of a SAS date variable. The SAS User's Guide: Basics (1989) gives information on how to enter or convert dates into a SAS form. Some of the macros require information on the number of observations and the number of seasons per year.

The SAS macros found in Appendix D were designed to give the user a framework for the detection of trends in water quality constituents. They can (and in many cases should) be modified to suit the users needs. This appendix covers four issues that users need to be familiar with, and is designed for the inexperienced macro user. The four issues discussed are as follows:

(1) Preparation of the data set.
(2) Naming of macro variables.
(3) Saving files from a macro for later use.
(4) Graphics modification.

The macros themselves are found in Appendix D, and many of the references will be to specific lines of text within the macro. There are comments throughout each of the macros that are designed to let the user know what function is being performed. These comments are always preceded by an asterisk (*), and are in most cases self-explanatory.

Data Preparation

For the macro programs to run properly, the data sets must be in a form the SAS system can read. Actual data set entry is described in detail in the SAS User's Guide: Basics (1989), and should provide the user with the information needed to transform variables (such as date) into SAS variables. Once SAS has the variables in a data set, they can be modified as needed for each macro.

The only requirements to start the macro Basics are: a data set name, a date variable (representing sequential time), the variables to be studied, a title to print on output pages, and whether or not your machine has graphics capabilities (SASGRAPH). Each of these requirements is explained in the comments found at the beginning of the macro Basics. To further assist the first-time user, we will explain how we prepared the data set for the total phosphorus case study.

The data for Falls Lake came from the North Carolina Division of Environmental Management (DEM). It was part of a larger data set containing information on many water quality variables. We chose to export the date variable, total phosphorus, total nitrogen, and observation number.

All of the variables in the data set were in a form that the SAS system could read, so they did not have to be transformed prior to entering them into a SAS data set. The date was in numeric form, with a start date of January 1, 1960 equal to 1, and all subsequent days were numerically ordered from this point (earlier dates would have negative values moving back in time from 1/1/60). This is the same start date as the SAS date variable; thus the date variable did not require transformation.

Naming of Macro Variables

Each of the macros run requires the user to list names for the variables to be used within a given run. The variables that require names are listed in parentheses directly after the macro name. These variables are also explained in the comment section at the beginning of each macro program.

The actual naming process occurs inside parentheses immediately following the calling of a macro program. For example, the naming of the variables for the macro program Basics can be seen on page 4 of Appendix D. Here, the data set to be

used is FALLS; the variable to be examined is TP; the time variable is called DATE; the title of the program is FALLS TEST DATA; and the program does have graphics capability. Each of the macro variables requested is explained in the comment section found on the second page of Appendix D.

The variable names chosen must be identical to an existing variable name in the data set being used. However, the variable names used can change each time the program is run, as illustrated in the difference between the two correlation macro programs found in Appendix D. There is no difference between these two macro programs; they simply use different variables to run the tests. The first CORR program uses the FALLS data set and the TP variable, while the second used the newly created C.ADJUST data set and ADJUSTED variable.

Saving Files

As with any SAS program, if the user creates a data set to make it permanent, the new data set must have a two level SAS name. Thus, the new data set created in the ADJUST program is designated C.ADJUST, so that it can be used for later programs. If the user is interested only in creating temporary files, then the two level designation found in many of the macros is not necessary.

The macros for this case study were run in a Batch mode, on a machine that did not have expanded memory. Therefore, many of the temporary data sets needed for subsequent macro programs had to be stored outside the SAS-PC system. We used a "C.name" to designate them. The user is referred to the SAS User's Guide: Basics (1989) for further explanation of saving data sets.

Graphics

The system used in this case study included the SASGRAPH program for graphics. Hence, the designation of "YES" for the graphics option in each of the macros. The user is referred to her/his specific system for details on device name and the other specifics for the graphics system. It should be noted that the macros allow considerable room for graphics modification, and should be adapted to suit the needs of the user.

If "NO" is designated for the graphics option, the program will still give the user a few basic graphical illustrations of the data set. The only option omitted is the bivariate time series graph, which is replaced by a printer plot of the data over time.

OBS	OBS	DATE	TP	TN	NEWDATE	MONTH	DAY	YEAR
1	1	8426	.	.	01/26/83	01	26	83
2	2	8456	.	.	02/25/83	02	25	83
3	3	8486	.	.	03/27/83	03	27	83
4	4	8516	0.07	0.58	04/26/83	04	26	83
5	5	8539	0.04	0.36	05/19/83	05	19	83
6	6	8579	0.05	0.36	06/28/83	06	28	83
7	7	8609	0.04	0.31	07/28/83	07	28	83
8	8	8643	0.04	0.51	08/31/83	08	31	83
9	19	8670	0.03	0.42	09/27/83	09	27	83
10	10	8699	0.03	0.71	10/26/83	10	26	83
11	11	8733	0.03	0.87	11/29/83	11	29	83
12	12	8748	.	.	12/14/83	12	14	83
13	13	8774	0.07	0.85	01/09/84	01	09	84
14	14	8805	0.09	0.76	02/09/84	02	09	84
15	15	8839	0.10	0.72	03/14/84	03	14	84
16	16	8873	0.07	0.65	04/17/84	04	17	84
17	17	8901	0.04	0.45	05/15/84	05	15	84
18	18	8931	0.03	0.41	06/14/84	06	14	84
19	19	8966	0.02	0.41	07/19/84	07	19	84
20	20	8994	0.05	0.41	08/16/84	08	16	84
21	21	9026	0.04	0.51	09/17/84	09	17	84
22	22	9062	0.04	0.62	10/23/84	10	23	84
23	23	9085	0.04	0.73	11/15/84	11	15	84
24	24	9118	0.03	0.63	12/18/84	12	18	84
25	25	9161	0.03	0.85	01/30/85	01	30	85
26	26	9183	0.09	0.93	02/21/85	02	21	85
27	27	9210	0.05	0.71	03/20/85	03	20	85
28	28	9238	0.03	0.50	04/17/85	04	17	85
29	29	9252	0.03	0.33	05/01/85	05	01	85
30	30	9294	0.02	0.41	06/12/85	06	12	85
31	31	9322	0.04	0.41	07/10/85	07	10	85
32	32	9356	0.03	0.51	08/13/85	08	13	85
33	33	9398	0.02	0.41	09/24/85	09	24	85
34	34	9427	0.02	0.73	10/23/85	10	23	85
35	35	9449	0.03	0.94	11/14/85	11	14	85
36	36	9482	0.05	0.74	12/17/85	12	17	85
37	37	9512	0.04	0.76	01/16/86	01	16	86
38	38	9547	0.06	0.87	02/20/86	02	20	86
39	39	9567	0.06	0.80	03/12/86	03	12	86
40	40	9614	0.04	0.40	04/28/86	04	28	86
41	41	9643	0.04	0.31	05/27/86	05	27	86
42	42	9671	0.03	0.41	06/24/86	06	24	86
43	43	9693	0.02	0.41	07/16/86	07	16	86
44	44	9714	0.03	0.41	08/06/86	08	06	86
45	45	9769	0.01	0.41	09/30/86	09	30	86
46	46	9797	0.02	0.61	10/28/86	10	28	86
47	47	9818	0.03	0.94	11/18/86	11	18	86
48	48	9845	0.02	0.88	12/15/86	12	15	86
49	49	9876	.	.	01/15/87	01	15	87
50	50	9895	0.08	0.85	02/03/87	02	03	87
51	51	9945	0.08	0.55	03/25/87	03	25	87
52	52	9958	0.04	0.44	04/07/87	04	07	87
53	53	9994	0.04	0.41	05/13/87	05	13	87
54	54	10016	0.03	0.41	06/04/87	06	04	87

Table C1. Completed Data Set

55	55	10043	0.04	0.21	07/01/87	07	01	87
56	56	10085	0.02	0.41	08/12/87	08	12	87
57	57	10119	0.03	0.52	09/15/87	09	15	87
58	58	10148	0.02	0.81	10/14/87	10	14	87
59	59	10178	.	.	11/13/87	11	13	87
60	60	10208	.	.	12/13/87	12	13	87

UNIVARIATE PROCEDURE

ariable=TP

Moments

N	53	Sum Wgts	53
Mean	0.040943	Sum	2.17
Std Dev	0.020406	Variance	0.000416
Skewness	1.222268	Kurtosis	1.013003
USS	0.1105	CSS	0.021653
CV	49.83928	Std Mean	0.002803
T:Mean=0	14.60717	Prob>|T|	0.0001
Sgn Rank	715.5	Prob>|S|	0.0001
Num ^= 0	53		
W:Normal	0.859662	Prob<W	0.0001

Quantiles(Def=5)

100% Max	0.1	99%	0.1
75% Q3	0.05	95%	0.09
50% Med	0.04	90%	0.07
25% Q1	0.03	10%	0.02
0% Min	0.01	5%	0.02
		1%	0.01
Range	0.09		
Q3-Q1	0.02		
Mode	0.03		

Extremes

Lowest	Obs	Highest	Obs
0.01(	45)	0.08(	50)
0.02(	58)	0.08(	51)
0.02(	56)	0.09(	14)
0.02(	48)	0.09(	26)
0.02(	46)	0.1(	15)

Missing Value	
Count	7
% Count/Nobs	11.67

Table C2. Univariate Moment Statistics

FALLS TEST DATA
PRINT OF DATA USED IN CORRELOGRAM
19:** Thursday, September 20, 1990

OBS	LAGGED TP	CORRELATION	STE	UPPER LIMIT	LOWER LIMIT
1	0	1.00000	0.12910	0.25820	-0.25820
2	1	0.60342	0.22361	0.44721	-0.44721
3	2	0.22298	0.24927	0.49855	-0.49855
4	3	-0.09027	0.25258	0.50515	-0.50515
5	4	-0.26161	0.25311	0.50623	-0.50623
6	5	-0.26187	0.25758	0.51516	-0.51516
7	6	-0.24185	0.26198	0.52396	-0.52396
8	7	-0.28056	0.26567	0.53135	-0.53135
9	8	-0.17439	0.27057	0.54113	-0.54113
10	9	0.01384	0.27243	0.54487	-0.54487
11	10	0.22991	0.27245	0.54489	-0.54489
12	11	0.46544	0.27566	0.55132	-0.55132
13	12	0.56642	0.28846	0.57692	-0.57692
14	13	0.39239	0.30644	0.61288	-0.61288
15	14	0.11456	0.31470	0.62940	-0.62940
16	15	-0.03949	0.31540	0.63079	-0.63079

Table C3. Correlogram Statistics

OBS	LAGGED ADJUSTED	CORRELATION	STE	UPPER LIMIT	LOWER LIMIT
1	0	1.00000	0.12910	0.25820	-0.25820
2	1	0.10560	0.22361	0.44721	-0.44721
3	2	0.16765	0.22444	0.44887	-0.44887
4	3	-0.11928	0.22651	0.45303	-0.45303
5	4	-0.07445	0.22756	0.45512	-0.45512
6	5	0.09431	0.22796	0.45593	-0.45593
7	6	-0.01524	0.22861	0.45723	-0.45723
8	7	0.02570	0.22863	0.45726	-0.45726
9	8	-0.00180	0.22868	0.45736	-0.45736
10	9	-0.00711	0.22868	0.45736	-0.45736
11	10	0.16451	0.22868	0.45737	-0.45737
12	11	0.02589	0.23065	0.46129	-0.46129
13	12	-0.10352	0.23069	0.46139	-0.46139
14	13	-0.20313	0.23147	0.46294	-0.46294
15	14	-0.02947	0.23442	0.46884	-0.46884
16	15	-0.16716	0.23448	0.46896	-0.46896

Table C4. Correlogram Statistics ("Adjusted")

UNIVARIATE PROCEDURE

Variable=TN

Moments

N	53	Sum Wgts	53
Mean	0.576604	Sum	30.56
Std Dev	0.200758	Variance	0.040304
Skewness	0.340677	Kurtosis	-1.1916
USS	19.7168	CSS	2.095789
CV	34.81726	Std Mean	0.027576
T:Mean=0	20.90949	Prob>|T|	0.0001
Sgn Rank	715.5	Prob>|S|	0.0001
Num ^= 0	53		
W:Normal	0.90432	Prob<W	0.0002

Quantiles(Def=5)

100% Max	0.94	99%	0.94
75% Q3	0.74	95%	0.93
50% Med	0.51	90%	0.87
25% Q1	0.41	10%	0.36
0% Min	0.21	5%	0.31
		1%	0.21
Range	0.73		
Q3-Q1	0.33		
Mode	0.41		

Extremes

Lowest	Obs	Highest	Obs
0.21(	55)	0.87(	38)
0.31(	41)	0.88(	48)
0.31(	7)	0.93(	26)
0.33(	29)	0.94(	35)
0.36(	6)	0.94(	47)

Missing Value	
Count	7
% Count/Nobs	11.67

Table C5. Univariate Moment Statistics for Nitrogen

Appendix D
List of SAS Programs and Files on Disk

Basics.sas - provides basic statistics and plots
Boxplt.sas - provides boxplots (can be by season or by year)
Corr.sas - provides estimates of serial correlation and plot of correlogram
Kens.sas - calculates Kendall test statistics (calls FORTRAN routines)
Adjust.sas - detrends and deseasonalizes data series (for serial correlation check)
Corradj.sas - like Corr.sas; provides serial correlations for "adjusted" series.

Kendall2,3,4.exe,for - FORTRAN programs used to calculate Kendall statistics.
"Kendall2" is structured for up to 20 years of 52-week seasons.
"Kendall3" is structured for up to 86 years of 12-month seasons, and may also be chosen for quarterly-season data.
"Kendall4" is structured for up to 1040 years of 1-season (annual) data.

Falls.dat - Falls Reservoir data set used for examples in the guidance manual.

Out.dat, Tauin.ssd, Temp.dat - misc. files created during Falls Reservoir examples.

Appendix D
SAS Program Listings
Basics.sas

The following are graphic interface lines that set up the communication with the particular graphics device used. The asterisk (*) in front of an option turns that option off; If the asterisk for the second GOPTIONS line is removed then that set of specifications determines the graphics output. In this example, the graphics are being sent to a laserjet printer (HPLJS2).

```
GOPTIONS RESET=ALL DEVICE=HPLJS2 NOROTATE GACCESS='SASGASTD>LPT1:'
   HBY=1 FBY=CENTX;
*GOPTIONS RESET=ALL DEVICE=HPLJS2 ROTATE GACCESS='SASGASTD>LPT1:'
   HBY=1 FBY=CENTX HSIZE= 5.5 VSIZE =8 NOBORDER
    HORIGIN=1.325 VORIGIN=1 NOFILL NOPOLYGONFILL
    FTEXT=CENTX HTEXT=1  ;
OPTIONS PAGESIZE=56  MPRINT  MISSING=' ';

DATA FALLS;
 INFILE 'A:FALLS1.DAT' FIRSTOBS=2 end=eof;
 INPUT OBS DATE TP TN ;
```

Select display device (e.g., printer)

Identify data set and variables

The following are program lines to modify the data set to create a NEWDATE variable, insert points (.) for missing values, and to consolidate all the day values to the same day each month (for ease in visual presentation_). These changes may not be necessary for other data sets.

```
 NEWDATE= PUT(DATE,MMDDYY8.);
  MONTH=SUBSTR(NEWDATE,1,2);
  DAY=SUBSTR(NEWDATE,4,2);
  YEAR=SUBSTR(NEWDATE,7,2);

output;

if eof then do;
tp=.;
 month='01';day='01';year='83';output;
 month='02';day='01';year='83';output;
 month='03';day='01';year='83';output;
 month='12';day='01';year='83';output;
 month='11';day='01';year='87';output;
month='12';day='01';year='87';output;
end;

proc sort data=falls;by year month day;

proc means data=falls noprint;
by year month ;
var tp;
output out=falls mean=tp;
```

Identify variable for trend analysis

```
data falls;
set falls;
length mo da yr 8.;
mo=month;
da=1;
yr=year;
date=mdy(mo,da,yr);

 RUN;

%MACRO BASICS(DATA,VAR,TIME,TITLE1,GRAPHICS);
************************************************************
  BASICS - BASIC STATISTICS, HISTOGRAM AND BIVARIATE PLOT


   INPUTS:  DATA = SAS DATA SET CONTAINING TEST INFORMATION
             VAR  = SINGLE VARIABLE OF INTEREST FOR WHICH THE
                    DIAGNOSTICS WILL BE PERFORMED
            TIME = SAS VARIABLE REPRESENTING SEQUENTIAL TIME
          TITLE1 = FIRST OUTPUT TITLE
        GRAPHICS = YES FOR FANCY GRAPHICS, NO FOR PRINTER PLOTS
             NOTE: IF YES, THEN THE USER MUST PROVIDE
             AT MINIMUM A DEVICE DRIVER AND POSSIBLY OTHER
             OTHER GRAPHICS CONTROLS (E.G. GACCESS,
             HSIZE, BORDER, ETC.) IN A GOPTIONS STATEMENT


    DATA ASSUMPTIONS:
           - ONE OBSERVATION PER TIME PERIOD
           - MISSING OBSERVATIONS ARE ALLOWED:

************************************************************;
***  PLOT DATA, AND OUTPUT STATISTICS              ***;
************************************************************;

PROC PRINT DATA=&DATA;
TITLE1 "&TITLE1";

PROC UNIVARIATE NORMAL PLOT;
  VAR &VAR;
TITLE2 "BASIC STATISTICS";

PROC CHART DATA=&DATA;
  VBAR &VAR/ MIDPOINTS=.01 TO .10 BY .005;
TITLE2 "FREQUENCY HISTOGRAM FOR &VAR";

PROC SORT DATA=&DATA; BY &TIME;

PROC REG DATA=&DATA; MODEL &VAR=&TIME;
OUTPUT OUT=OUT P=P;
TITLE1 "&TITLE1";
TITLE2 "LINEAR REGRESSION : &VAR=&TIME";
```

```
%IF &GRAPHICS=YES %THEN %DO;

SYMBOL1 C=WHITE V=NONE L=1 I=JOIN W=1;
SYMBOL2 C=WHITE V=NONE L=2 I=JOIN W=1;

DATA TEMP;
SET OUT;
KEEP CLASS PVAR &TIME;
CLASS=1;PVAR=&VAR;OUTPUT;
CLASS=2;PVAR=P;OUTPUT;

PROC FORMAT;
VALUE FX
1="OBSERVED VALUES "
2='PREDICTED VALUES'
;

PROC GPLOT DATA=TEMP;
PLOT PVAR*&TIME=CLASS /HAXIS=AXIS1 VAXIS=AXIS2 LEGEND=LEGEND1;
AXIS1 LABEL=(H=1 F=CENTX 'Date')
    WIDTH=1
    VALUE=(F=CENTX);
AXIS2 LABEL=(H=1 R=0 A=90 F=CENTX "&VAR") WIDTH=1
    VALUE=(F=CENTX);
LEGEND1 LABEL=(H=1 F=CENTX 'Legend: ')
    VALUE=(F=CENTX);
FORMAT &TIME  MMDDYY8. CLASS FX.;
TITLE1 H=1 F=CENTX "&TITLE1";
TITLE2  F=CENTX H=1
'Plot of Observed and Linear Regression Model Predictions Against Time';

%END;

%ELSE %DO;

 PROC PLOT DATA=OUT;
 PLOT &VAR*&TIME='O' P*&TIME='P'/OVERLAY;
 LABEL &TIME ='TIME';
 FORMAT &TIME  MMDDYY8.;
 TITLE2 "PLOT OF &VAR (O) AND PREDICTED VALUES(P) AGAINST TIME";
%END;
 RUN;
%MEND BASICS;
```

The following are the program variables chosen for the example run. The variables can be changed for each run of the macro, which is invoked by "%BASICS" followed by the necessary variables in parentheses.

```
%BASICS(DATA=FALLS,
      VAR=TP,
      TIME=DATE,
      TITLE1=Figure 3.4. FALLS TEST DATA,
      GRAPHICS=YES
      );
```

← Variables in the Basics Macro are defined here.

Boxplt.sas

The following graphic interface lines determine the features of the graphic output (screen display or printer output). For example, Figure 3.5 was printed (on the HP laserjet printer) using the second GOPTIONS statement; this statement was invoked by simply removing the asterisk at the beginning of the line.

```
GOPTIONS RESET=ALL DEVICE=HPLJS2 GACCESS='SASGASTD>LPT1:' HBY=1
  FBY=CENTX ;
*GOPTIONS RESET=ALL DEVICE=HPLJS2 ROTATE GACCESS='SASGASTD>LPT1:'
    HBY=1 FBY=CENTX HSIZE=8 VSIZE=5.5 NOBORDER
     HORIGIN=1.5 VORIGIN=1.5 NOFILL NOPOLYGONFILL
     FTEXT=CENTX HTEXT=1  ;
OPTIONS PAGESIZE=56  MPRINT  MISSING=' ' ;

DATA FALLS;
 INFILE 'A:FALLS1.DAT' FIRSTOBS=2 end=eof;
 INPUT OBS DATE TP TN ;
```

Identify data set and variables

The following are program lines to modify the data set to create a NEWDATE variable, insert points (.) for missing values, and to consolidate all of the day values to the same day each month (for ease in visual presentation). These manipulations may not be necessary for other data sets.

```
 NEWDATE= PUT(DATE,MMDDYY8.);
  MONTH=SUBSTR(NEWDATE,1,2);
  DAY=SUBSTR(NEWDATE,4,2);
  YEAR=SUBSTR(NEWDATE,7,2);

output;

if eof then do;
tp=.;
 month='01';day='01';year='83';output;
 month='02';day='01';year='83';output;
 month='03';day='01';year='83';output;
 month='12';day='01';year='83';output;
 month='11';day='01';year='87';output;
month='12';day='01';year='87';output;
end;

proc sort data=falls;by year month day;
proc means data=falls noprint;
by year month ;
var tp;
output out=falls mean=tp;

data falls;
set falls;
length mo da yr 8.;
mo=month;
da=1;
yr=year;
```

Identify variable for analysis

```
date=mdy(mo,da,yr);
RUN;

************************************************************
   SAS MARCOS FOR CREATING A BOXPLOT USING SAS GPLOT
    PROCEDURES.  MACROS NEEDED INCLUDE:
              1. %MACRO NOBS
              2. %MACRO ORDER
              3. %MACRO BOXVARS
              4. %MACRO BOXPLOT

AN EXAMPLE CALL USING THE MOST INTERESTING OPTIONS IS PROVIDED.

DO NOT USE A STATEMENT STYPE CALL.  IF A PARAMETER VALUE CONTAINS
PARENTHESES, COMMAS OR EQUALS SIGNS, MAKE IT THE ARGUMENT OF THE
%STR FUNCTION: PARAMETER=%STR(VALUE).

THE MACROS ARE TAKEN FROM A PAPER IN SUGI 10 PROCEEDINGS, P. 890.
AUTHOR: ANN OLMSTED, SYNTEX RESEARCH

DESCRIPTION:
 THE BASIC BOX:
 FOR EACH VALUE OF YOUR CLASS VARIABLE, WHICH MUST BE NUMBERIC,
 THE PROGRAM DRAWS A RECTANGLE WITH A LINE PASSING THROUGH IT AT THE
 PERCENTILE &MIDDLE, LOWER AND UPPER EDGES AT PERCENTILES
 &LO_EDGE AND &HI_EDGE, AND WHISKERS EXTENDING TO &LO_WHISK AND
 &HI_WISK. A PLUS SIGN MARKS THE MEAN.

 &LO_WHISK, &LO_EDGE, &MIDDLE, &HI_EDGE, AND &HI_WHISK DEFAULT
 TO MIN, P25, P50, P75, AND MAX.  IF YOU SPECIFY THEIR VALUES,
 YOU MAY CHOSE FROM: MIN, P1, P5, P10, P25, P50, MEAN, P75,
 P90, P95, P99, MAX.
 THE PERCENTILES ARE COMPUTED BY PROC UNIVARIATE USING THE DEFAULT
 METHOD.

VARIABLE BOX WIDTH:
 YOU CAN CONTROL THE WIDTH OF THE BOXES THROUGH THESE PARAMETERS:
 K -- HALFWIDTH OF WIDEST BOX. IF KA IS NOT GIVEN, THE
    PROGRAM WILL SET IT TO .9*(MIN SPACING BET. CLASS VALUES)
 FN-- ANY FUNCTION OF THE GROUP SIZE, E.G., 1, SQRT(N), N.
    BOXWIDTHS WILL BE PROPORTIONAL TO F(N). THE DEFAULT
    IS 1 (CONSTANT BOXWIDTH)
 OVERLAP -- SET TO 1 (TRUE) TO PERMIT BOXES TO OVERLAP, OR
              0 (FALSE) TO ENABLE THE PROGRAM TO OVERRIDE YOUR
              VALUE OF K IF IT WOULD RESULT IN OVERLAP. THE
              DEFALUT IS 0 (NO OVERLAP PERMITTED)

WAISTS (CONFIDENCE INTERVALS FOR THE MEDIAN).
 THE BOXES WILL BE NOTCHED TO INDICATE A CONFICENCE INTERVAL
 FOR THE MEDIAN IF THE MACRO IS CALLED WITH WAIST=1. SET
 PARAMETER F (DEPTH OF THE NOTCH AS A FRACTION OF THE HALFWIDTH)
 TO A VALUE BETWEEN 0 AND 1.

 WAISTTYP=ORDER:
```

```
SET LEVELTAR TO YOUR DESIRED MINIMUM CONFIDENCE LEVEL (E.G.,
.90 OR .95).  THE PROGRAM GOES INTO A LOOP TO SELECT INTEGERS
R,S (1 <=R <S <= N) S.T.
   P ( X(R) <=MEDIAN <=X(S) ) >= &LEVELTAR.
THEN USES THE SAMPLE ORDER STATISTICS X(R), X(S) TO DEFINE THE
NOTCH.

 WAISTTYP=TUKEY:
SET TUKCONST TO THE DESIRED MULTIPLE OF THE GAUSSIAN ASYMPTOTIC
STANDARD DEVIATION OF THE MEDIAN S_HAT=
(1/1.08)(P75-P25)/SQRT(N).
THE VALUES P50 - &TUKCONST*S_HAT, P50 + &TUKCONST*S_HAT ARE
USED TO DEFINE THE NOTCH.
*************************************************************************
;


%MACRO NOBS(DATA=_LAST_);
DATA _NULL_;
 POINT=1;
 SET &DATA POINT=POINT NOBS=NOBS;
 PUT NOBS " OBS IN DATASET &DATA " /;
 %GLOBAL NOBS;
 CALL SYMPUT('NOBS',TRIM(LEFT(NOBS)));
 STOP;
 RUN;
%MEND NOBS;

%MACRO ORDER(P=.5,LEVELTAR=.90);
PROC SORT DATA=WORK OUT=SORTDATA;
 BY &CLASS &VAR;
DATA RANKDATA N (KEEP=&CLASS RANK RENAME=(RANK=N));
 SET SORTDATA;
 BY &CLASS;
 IF(FIRST.&CLASS) THEN DO;
  RANK=0;
 END;
 RANK+1;
 OUTPUT RANKDATA;
 IF(LAST.&CLASS) THEN DO;
  OUTPUT N;
 END;

 %LET PROB=(PROBBNML(&P,N,S-1) - PROBBNML(&P,N,R-1) );
 DATA RS (KEEP=N &CLASS R S LEVEL);
  SET N;
  R=FLOOR( (N+1)*&P) - ( (N+1)*&P=FLOOR((N+1)*&P)  );
  S=R+1;
  STEPR=1;
  DO WHILE ( (&PROB <&LEVELTAR) AND ((R>1) OR (S<N)) );
    IF ( STEPR) THEN DO;
      R=R-(R>1);
    END; ELSE DO;
      S=S+(S<N);
```

```
    END;
    STEPR=1-STEPR;
  END;
  LEVEL=&PROB;
  IF ( LEVEL < &LEVELTAR) THEN DO;
   PUT &CLASS= N= R= S= LEVEL=
    /"NOTE: FAILURE TO ACHIEVE CONF. LEVEL &LEVELTAR" /;
  END;
DATA WAIST(KEEP=&CLASS N R S LEVEL RVAL SVAL);
  MERGE RANKDATA RS;
  BY &CLASS;
  RETAIN RVAL;
  IF(RANK=R) THEN DO;
   RVAL=&VAR;
  END;ELSE IF ( RANK=S) THEN DO;
   SVAL=&VAR;
   OUTPUT;
  END;
PROC PRINT DATA=WAIST LABEL;
VAR &CLASS N R S RVAL SVAL LEVEL;
 LABEL N   = "n"
     R   = "r"
     S   = "s"
     RVAL= "x(r)"
     SVAL= "x(s)";
TITLE1 "DATASET: WAIST(SOURCE: &DATA),VARIABLE: &VAR";
TITLE2 "(X(R),X(S)) FORMS A LEVEL LEVEL CONFIDENCE INTERVAL FOR";
TITLE3 "QUANTILE &P";
TITLE4 "THE TARGET CONFIDENCE LEVEL WAS &LEVELTAR";
%MEND ORDER;

%MACRO BOXVARS;
 &LO_WHISK
%IF ( &LO_EDGE NE &LO_WHISK) %THEN %DO;
  &LO_EDGE
%END;
%IF ( &MIDDLE NE &LO_EDGE) %THEN %DO;
 &MIDDLE
%END;
%IF( &HI_EDGE NE &MIDDLE) %THEN %DO;
 &HI_EDGE
%END;
%IF ( &HI_WHISK NE &HI_EDGE) %THEN %DO;
 &HI_WHISK
%END;
%MEND BOXVARS;

%LET LARGENUM=1E23;
%LET SMALLNUM=-1E23;
%MACRO BOXPLOT(DATA=_LAST_,OUT=,     /* INPUT AND OUTPUT FILES      */
        CLASS=,VAR=,                 /* CLASS AND PLOTTING VARIABLES */
                                     /* MUST BE NUMERIC             */
        K=,                          /* HALFWIDTH OF WIDEST BOX     */
        FN=1,                        /* 1/SQRT(N)/N OR PROPORTIONAL */
```

```
                                    /*  TO F(N)                 */
        WAIST=0,                    /* 0/1=NO WAISTS/WAISTS      */
        WAISTTYP=ORDER,             /* ORDER/TUKEY               */
        LEVELTAR=.90,               /* TARGET CONF. LEVEL IF WAISTTYP */
                                    /*  =ORDER                   */
        TUKCONST=1.7,               /* WAIST INT=MED +- TUKCONST*S_HAT */
                                    /* IF WAISTTYP=TUKEY         */
        F=1,                        /* NOTCH DEPTH (AS FRACTION OF  */
                                    /* HALFWIDTH FOR WAISTED PLOTS  */
        LO_WHISK=MIN,               /* PLOTTED RANGES AND PERCENTILES */
        LO_EDGE=P25,
        MIDDLE=P50,
        HI_EDGE=P75,
        HI_WHISK=MAX,
        KLUDGE=0,
        LO_MARK=P25,
        HI_MARK=P75,
        V_LO=,V_HI=,V_BY=,          /* PLOTTING RANGES FOR VERTICLE  */
                                    /*  AXIS                     */
        CONNECT=1,                  /* 0/1 -- SET TO 1 TO CONNECT   */
        OVERLAP=0,                  /* 0/1 -- SET TO 1 TO ALLOW OVERLAP */
      TITLE1=,TITLE2=,              /* TITLES                    */
        CLASSFMT=,                  /* FORMAT FOR CLASS VARIABLE    */
        VARFMT=,                    /* FORMAT FOR PLOTTED VARIABLE  */
        CLASSLAB=,                  /* CLASS VARIABLE LABEL      */
        VARLAB=,                    /* PLOTTED VARIABLE LABEL    */
        VERBOSE=1                   /* 0/1 -- SET TO 1 FOR PRINTOUTS */
 );
%IF ( &MIDDLE NE P50 ) AND (&WAIST) %THEN %DO;
 %PUT PROGRAM DOES NOT COMPUTE CONFIDENCE INTERVALS FOR &MIDDLE;
 %LET WAIST=0;
%END;
%IF( NOT &WAIST) %THEN %DO;
 %LET WAISTTYP=;
%END;
%IF( %QUOTE(&V_LO&V_HI&V_BY) NE) %THEN %DO;
 %IF( %QUOTE(&V_LO)= ) OR
   ( %QUOTE(&V_HI)= ) OR
   ( %QUOTE(&V_BY)= ) %THEN %DO;
    %PUT INCOMPLETE VAXIS SPECIFICATION IGNORED;
    %LET V_LO=; %LET V_HI=; %LET V_BY=;
 %END;
%END;

DATA WORK;
 SET &DATA(KEEP=&CLASS &VAR);
 IF(&CLASS > .Z) AND ( &VAR > .Z);
PROC SORT DATA=WORK;
 BY &CLASS &VAR;
PROC UNIVARIATE DATA=WORK NOPRINT;
 BY &CLASS;
 OUTPUT OUT=SUMMARY
  MEAN=MEAN N=N MIN=MIN P1=P1 P5=P5 P10=P10 Q1=P25 MEDIAN=P50
  Q3=P75 P90=P90 P95=P95 P99=P99 MAX=MAX;
```

```
%IF( &WAISTTYP=ORDER) %THEN %DO;
 %ORDER(LEVELTAR=&LEVELTAR)
%END;

DATA _NULL_;
 SET SUMMARY(KEEP=&CLASS N) END=LASTOBS;
 RETAIN KCOMP &LARGENUM FMAX &SMALLNUM;
 SPAN=DIF(&CLASS);
 IF( _N_ > 1) THEN DO;
  KCOMP=MIN(KCOMP,SPAN/2);
 END;
 FMAX=MAX(FMAX,&FN);
 IF(LASTOBS) THEN DO;
  CALL SYMPUT('KCOMP',TRIM(LEFT(KCOMP)));
  PUT ' MAXIMUM POSSIBLE HALFWIDTH FOUND TO BE : ' KCOMP /;
  CALL SYMPUT('FMAX',TRIM(LEFT(FMAX)));
  PUT ' FMAX: ' FMAX /;
 END;
  RUN;

%LET TRIM_HI=0;
%LET TRIM_LO=0;
DATA BOXDATA(KEEP=&CLASS X Y);
%IF(&WAISTTYP NE ORDER) %THEN %DO;
 SET SUMMARY END=LASTOBS;
%END; %ELSE %DO;
 MERGE SUMMARY WAIST END=LASTOBS;
%END;
 BY &CLASS;
 RETAIN KCOMP &KCOMP K &K HALFWID;
 IF( _N_ =1) THEN DO;
 IF( K<=0) THEN DO;
  K=.9*KCOMP;
 END;ELSE DO;
%IF ( NOT &OVERLAP) %THEN  %DO;
   K=MIN(K , 0.9*KCOMP);
%END; %ELSE %DO;
  IF(K > KCOMP) THEN DO;
   PUT ' NOTE: K=' K ' KCOMP=' KCOMP ' BOXES WILL OVERLAP' /;
  END;
%END;
  END;* K > 0 ELSE;
 END;
 IF(FIRST.&CLASS) THEN DO;
  HALFWID=K * &FN/&FMAX;
%IF ( &WAISTTYP=TUKEY) %THEN %DO;
 S_HAT=(1/1.08)*(P75-P25)/SQRT(N);
 L2=&TUKCONST*S_HAT;
 RVAL=P50-L2;
 SVAL=P50+L2;
%END;
 END;
```

```
X=&CLASS              ;Y=&LO_WHISK; OUTPUT;
X=&CLASS              ;Y=&LO_EDGE ; OUTPUT;
X=&CLASS-HALFWID;Y=&LO_EDGE;OUTPUT;
%IF (&WAIST) %THEN %DO;
 X=&CLASS-HALFWID;Y=RVAL;OUTPUT;
 X=&CLASS-(1-&F)*HALFWID;Y=&MIDDLE;OUTPUT;
 X=&CLASS-HALFWID;Y=SVAL;OUTPUT;
%END;
  X=&CLASS-HALFWID;Y=&HI_EDGE;OUTPUT;
  X=&CLASS;Y=&HI_EDGE;OUTPUT;
  X=&CLASS;Y=&HI_WHISK;OUTPUT;
  X=&CLASS;Y=&HI_EDGE;OUTPUT;
  X=&CLASS+HALFWID;Y=&HI_EDGE;OUTPUT;
%IF ( &WAIST) %THEN %DO;
  X=&CLASS + HALFWID; Y=SVAL;OUTPUT;
%END;
  X=&CLASS + (1-&WAIST*&F)*HALFWID;Y=&MIDDLE;OUTPUT;
  X=&CLASS - (1-&WAIST*&F)*HALFWID;Y=&MIDDLE;OUTPUT;
  X=&CLASS + (1-&WAIST*&F)*HALFWID;Y=&MIDDLE;OUTPUT;
%IF (&WAIST) %THEN %DO;
  X=&CLASS + HALFWID; Y=RVAL; OUTPUT;
%END;
 X=&CLASS + HALFWID; Y=&LO_EDGE; OUTPUT;
 X=&CLASS; Y=&LO_EDGE; OUTPUT;
%IF( %QUOTE(&V_LO) NE ) %THEN %DO;
 %PUT V_LO=&V_LO;
 RETAIN MAXHI MAXHI_ED &SMALLNUM MINLO MINLO_ED &LARGENUM;
 MAXHI=MAX(&HI_WHISK,MAXHI);
 MAXHI_ED=MAX(&HI_EDGE,MAXHI_ED);
 MINLO_ED=MIN(&LO_EDGE,MINLO_ED);
 MINLO=MIN(&LO_WHISK,MINLO);
 IF ( LASTOBS ) THEN DO;
  REF=&V_BY * FLOOR( (MAXHI_ED - MINLO_ED)/&V_BY);
  C_HI=REF+&V_BY *
     ( FLOOR( (MAXHI_ED-REF)/&V_BY) + 1);
  C_LO=REF-&V_BY*
     (FLOOR( (REF-MINLO_ED)/&V_BY) + 1);
  V_HI=MAX(&V_HI,C_HI);
  V_LO=MIN(&V_LO,C_LO);
  CALL SYMPUT('V_HI',TRIM(LEFT(V_HI))); PUT V_HI= ;
  CALL SYMPUT('V_LO',TRIM(LEFT(V_LO))); PUT V_LO= ;
  TRIM_HI=(MAXHI > V_HI);PUT MAXHI= ;
  TRIM_LO=(MINLO <V_LO);PUT MINLO = ;
  CALL SYMPUT('TRIM_HI',TRIM(LEFT(TRIM_HI))); PUT TRIM_HI = ;
   CALL SYMPUT('TRIM_LO',TRIM(LEFT(TRIM_LO))); PUT TRIM_LO = ;
END;
   RUN;
%END;

%IF( &TRIM_HI) OR (&TRIM_LO) %THEN %DO;
DATA BOXDATA (KEEP=&CLASS X Y TRIMMED)
 TRIM (KEEP=SRTCLASS X Y RENAME=(SRTCLASS=&CLASS));
 SET BOXDATA END=LASTOBS;
RETAIN SRTCLASS 4E23 V_HI &V_HI V_LO &V_LO;
```

```
TRIMMED=0;
%IF ( &TRIM_HI) %THEN %DO;
IF(Y > V_HI) THEN DO;
 Y=V_HI; TRIMMED=1;TRIMCT+1;OUTPUT TRIM;
END;
%END;
%IF(&TRIM_LO) %THEN %DO;
 IF(Y< V_LO) THEN DO;
 Y=V_LO;TRIMMED=1;TRIMCT+1;OUTPUT TRIM;
END;
%END;
IF(LASTOBS) THEN DO;
 PUT TRIMCT ' Y-VALUES TRIMMED' /;
END;
OUTPUT BOXDATA;

PROC PRINT DATA=TRIM LABEL N;
 VAR X Y &CLASS;
 LABEL X="X (&CLASS)"
      Y="Y(V_HI/V_LO)"
      &CLASS="FAKE &CLASS VALUE";
 TITLE1 "CONTENTS OF TRIM (SOURSE DS &DATA, OUTPUT DS &OUT)";
%END;
%IF ( &VERBOSE) %THEN %DO;
PROC PRINT DATA=BOXDATA;
 TITLE1 "CONTENTS OF BOXDATA( SOURCE DS &DATA, OUTPUT DS &OUT)";
%END;

DATA MEANS (KEEP=&CLASS X Y);
 SET SUMMARY(KEEP=&CLASS MEAN RENAME=(&CLASS=X MEAN=Y));
 RETAIN &CLASS 1E23;
%IF(&VERBOSE) %THEN %DO;
PROC PRINT DATA=MEANS LABEL;
 VAR X Y &CLASS;
 LABEL X="X(&CLASS)"
      Y="Y(MEAN)"
      &CLASS="FAKE &CLASS VALUE";
TITLE1 "CONTENTS OF MEANS (SOURCE DS &DATA, OUTPUT DS &OUT)";
%END;

%IF(&CONNECT) %THEN %DO;
DATA CONNECT(KEEP=&CLASS X Y);
 SET SUMMARY(KEEP=&CLASS &MIDDLE RENAME=(&CLASS=X &MIDDLE=Y));
 RETAIN &CLASS 2E23;
 %IF(&VERBOSE) %THEN %DO;
PROC PRINT DATA=CONNECT LABEL;
 VAR X Y &CLASS;
 LABEL X="X(&CLASS)"
      Y="Y(&MIDDLE)"
      &CLASS="FAKE &CLASS VALUE";
TITLE1 "CONTENTS OF CONNECT(SOURCE DS &DATA, OUTPUT DS &OUT)";
%END;
%END;
```

```
%IF(&KLUDGE) %THEN %DO;
DATA KLUDGE(KEEP=SRTCLASS X Y LABEL RENAME=(SRTCLASS=&CLASS));
 MERGE WORK(RENAME=(&VAR=Y)) SUMMARY (KEEP=&CLASS &LO_MARK &HI_MARK);
 BY &CLASS;
 RETAIN SRTCLASS 3E23;
 LENGTH LABEL $20;
 X=&CLASS;
 IF(Y<&LO_MARK) THEN DO;
  LABEL="< &LO_MARK";
  %IF(%QUOTE(&V_LO) NE ) %THEN %DO;
   IF(Y < &V_LO) THEN DO;
    LABEL=TRIM(LABEL)||',OFF PLOT';
   END;
  %END;
  OUTPUT;
END; ELSE IF (Y > &HI_MARK) THEN DO;
  LABEL="> &HI_MARK";
  %IF ( %QUOTE(&V_HI) NE ) %THEN %DO;
  IF ( Y > &V_HI ) THEN DO;
   LABEL=TRIM(LABEL)||', OFF PLOT';
  END;
 %END;
  OUTPUT;
END;
%IF( &VERBOSE) %THEN %DO;
PROC PRINT DATA=KLUDGE LABEL;
  VAR X Y LABEL &CLASS;
 LABEL X="X(&CLASS)"
      Y="Y(<&LO_MARK,>&HI_MARK)"
      &CLASS="FAKE &CLASS VALUE";
TITLE1 "CONTENTS OF KLUDGE(SOURCE DS &DATA, OUTPUT DS &OUT)";
%END;
%END;

DATA CBOXDATA;
  SET BOXDATA MEANS
%IF(&CONNECT) %THEN %DO;
  CONNECT
%END;
%IF(&KLUDGE) %THEN %DO;
 KLUDGE
%END;
%IF(&TRIM_HI OR &TRIM_LO) %THEN %DO;
  TRIM
%END;
;


%MACRO LABEL;
 LABEL Y="&VARLAB";
 LABEL X="&CLASSLAB";;
%MEND LABEL;
%MACRO FORMAT;
%IF(&VARFMT NE ) %THEN %DO; FORMAT Y &VARFMT; %END;
```

```
%IF(&CLASSFMT NE ) %THEN %DO; FORMAT X &CLASSFMT; %END;
%MEND FORMAT;

***************
 ADD GOPTIONS
**************;

%NOBS(DATA=SUMMARY)

%LET S=1;
SYMBOL&S R=&NOBS I=JOIN L=1 COLOR=WHITE W=2;* BOXES;

%LET S=%EVAL(&S+1);
SYMBOL&S R=1 I=NONE COLOR=WHITE V=PLUS W=2;* MEANS;
%IF(&CONNECT) %THEN %DO;
%LET S=%EVAL(&S+1);
SYMBOL&S R=1 I=JOIN L=1 COLOR=WHITE W=2;* CONNECT MEDIANS;
%END;

%IF (&KLUDGE) %THEN %DO;
%LET S=%EVAL(&S+1);
SYMBOL&S R=1 I=NONE COLOR=WHITE V=X W=2; * MAKE LKUDGE MARKS;
%END;

%IF(&TRIM_HI OR &TRIM_LO) %THEN %DO;
%LET S=%EVAL(&S+1);
SYMBOL&S R=1 I=NONE COLOR=WHITE V=TRIANGLE W=2;*MAKE TRIM MARKS;
%END;

PROC GPLOT DATA=CBOXDATA ;
PLOT Y*X=&CLASS/NOLEGEND HAXIS=AXIS1 VAXIS=AXIS2
;
AXIS1 LABEL=(H=1 F=CENTX &CLASSLAB)
    WIDTH=2
    VALUE=(F=CENTX);
AXIS2 LABEL=(H=1 R=0 A=90 F=CENTX
    &VARLAB ) WIDTH=2
    ORDER=&V_LO TO &V_HI BY &V_BY
    VALUE=(F=CENTX);
TITLE1  F=SWISS H=1 "&TITLE1";
TITLE2  F=SWISS H=1 "&TITLE2";
%FORMAT;
RUN;
%MEND BOXPLOT;

%BOXPLOT(
 DATA=FALLS,
 CLASS=MO,
 VAR=TP,
 CLASSFMT=,
 CLASSLAB=%STR('MONTH'),
 VARLAB=%STR('TOTAL PHOSPHORUS (mg/L)'),
 VERBOSE=0,
 F=.5,
```

```
WAISTTYP=TUKEY,
CONNECT=0,
WAIST=1,
TITLE1=Figure 3.5  Falls Lake Data,
TITLE2=Seasonal Boxplots for Total Phosphorus,
V_LO= 0,V_HI=.12,V_BY=.02
)
```

The lines above in the %BOXPLOT parentheses are used to: identify the dataset (DATA=FALLS), define the box classes (usually year or month; here CLASS=MO), identify the trend water quality variable (VAR=TP), define graph axis labels (CLASSLAB=%STR('MONTH'), VARLAB=%STR('TOTAL PHOSPHORUS (mg/L)')), define graph titles (TITLE1=Figure 3.5 Falls Lake Data, TITLE2=Seasonal Boxplots for Total Phosphorus), and specify the graph axis ranges and increments (V_LO= 0,V_HI=.12,V_BY=.02). The increments were determined by trial and error printouts; see Figure 3.5 in the text for the final result.

Corr.sas

The following graphic interface lines determine the features of the graphic output (screen display or printer output). For example, Figure 3.7 was printed (on the HP laserjet printer) using the second GOPTIONS statement; this statement was invoked by simply removing the asterisk at the beginning of the line.

```
GOPTIONS RESET=ALL DEVICE=EGA ROTATE GACCESS='SASGASTD>LPT1:' HBY=1
  FBY=CENTX ;
*GOPTIONS RESET=ALL DEVICE=HPLJS2 GACCESS='SASGASTD>LPT1:'
    HBY=1 FBY=CENTX HSIZE= 5.5 VSIZE =8 NOBORDER
     HORIGIN=1.325 VORIGIN=1 NOFILL NOPOLYGONFILL
     FTEXT=CENTX HTEXT=1  ;
OPTIONS PAGESIZE=56  LINESIZE=80 MPRINT  MISSING=' ';

DATA FALLS;
 INFILE 'A:FALLS1.DAT' FIRSTOBS=2 end=eof;
 INPUT OBS DATE TP TN ;
```

Identify data set and variables

The following are program lines to modify the data set to create a NEWDATE variable, insert points (.) for missing values, and to consolidate all of the day values to the same day each month (for ease in visual presentation). These manipulations may not be necessary for other data sets.

```
 NEWDATE= PUT(DATE,MMDDYY8.);
  MONTH=SUBSTR(NEWDATE,1,2);
  DAY=SUBSTR(NEWDATE,4,2);
  YEAR=SUBSTR(NEWDATE,7,2);

output;

if eof then do;
tp=.;
 month='01';day='01';year='83';output;
 month='02';day='01';year='83';output;
 month='03';day='01';year='83';output;
 month='12';day='01';year='83';output;
 month='11';day='01';year='87';output;
month='12';day='01';year='87';output;
end;

proc sort data=falls;by year month day;

proc means data=falls noprint;
by year month ;
var tp;
output out=falls mean=tp;
```

Identify variable for analysis

```
data falls;
set falls;
length mo da yr 8.;
```

```
mo=month;
da=1;
yr=year;
date=mdy(mo,da,yr);

 RUN;

%MACRO CORR(DATA,VAR,TIME,NOBS,TITLE1);
************************************************************
  CORRELOGRAM: PLOT AND PRINT


   INPUTS:  DATA = SAS DATA SET CONTAINING TEST INFORMATION
         VAR  = SINGLE VARIABLE OF INTEREST FOR WHICH THE
              DIAGNOSTICS WILL BE PERFORMED
         TIME = SAS VARIABLE REPRESENTING SEQUENTIAL TIME
         NOBS = NUMBER OF OBSERVATIONS
              (MISSING + NONMISSING)
        TITLE1 = FIRST OUTPUT TITLE


    DATA ASSUMPTIONS:
          - ONE OBSERVATION PER TIME PERIOD
          - MISSING OBSERVATIONS ARE ALLOWED:

************************************************************;
***  PLOT DATA, AND OUTPUT STATISTICS              ***;
************************************************************;

PROC SORT DATA=&DATA;BY &TIME;

DATA CORR;
SET &DATA;
LAG0=&VAR;
LAG1=LAG1(&VAR);
LAG2=LAG2(&VAR);
LAG3=LAG3(&VAR);
LAG4=LAG4(&VAR);
LAG5=LAG5(&VAR);
LAG6=LAG6(&VAR);
LAG7=LAG7(&VAR);
LAG8=LAG8(&VAR);
LAG9=LAG9(&VAR);
LAG10=LAG10(&VAR);
LAG11=LAG11(&VAR);
LAG12=LAG12(&VAR);
LAG13=LAG13(&VAR);
LAG14=LAG14(&VAR);
LAG15=LAG15(&VAR);

PROC CORR DATA=CORR noprint out=out ;
VAR LAG0 LAG1 LAG2 LAG3 LAG4 LAG5 LAG6 LAG7 LAG8 LAG9 LAG10
LAG11 LAG12 LAG13 LAG14 LAG15;
```

```
DATA OUT;
SET OUT;
KEEP LAG0-LAG15 STE0-STE15 UPPER0-UPPER15 LOWER0-LOWER15;
ARRAY V1{16} LAG0-LAG15;
ARRAY V2{16} STE0-STE15;
ARRAY V3{16} UPPER0-UPPER15;
ARRAY V4{16} LOWER0-LOWER15;
LENGTH OBSNUM 8.;
OBSNUM=&NOBS;
IF _TYPE_='CORR' and _NAME_='LAG0';
DO I=1 TO 16;
 K=I-1;
 IF K=0 THEN DO;
  STE0=1/SQRT(OBSNUM);
 END;
 ELSE DO;
  SUM=0;
   DO J=1 TO K ;
    SUM=SUM + (V1{J}**2);
   END;
  V2{I}=SQRT(1+2*SUM)*(1/SQRT(OBSNUM) );
 END;
END;
DO I=1 TO 16;
  V3{I}=2*V2{I};
  V4{I}=-2*V2{I};
END;

DATA PLT;
SET OUT;
ARRAY V1{16} LAG0-LAG15;
ARRAY V2{16} UPPER0-UPPER15;
ARRAY V3{16} LOWER0-LOWER15;
ARRAY V4{16} STE0-STE15;
LENGTH X 8.;
KEEP VAR UP LOW X STE;
DO I=1 TO 16;
 X=I-1;
 VAR=V1{I};
 UP=V2{I};
 LOW=V3{I};
 STE=V4{I};
 OUTPUT;
END;

PROC PRINT DATA=PLT LABEL;
VAR X VAR STE UP LOW;
LABEL VAR='CORRELATION'
     X="LAGGED &VAR"
     UP='UPPER LIMIT'
    LOW='LOWER LIMIT'
;
TITLE1 "&TITLE1";
TITLE2 ' PRINT OF DATA USED IN CORRELOGRAM';
```

```
PROC GPLOT DATA=PLT;
PLOT VAR*X='o' UP*X='U' LOW*X='L'/ VAXIS=-1.0 TO 1.0 BY 0.20 VREF=0.0
 OVERLAY;
LABEL VAR='CORRELATION'
    X="LAGGED &VAR"
    UP='UPPER LIMIT'
   LOW='LOWER LIMIT';
TITLE2 'CORRELOGRAM WITH UPPER AND LOWER 95% CONFIDENCE LIMITS';

 RUN;
%MEND CORR;
```

The following statement in parentheses defines the data set, variables, and figure title for the macro. This program was used to create Figure 3.7 in the text.

```
%CORR(DATA=FALLS,
      VAR=TP,
      TIME=DATE,
      NOBS=60,
      TITLE1=Figure 3.7. FALLS LAKE DATA
      );
```

Kens.sas

The following graphic interface lines determine the features of the graphic output (screen display or printer output). The second GOPTIONS statement can be invoked by simply removing the asterisk at the beginning of the line.

```
GOPTIONS RESET=ALL DEVICE=EGA GACCESS='SASGASTD>LPT1:' HBY=1
  FBY=CENTX ;
*GOPTIONS RESET=ALL DEVICE=HPLJS2 GACCESS='SASGASTD>LPT1:'
   HBY=1 FBY=CENTX HSIZE= 5.5 VSIZE =8 NOBORDER
    HORIGIN=1.325 VORIGIN=1 NOFILL NOPOLYGONFILL
    FTEXT=CENTX HTEXT=1  ;
OPTIONS PAGESIZE=56  MPRINT  MISSING=' ';

DATA FALLS;
 INFILE 'A:FALLS1.DAT' FIRSTOBS=2 end=eof;
 INPUT OBS DATE TP TN ;
```

Identify data set and variables

```
 NEWDATE= PUT(DATE,MMDDYY8.);
  MONTH=SUBSTR(NEWDATE,1,2);
  DAY=SUBSTR(NEWDATE,4,2);
  YEAR=SUBSTR(NEWDATE,7,2);

output;

if eof then do;
tp=.;
 month='01';day='01';year='83';output;
 month='02';day='01';year='83';output;
 month='03';day='01';year='83';output;
 month='12';day='01';year='83';output;
 month='11';day='01';year='87';output;
month='12';day='01';year='87';output;
end;

proc sort data=falls;by year month day;
proc means data=falls noprint;
by year month ;
var tp;
output out=falls mean=tp;
```

Identify variable for analysis

```
data falls;
set falls;
length mo da yr 8.;
mo=month;
da=1;
yr=year;
date=mdy(mo,da,yr);
```

```
%MACRO KENDALL(DATA,NOBS,VAR,TIME,NSEASONS,TITLE1);
************************************************************
  KENDALL TAU - TEST FOR RANDOMNESS AGAINST TREND
        - INCLUDES ADJUSTMENT FOR TIED VALUES OF X
        - EXECUTES A FORTRAN PROGRAM THAT CALCULATES
          KENDALLS TAU:  THE PROGRAM CAN CALCULATE

          A SIMPLE KENDALL (NSEASONS=1), A SEASONAL
          KENDALL (NSEASONS>1), OR A SEASONAL KENDALL
          FOR AUTOCORRELATED DATA (NSEASONS>1)
        - THE FORTRAN PROGRAM WAS OBTAINED FROM
          J.R. SLACK, U.S. DEPARTMENT OF THE INTERIOR,
          345 MIDDLEFIELD RD.,MS496,MENLO PARK,CA 94025
        - THE PROGRAM IS MOST EFFECTIVE IF SAS IS
          EXECUTED USING -XWAIT OFF WHEN CALLED FROM DOS:
          OR, -XWAIT OFF CAN BE STORED IN THE
          CONFIG.SAS FILE

  INPUTS:     DATA = SAS DATA SET CONTAINING TEST INFORMATION
              NOBS = TOTAL NUMBER OF OBSERVATIONS
                          (MISSING+NONMISSING)
              VAR  = SINGLE VARIABLE OF INTEREST AGAINST WHICH THE
                          TEST WILL BE PERFORMED
              TIME = VARIABLE REPRESENTING SEQUENTIAL TIME
              NSEASONS = NUMBER OF SEASONS **** NOTE: IF A SIMPLE
                          KENDALLS TAU STATISTIC IS DESIRED, THEN
              SET NSEASONS=1
              TITLE1 = FIRST OUTPUT TITLE


   DATA ASSUMPTIONS:
        - ONE OBSERVATION PER TIME PERIOD
        - MUST HAVE EQUAL NUMBER OF DATA VALUES
          PER YEAR:I.E. MUST HAVE SEQUENTIAL
          DATA (E.G. 5 YEARS OF MONTHLY DATA
          = 60 OBSERVATIONS)
        - MISSING OBSERVATIONS ARE ALLOWED
        - NO MORE THAN 1040 OBSERVATIONS ALLOWED
        - NOBS/NSEASONS MUST BE A WHOLE NUMBER.
************************************************************;
PROC SORT DATA=&DATA;BY &TIME;

DATA ONE;
SET &DATA;
FILE 'TEMP.DAT';
LENGTH T1 T2 8.;
ZERO=0;
IF &VAR=. THEN &VAR=-99999.;
T1=&NOBS;
T2=&NSEASONS;
IF _N_=1 THEN DO;
 PUT @1 ZERO 1. T1 2-6 T2 7-9;
END;
PUT @ 1 &VAR 10.3;
```

-XWAIT is not essential; it simply eliminates the need to hit the "Return" key to resume SAS.

```
RUN;
```

The next line calls the Fortran program (Kendall4). If Kendall4 is not on the SAS directory (e.g., it is on c:\trend), then you must change this line to reflect the correct location (e.g., X 'C:\TREND\KENDALL4';).

```
X 'KENDALL4';
RUN;

DATA TAUIN;
INFILE 'OUT.DAT';
INPUT TAU PWITHOUT PWITH SLOPE ;

PROC PRINT DATA=TAUIN LABEL;
VAR TAU PWITHOUT PWITH SLOPE;
LABEL TAU='TAU STATISTIC'
    PWITHOUT='P-VALUE WITHOUT SERIAL CORRELATION'
    PWITH=  'P-VALUE WITH SERIAL CORRELATION'
    SLOPE='SLOPE STATISTIC';

TITLE1 "&TITLE1";
TITLE2 'KENDALL TAU';
RUN;
%MEND;
```

The following statement in parentheses defines the data set, variables, and figure title for the macro.

```
%KENDALL(DATA=falls,
      NOBS=60,
      VAR=tp,
      TIME=DATE,
      NSEASONS=1,
      TITLE1=TEST DATA
      );
```

Adjust.sas

The following graphic interface lines determine the features of the graphic output (screen display or printer output). The second GOPTIONS statement can be invoked by simply removing the asterisk at the beginning of the line.

```
LIBNAME C'C:\';
GOPTIONS RESET=ALL DEVICE=EGA ROTATE GACCESS='SASGASTD>LPT1:' HBY=1
  FBY=CENTX ;
*GOPTIONS RESET=ALL DEVICE=HPLJS2 ROTATE GACCESS='SASGASTD>LPT1:'
    HBY=1 FBY=CENTX HSIZE= 5.5 VSIZE =8 NOBORDER
     HORIGIN=1.325 VORIGIN=1 NOFILL NOPOLYGONFILL
     FTEXT=CENTX HTEXT=1  ;
OPTIONS PAGESIZE=56  MPRINT  MISSING=' ';

DATA FALLS;
 INFILE A:FALLS1.DAT' FIRSTOBS=2 end=eof;
 INPUT OBS DATE TP TN ;

 NEWDATE= PUT(DATE,MMDDYY8.);
  MONTH=SUBSTR(NEWDATE,1,2);
  DAY=SUBSTR(NEWDATE,4,2);
  YEAR=SUBSTR(NEWDATE,7,2);

output;

if eof then do;
tp=.;
 month='01';day='01';year='83';output;
 month='02';day='01';year='83';output;
 month='03';day='01';year='83';output;
 month='12';day='01';year='83';output;
 month='11';day='01';year='87';output;
month='12';day='01';year='87';output;
end;

proc sort data=falls;by year month day;

proc means data=falls noprint;
by year month ;
var tp;
output out=falls mean=tp;

data falls;
set falls;
length mo da yr 8.;
mo=month;
da=1;
yr=year;
date=mdy(mo,da,yr);
```

Identify data set and variables

Identify variable for analysis

```
%MACRO ADJUST(DATA,VAR,TIME,SEASON,NSEASONS,NOBS,TITLE1);
*************************************************************
  ADJUST - PROGRAM TO DETREND AND DESEASONALIZE THE RAW DATA
      - THE SLOPE ESTIMATE NEEDED
        FOR DETRENDING IS OBTAINED FROM THE OUTPUT
        OF THE KENS MACRO
      - THE OUTPUT DATA SET IS CALLED: ADJUST,
        THE NEW VARIABLE IS CALLED: ADJUSTED
      - THE ADJUST DATA SET MAY BE ENTERED INTO THE KEN MACRO

   INPUTS:  DATA = SAS DATA SET CONTAINING RAW DATA
          VAR  = SINGLE VARIABLE OF INTEREST FOR WHICH THE
               DIAGNOSTICS WILL BE PERFORMED
          TIME = SAS VARIABLE REPRESENTING SEQUENTIAL TIME
        SEASON = NAME OF THE SAS VARIABLE REPRESENTING
               THE SEAONAL TIME COMPONENT (E.G., MONTH)
       NSEASON = THE NUMBER OF SEASONS PER YEAR
          NOBS = TOTAL NUMBER OF OBSERVATIONS
               (MISSING + NONMISSING)
        TITLE1 = FIRST OUTPUT TITLE


    DATA ASSUMPTIONS:
          - ONE OBSERVATION PER TIME PERIOD
          - MISSING OBSERVATIONS ARE ALLOWED:
*************************************************************************;

PROC SORT DATA=&DATA:
BY &SEASON;

PROC UNIVARIATE DATA=&DATA NOPRINT;
 VAR &VAR;
BY &SEASON;
OUTPUT  OUT=OUT1 MEDIAN=S_MEDIAN;

*********************************************************************
****  ADJUST THE DATA BY SUBTRACTING THE MEDIAN SEASONAL VALUE
*********************************************************************;

DATA ADJUST;
MERGE &DATA OUT1;
BY &SEASON;
ADJUSTED=&VAR-S_MEDIAN;
YEAR=YEAR(&TIME);

*********************************************************************
****  ADJUST THE DATA BY SUBTRACTING THE TREND LINE
*********************************************************************;

PROC SORT DATA=ADJUST;
BY &TIME;

DATA TEMP;
```

```
SET ADJUST;
FILE 'TEMP.DAT';
LENGTH T1 T2 8.;
ZERO=0;
IF ADJUSTED=. THEN ADJUSTED=-99999.;
T1=&NOBS;
T2=&NSEASONS;
IF _N_=1 THEN DO;
 PUT @1 ZERO 1. T1 2-6 T2 7-9;
END;
PUT @ 1 ADJUSTED 10.3;
 run;
```

The next line calls the Fortran program (Kendall3). If Kendall3 is not on the SAS directory (e.g., it is on c:\trend), then you must change this line to reflect the correct location (e.g., X 'C:\TREND\KENDALL3';).

```
X 'A:KENDALL3';
 run;

DATA TAUIN;
INFILE 'OUT.DAT';
INPUT TAU PWITHOUT PWITH SLOPE ;

DATA ADJUST;
SET ADJUST;
IF _N_=1 THEN SET TAUIN;

PROC SORT DATA=ADJUST;
BY YEAR;

DATA C.ADJUST;
SET ADJUST;
BY YEAR;
RETAIN YRCNT 0 NEWTIME;
LENGTH SEANUM 8.;
SEANUM=&NSEASONS;
IF FIRST.YEAR THEN DO;
 YRCNT=YRCNT+1;
 NEWTIME=YRCNT;
 ADJUSTED=ADJUSTED-(SLOPE*NEWTIME);
 OUTPUT;
END;
ELSE DO;
 NEWTIME=NEWTIME+(1/SEANUM);
 ADJUSTED=ADJUSTED-(SLOPE*NEWTIME);
 OUTPUT;
END;

PROC SORT DATA=C.ADJUST;BY &TIME;

PROC PRINT DATA=C.ADJUST;
VAR &TIME ADJUSTED  SLOPE S_MEDIAN;
FORMAT &TIME MMDDYY8.;
```

```
TITLE1 "&TITLE1" ;
TITLE2 'PRINT OF ADJUSTED DATA';

PROC PLOT DATA=C.ADJUST;
PLOT ADJUSTED*&TIME;
FORMAT &TIME MMDDYY8.;
TITLE2 'PLOT OF ADJUSTED DATA';

RUN;
%MEND ADJUST;
```

The following statement in parentheses defines the data set, variables, and figure title for the macro.

```
%ADJUST(DATA=FALLS,
      VAR=TP,
      TIME=DATE,
      SEASON=MONTH,
      NSEASONS=12,
      NOBS=60,
      TITLE1=FALLS TEST DATA
      );
```

Corradj.sas

The following graphic interface lines determine the features of the graphic output (screen display or printer output). The second GOPTIONS statement can be invoked by simply removing the asterisk at the beginning of the line.

```
LIBNAME C'C:\';
GOPTIONS RESET=ALL DEVICE=EGA ROTATE GACCESS='SASGASTD>LPT1:' HBY=1
  FBY=CENTX ;
*GOPTIONS RESET=ALL DEVICE=HPLJS2 ROTATE GACCESS='SASGASTD>LPT1:'
    HBY=1 FBY=CENTX HSIZE= 5.5 VSIZE =8 NOBORDER
     HORIGIN=1.325 VORIGIN=1 NOFILL NOPOLYGONFILL
     FTEXT=CENTX HTEXT=1  ;
OPTIONS PAGESIZE=56  LINESIZE=80 MPRINT  MISSING=' ';

DATA FALLS;
 INFILE 'A:FALLS1.DAT' FIRSTOBS=2 end=eof;
 INPUT OBS DATE TP TN ;

 NEWDATE= PUT(DATE,MMDDYY8.);
  MONTH=SUBSTR(NEWDATE,1,2);
  DAY=SUBSTR(NEWDATE,4,2);
  YEAR=SUBSTR(NEWDATE,7,2);

output;

if eof then do;
tp=.;
 month='01';day='01';year='83';output;
 month='02';day='01';year='83';output;
 month='03';day='01';year='83';output;
 month='12';day='01';year='83';output;
 month='11';day='01';year='87';output;
month='12';day='01';year='87';output;
end;

proc sort data=falls;by year month day;

proc means data=falls noprint;
by year month ;
var tp;
output out=falls mean=tp;

data falls;
set falls;
length mo da yr 8.;
mo=month;
da=1;
yr=year;
date=mdy(mo,da,yr);
```

Identify data set and variables

Identify variable for analysis

```
RUN;

%MACRO CORR(DATA,VAR,TIME,NOBS,TITLE1);
************************************************************
  CORRELOGRAM: PLOT AND PRINT


   INPUTS:  DATA = SAS DATA SET CONTAINING TEST INFORMATION
         VAR  = SINGLE VARIABLE OF INTEREST FOR WHICH THE
              DIAGNOSTICS WILL BE PERFORMED
         TIME = SAS VARIABLE REPRESENTING SEQUENTIAL TIME
         NOBS = NUMBER OF OBSERVATIONS
              (MISSING + NONMISSING)
        TITLE1 = FIRST OUTPUT TITLE


    DATA ASSUMPTIONS:
           - ONE OBSERVATION PER TIME PERIOD
           - MISSING OBSERVATIONS ARE ALLOWED:

************************************************************;
*** PLOT DATA, AND OUTPUT STATISTICS              ***;
************************************************************;

PROC SORT DATA=&DATA;BY &TIME;

DATA CORR;
SET &DATA;
LAG0=&VAR;
LAG1=LAG1(&VAR);
LAG2=LAG2(&VAR);
LAG3=LAG3(&VAR);
LAG4=LAG4(&VAR);
LAG5=LAG5(&VAR);
LAG6=LAG6(&VAR);
LAG7=LAG7(&VAR);
LAG8=LAG8(&VAR);
LAG9=LAG9(&VAR);
LAG10=LAG10(&VAR);
LAG11=LAG11(&VAR);
LAG12=LAG12(&VAR);
LAG13=LAG13(&VAR);
LAG14=LAG14(&VAR);
LAG15=LAG15(&VAR);

PROC CORR DATA=CORR noprint out=out ;
VAR LAG0 LAG1 LAG2 LAG3 LAG4 LAG5 LAG6 LAG7 LAG8 LAG9 LAG10
LAG11 LAG12 LAG13 LAG14 LAG15;

DATA OUT;
SET OUT;
KEEP LAG0-LAG15 STE0-STE15 UPPER0-UPPER15 LOWER0-LOWER15;
ARRAY V1{16} LAG0-LAG15;
ARRAY V2{16} STE0-STE15;
```

```
ARRAY V3{16} UPPER0-UPPER15;
ARRAY V4{16} LOWER0-LOWER15;
LENGTH OBSNUM 8.;
OBSNUM=&NOBS;
IF _TYPE_='CORR' and _NAME_='LAG0';
DO I=1 TO 16;
 K=I-1;
 IF K=0 THEN DO;
  STE0=1/SQRT(OBSNUM);
 END;
 ELSE DO;
  SUM=0;
  DO J=1 TO K ;
   SUM=SUM + (V1{J}**2);
  END;
 V2{I}=SQRT(1+2*SUM)*(1/SQRT(OBSNUM) );
 END;
END;
DO I=1 TO 16;
 V3{I}=2*V2{I};
 V4{I}=-2*V2{I};
END;

DATA PLT;
SET OUT;
ARRAY V1{16} LAG0-LAG15;
ARRAY V2{16} UPPER0-UPPER15;
ARRAY V3{16} LOWER0-LOWER15;
ARRAY V4{16} STE0-STE15;
LENGTH X 8.;
KEEP VAR UP LOW X STE;
DO I=1 TO 16;
 X=I-1;
 VAR=V1{I};
 UP=V2{I};
 LOW=V3{I};
 STE=V4{I};
 OUTPUT;
END;

PROC PRINT DATA=PLT LABEL;
VAR X VAR STE UP LOW;
LABEL VAR='CORRELATION'
    X="LAGGED &VAR"
    UP='UPPER LIMIT'
   LOW='LOWER LIMIT'
;
TITLE1 "&TITLE1";
TITLE2 ' PRINT OF DATA USED IN CORRELOGRAM';

PROC PLOT DATA=PLT;
PLOT VAR*X='*' UP*X='U' LOW*X='L'/ VAXIS=-1.0 TO 1.0 BY 0.20 VREF=0.0
 OVERLAY;
LABEL VAR='CORRELATION'
```

```
    X="LAGGED &VAR"
    UP='UPPER LIMIT'
   LOW='LOWER LIMIT';
TITLE2 'CORRELOGRAM WITH UPPER AND LOWER 95% CONFIDENCE LIMITS';

 RUN;
%MEND CORR;
```

The following statement in parentheses defines the data set, variables, and figure title for the macro. This program was used to create Figure 3.15 in the text.

```
%CORR(DATA=C.ADJUST,
      VAR=ADJUSTED,
      TIME=DATE,
      NOBS=60,
      TITLE1=FALLS TEST DATA
      );
```