Review of Environmental Monitoring Methods: Trend Detection

Trent L. McDonald[*], Bryan F. J. Manly[*], and Ryan M. Nielson[*]

[*]Western EcoSystems Technology, Inc., 2003 Central Ave, Cheyenne WY 82001.

Contact author: Trent McDonald, Western EcoSystems Technology, Inc., 2003 Central Ave, Cheyenne WY 82001; Phone: 1-307-634-1756; Email: tmcdonald@west-inc.com

Abstract: Detecting trend in an environmental parameter can be an important objective of environmental surveys. However, the best analysis to apply to a particular set of monitoring data is often unclear, due to the large number of available analyses and recent advances in this branch of statistics. This paper attempts to clarify trend detection analyses by reviewing and summarizing the general types, their characteristics, and examples of each. Based on our literature review, we categorized trend analyses into the 3 general classes of *design-based, parametric*, and *non-parametric*. In addition to describing these classes, we applied 6 general purpose trend analyses to a single data set to illustrate their computation. These 6 analyses consisted of 2 design-based methods, 2 parametric, and 2 non-parametric methods. All were applied to the same set of monitoring data on dissolved reactive phosphorus (DRP) in 16 rivers on the south island of New Zealand. Throughout, special attention was paid to spatial and temporal correlation in the collected data, and relatively simple adjustments or alternatives to each method are described that ameliorate the deleterious effects that correlation has on some analyses.

Keywords: finite population, rotating panel, CUSUM, Mann-Kendall, long-term change, abrupt change

1. Introduction

A primary objective of most long-term environmental monitoring surveys is to detect and estimate trends in the parameters they measure. This objective is so prevalent that statements about trend pervade the publications of national and international agencies concerned with environmental conditions (Esterby, 1993). Yet designing a long-term monitoring program to estimate trend is not easy. Among other things, researchers must decide upon the definition of sampling units, how to disperse the sample through space and time, how to construct an accurate sampling frame, whether to conduct equiprobable sampling or variable probability sampling, whether to stratify and how, and what to do with inaccessible or non-responsive units. Once data are collected researchers must decide which parameters in the data are important, what trend analyses are appropriate, which trend analysis is best, and whether spatial and temporal correlation will cause problems during analysis.

Trend analyses in particular, are deceptively difficult to conceptualize and carry out with confidence. Consider that the very definition of trend itself is not fixed. Trends can be defined as relatively sudden shifts caused by a single event (e.g., contamination or cleanup), or monotonic changes occurring over multiple time periods (Hirsch et al., 1991). Moreover, both types of trend (sudden and monotonic) can exist in parameters defined on individual members of a population (i.e., gross trend), or on the population as a whole (i.e., net trend) (Urquhart et al., 1998; McDonald, 2003). Taken together, these survey design and analysis issues can make optimum choices unclear.

To add to the confusion, over the past decade there has been an abundance of statistical research into the design of long-term monitoring surveys and the analysis of data that they collect. In this literature, there exists an abundance of slightly or fundamentally different analyses that all purport to detect trend, with few general recommendations. McDonald (2003) summarized the salient features of rotational and non-rotational effort sampling designs commonly used for long-term monitoring, but no similar non-technical summary of trend detection analyses is available.

For this paper, we conducted an extensive search of the statistics literature, targeting analyses that could be used to detect trend in data typically collected by long-term monitoring projects. We then classified the trend analyses we found into broad families based on the assumptions they make. Two key assumptions are about the form of the trend (e.g., whether it is linear, quadratic, etc.) and whether the measured response is assumed to follow a particular statistical distribution (e.g., parametric vs. nonparametric analyses). In this paper, we describe in detail 6 of these analyses that we feel have wide applicability, and illustrate each on a common data set. These mini case studies were designed to illuminate the assumptions and calculations of each method, but not all the mathematical details in the original papers. Throughout, certain aspects of environmental survey design and analysis, such as sample unit definition, frame construction, membership design, revisit design, etc, will be assumed. Many of those topics have been summarized in McDonald (2003). Complete background on finite population surveys can be found in texts like Kish (1965), Cochran (1977), Krishnaiah (1988), Särndal (1992), Thompson (1992), Lohr (1999), and others.

2. Literature Survey

We conducted our literature review by searching the 2003 (version 11) Cumulative Index to Statistics (CIS) (published by the American Statistical Association and the Institute of Mathematical Statistics; www.statindex.org) for keywords ‘trend’, ‘change’, ‘environmental’, ‘monitoring’, both singly and in combination. A large number of other related phrases, such as ‘trend analysis’ and ‘change detection’ were also searched. At the time, the CIS contained full coverage of 162 core statistics journals, as well as partial coverage of numerous other non-core journals that have significant quantitative components. Most core journals were fully indexed back to 1975, with sporadic coverage beyond that. Our version the CIS contained articles through 2002, with partial coverage of 2003. In all, the CIS contained 250,786 records, including 182,390 articles from 1414 different journals, 41,990 articles in edited books and proceedings, and 11,518 books. Because we used CIS as our primary database, our literature search focused on articles in statistics journals, with only sporadic representation of non-statistics journals.

Among the results of our literature search, we found a few ‘essential’ volumes that we believe all statisticians involved in designing long-term monitoring surveys should have. The first two of these are conference proceedings. The conference entitled, Ecological Resource Monitoring: Change and Trend Detection, was held in 1996 and selected papers appear in a special issue of Ecological Applications (1998, Vol 8, No. 2) (see Edwards (1998) for a summary). The second conference, Environmental Monitoring Surveys Over Time, was held in 1998 and selected papers were published in a special issue of the Journal of Agricultural, Biological, and Environmental Statistics (1999, Vol 4, No. 4) (see Olsen (1999) for a summary). Another useful volume pertaining to design-based analyses for the statistically adept reader is Krishnaiah (1988). Model-assisted non-parametric analyses are well covered by Särndal et al. (1992). Berryman et al (1988) and Esterby (1993) contain useful summaries of many aspects of trend detection.

Examples we found of on-going large scale environmental surveys include the National Resources Inventory (NRI) (Nusser and Goebel, 1997; Nusser et al., 1998), the Forest Inventory and Analysis (FIA) (Leatherberry et al., 1995; Moisen and Edwards, 1999; Reams and Deusen, 1999; McRoberts and Hansen, 1999), the Forest Health Monitoring survey (FHM) (Eager et al., 1991), the National Wetlands Inventory (NWI) (Ernst et al., 1995), and various components of the Environmental Monitoring and Assessment Program (1990).

Categorization of the trend analyses we found into a small number of families was very difficult due to the diversity of assumptions and methods used. In the end, we decided to base our categorization of analyses on the types of assumptions they made, and not on any particular methodological technique. Based on our literature review, we categorized trend detection analyses into three families: pure *design-based* analyses, *parametric* analyses, and *nonparametric* analyses (Figure 1). The *design-based* family contained analyses that estimate a parameter at every time period, and summarized changes in those estimates through time. The key characteristics of design-based analyses were that they made few if any assumptions about the form of the trend and used properties of the sampling design (i.e., randomization) to make inferences about trend in the population. The family of *parametric* analyses was large and included ordinary linear regression models, generalized linear models, mixed linear models, and time series analysis. Parametric analyses make inferences about trend in a population's parameter by assuming that the relationship between mean responses and covariates has a particular form (e.g., linear), or by assuming statistical errors are random realizations from a particular distribution. We classified as parametric those analyses that calculate a parametric slope parameter for all sampled units and summarize that parameter or its distribution over all units in the population. We also classified the trend oriented model-assisted analyses of Särndal et al. (1992) as parametric because they assumed a particular (linear) model for trend, but we admit that calling these analyses parametric is a bit odd because they generally do not assume responses follow a particular distribution. Analyses in the *nonparametric* family do not assume any particular model for changes, nor do they assume responses follow a particular statistical distribution. The nonparametric family includes analyses such as the Wilcoxon rank sum test, the Kruskal-Wallis test (Conover, 1980; Hollander and Wolfe, 1999), the Mann-Kendall test (Mann, 1945; Kendall, 1975; Cabilio and Tilley, 1999), a host of modifications to the Mann-Kendall test to account for seasonality and serial correlation (Hirsch et al., 1982; Hirsch and Slack, 1984; Berryman et al., 1988; Lettenmaier, 1988; Hirsch et al., 1991), and the CUSUM technique (MacNally and Hart, 1997; Manly and Mackenzie, 2000; Manly and Mackenzie, 2003). Associated with the nonparametric procedures are a multitude of techniques for estimating the magnitude of trend once it is detected (Sen, 1968; Hirsch et al., 1991).

3. Trend Analyses

In this section, we describe 6 analyses that we feel have high potential for application in a wide variety of surveys. This by no means detracts for the utility of analyses that are not mentioned here. In fact, it may be the case that the best analysis for a particular response is not mentioned, or involves significant modifications to one of the reported routines. Outside the set of 6 analyses we chose to illustrate, a few are worth mentioning in passing as potentially useful. The first is the novel parametric and highly model dependent analyses of the North American Breeding Bird Survey (BBS) reported by Link and Sauer (1998; 2002). Data collected by the BBS would be practically impossible to analyze using a design-based approach because the sampling design is very difficult to quantify. Ericson (1988) and Woodward (1999) both report interesting examples of Bayesian model-based analyses. The analysis (and associated sampling scheme) published by Wikle *et al.* (1999) takes the same general approach as Guttorp (1994) and Huang (1996) in that their analysis attempts to adaptively locate sample points in space to minimize a variance criterion like average prediction variance. Once data are collected at a set of locations during occasion *t*, sample locations for occasion *t* + 1 are moved in an attempt to minimize prediction errors. With limited resources, this analysis (and design) might be worth considering if estimating an accurate map of the response is an important objective of the survey.

Our list of 6 general purpose trend analyses is comprised of the design-based Horvitz-Thompson and MVLUE analyses, the parametric unit-slope and linear model analyses, and the non-parametric Mann-Kendall and CUSUM procedures. We describe these analyses, and apply them to a common data set on dissolved reactive phosphorous (DRP) in 16 stream segments on the south island of New Zealand (Table 1, Figure 2)

3.1 DESIGN-BASED ANALYSES

No single reference exists for design-based analyses, but the general theory for these methods can be found in finite population sampling texts such as Cochran (1977), Cassel et al. (1977), Scheaffer et al. (1986), Thompson (1992), Sändarl et al. (1992), and Lohr (1999). Design-based analyses for trend are covered by Krishnaiah (1988), Binder and Hidiroglou (1988), Nijman et al. (1990), and Singh et al. (2001). To paraphrase Olsen et al. (1998), the key characteristic of design-based analyses is that they rely solely on repeated random realizations of the sampling mechanism as the basis for inference to unsampled units. Specifically, design-based analyses do not assume a statistical distribution for responses, nor do they assume trend has a particular form (e.g., linear). Rather, design-based analyses require data to be collected using a rigorous probability sample. The most useful characteristic of design-based analyses is their objectivity, unbiasedness, and lack of assumptions. These characteristics make it very difficult to argue against properly derived results from properly conducted design-based analyses.

Pure design-based analyses view each survey in time as a separate sample and attempt to summarize status of the resource at each occasion. Once status at each occasion is quantified, trend of the status estimates is inferred. These methods are generally simple to compute and interpret, and allow different or independent sample designs at each sample occasion. In the remainder of this section, we present the general purpose Horvitz-Thompson estimator that can be applied each time period to estimate current status, followed by the MVLUE estimator that can uses an assumed form for correlation of responses on the same unit through time to improve occasion-by-occasion estimates of trend.

3.1.1 Horvitz-Thompson Estimation

The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) of status at a particular time is well known. Let y_u represent the response from unit u at time t, which is assumed to be measured without error (t has been omitted, when possible, from the notation y_u for brevity). Assume that the population at sample occasion t has N_t units in it. Assume that a subset of population units, denoted by S_t, was collected at time t by a sample design such that the probability of unit u becoming a member of S_t was π_u prior to drawing the sample. Further, we assume the probability of sampling the pair of units u and v in S_t is π_{uv} prior to drawing the sample. Under these assumptions, the Horvitz-Thompson estimator of the mean of y_u at time t is,

$$\hat{y}_t = N_t^{-1}\sum_{u\in S_t}\frac{y_u}{\pi_u} \tag{1}$$

with estimated variance

$$\widehat{\text{var}}(\hat{y}_t) = N_t^{-2}\left(\sum_{u\in S_t}\frac{1-\pi_u}{\pi_u^2}y_u^2 + \sum_{u\in S_t}\sum_{v\in S_t, v\neq u}\frac{\pi_{uv}-\pi_u\pi_v}{\pi_{uv}\pi_u\pi_v}y_u y_v\right) \tag{2}$$

or,

$$\widehat{\text{var}}(\hat{y}_t) = -\frac{1}{2N_t^2}\sum_{u\in S_t}\sum_{v\in S_t}\frac{\pi_{uv}-\pi_u\pi_v}{\pi_{uv}}\left(\frac{y_u}{\pi_u}-\frac{y_v}{\pi_v}\right)^2 \tag{3}$$

(Cochran, 1977; Sarndal et al., 1992; Thompson, 1992). Equation (1) is unbiased for the population mean provided $\pi_u > 0$ for all u. Equations (2) and (3) are unbiased for the variance of the estimator in (1) provided $\pi_{uv} > 0$ for all u and v. Equation (2) is due to Horvitz and Thompson (1952), while

(3) is due to both Yates and Grundy (1953) and Sen (1953). While (2) and (3) can both produce negative variance estimates, anecdotal evidence suggests that (3) produces fewer negative estimates than (2).

Assuming S_t has n_t units in it and that $\pi_u = n_t/N_t$ and $\pi_{uv} = [n_t(n_t\text{-}1)]/[N_t(N_t\text{-}1)]$ for all units u and v (i.e., simple random sampling), an approximate 95% confidence interval for the true mean at time t is $\hat{\bar{y}}_t \pm 1.96\sqrt{\widehat{\text{var}}(\hat{\bar{y}}_t)}$. If $\pi_u = n_t/N_t$ for all u, but $\pi_{uv} \neq [n(n\text{-}1)]/[N(N\text{-}1)]$ for all uv (e.g., systematic sampling, GRTS sampling (Stevens and Olsen, 1999)), $\hat{\bar{y}}_t \pm 1.96\sqrt{\widehat{\text{var}}(\hat{\bar{y}}_t)}$ remains a reasonably accurate approximate confidence interval. Coverage of this confidence interval drops substantially below 95% when variance of the π_u's increase (McDonald, 1996). Stevens and Olsen (2002) provide a general purpose variance estimator based on spatial neighborhoods for this case, but properties of the confidence interval that uses this newer variance estimate are unstudied.

Assuming independent samples at times t and t-1, an estimate of the net change in average response is $\hat{\bar{y}}_t - \hat{\bar{y}}_{t-1}$ with estimated variance $\widehat{\text{var}}(\hat{\bar{y}}_t) + \widehat{\text{var}}(\hat{\bar{y}}_{t-1})$. If independent status estimates $\hat{\bar{y}}_1, \hat{\bar{y}}_2, \ldots, \hat{\bar{y}}_t$ are available and trends can be assumed linear, weighted least squares regression (Neter et al., 1985), with weights equal to $1/\widehat{\text{var}}(\hat{\bar{y}}_i)$, can be used to estimate linear trend. Under the same conditions as the previous paragraph (i.e., simple random or systematic sampling), $\hat{\bar{y}}_t$ is approximately normally distributed, and the significance of least squares trend estimates can be assessed using normal-theory regression. If variance of the π_u is large, or if the approximate normality of $\hat{\bar{y}}_t$ is otherwise in question, it may be possible to use randomization or bootstrap methods (see Manly, 1997) to assess significance of trend estimates. When regression is used to compute slope among design-based status estimates, we choose to classify the analysis as design-based even though we acknowledge a linear trend has been assumed.

For our example DRP data, the Horvitz-Thompson estimates of mean DRP were the simple sample averages at each occasion because river reaches were included in the survey with equal probability. The Horvitz-Thompson estimates of mean DRP appear in Table 2 (column $\bar{Y}_t^{HT}$) and Figure 3 alongside the MVLUE estimates described in the next section.

3.1.2 MVLUE Estimation

Although repeated estimation of status via the Horvitz-Thompson estimator is statistically valid and relatively easy to do, the above analysis does not make efficient use of data in all cases. To improve efficiency, yet retain many desirable design-based characteristics, the MVLUE estimator summarized in Binder (1988) can be used. The MVLUE estimator uses the correlation (if present) between current measurements and past measurements to enhance current estimates. The MVLUE analysis is slightly less design-based than the Horvitz-Thompson analysis, but the MVLUE is not entirely parametric under our definition of parametric. The MVLUE is slightly less design-based because it assumes that measurement error is present and that the correlation between responses taken on the same unit at 2 points in time is either known or has a particular form (Agarwal and Tikkiwal, 1975; Link and Doherty, 2002). However, the MVLUE does not meet our definition of parametric either because form of the trend is not assumed and the distribution of random measurement errors is unspecified, except that it is required to have a 0 mean.

Assuming that sampling units enter one of the study's sampling panels (McDonald, 2003) with uniform probability (i.e., equi-probable selection) and that the correlation between responses at two sampling occasions, labeled $\rho_{t,t-1}$, is known, the general form of the minimum variance linear unbiased estimator (i.e., MVLUE) for the true population mean at time *t* is

$$\hat{\bar{y}}_t = \psi_t \bar{y}_{tu} + (1-\psi_t)[\bar{y}_{tm} + \beta_t(\hat{\bar{y}}_{t-1} - \bar{y}_{tm-1,})]$$

where $\bar{y}_{tu}$ is the average response on units sampled at time *t* but not at time *t*-1 (so-called "unmatched" units), $\bar{y}_{tm}$ is the average response measured at time *t* on units sampled at time *t* and at time *t*-1 (so-called "matched" units), $\bar{y}_{tm-1,}$ is the average response measured at time *t*-1 on the same "matched" units, $\hat{\bar{y}}_{t-1}$ is the MVLUE estimate for the previous time period, and

$$\beta_t = \rho_{t,t-1}\frac{\sigma_t}{\sigma_{t-1}},$$

$$\frac{\psi_t}{1-\psi_t} = \frac{n_{tu}}{n_{tm}}(1-\rho^2_{t,t-1}) + \rho^2_{t,t-1}\frac{\psi_{t-1}}{1-\psi_{t-1}},$$

(Binder and Hidiroglou, 1988, eqn. 3.16). Sample size n_{tu} is the number of "unmatched" sample units at time *t*, n_{tm} is the number of "matched" sample units at time *t*, $n_{t-1,u}$ is the number of sample units at time *t*-1 that were not sampled at time *t*-2, and σ_t is the standard deviation of all responses in population at time *t*. For the first sampling occasion, letting ψ_1 = 1 will compute $\hat{\bar{y}}_1$ as the mean of all units sampled at time 1. The correlation $\rho_{t,t-1}$ will generally have to be estimated from strings of repeatedly sampled units using standard procedures like Pearson's or Spearman's correlation coefficient. With sufficient data, it is also possible to estimate $\rho_{t,t-1}$ using spatial variogram methods applied to 1-dimension (time) (Isaaks and Srivastava, 1989; Cressie, 1993) or time series auto-correlation functions (Pankratz, 1983). The variance of responses at each occasion will either need to be estimated from sample data using $(n_t - 1)^{-1}\sum_{u \in s_t}(y_{ut} - \hat{\bar{y}}_t)^2$, or assumed to be constant through time, in which case all σ_t's cancel.

The corresponding estimate of change from occasion *t*-1 to occasion *t* is,

$$\Delta\hat{\bar{y}}_t = \hat{\bar{y}}_t - \hat{\bar{y}}_{t-1}$$

which theoretically should be more efficient than the difference in Horvitz-Thompson estimates. An overall estimate of average change can be computed by averaging values of $\Delta\hat{\bar{y}}_t$ over time.

Estimation of the variance of $\hat{\bar{y}}_t$ and $\Delta\hat{\bar{y}}_t$ is difficult due to the need to know or estimate $\rho_{t,t-1}$. In fact, because of the need to estimate $\rho_{t,t-1}$ we did not find a closed-form variance estimator for $\hat{\bar{y}}_t$ that we felt comfortable reporting. However, provided all units were included in the sample with equal probability, it is possible to bootstrap re-sample the observed values in the same way that the original sample was drawn, and repeatedly calculate both $\hat{\bar{y}}_t$ and $\Delta\hat{\bar{y}}_t$. Variation in these repeatedly re-calculated values will provide a reasonable estimate of the variance of the original statistics.

MVLUE estimates for our example DRP data assumed the standard deviation of responses was constant through time, and consequently we estimated $\beta_t = \rho_{t,t-1}$. Correlation in DRP through time, i.e., $\rho_{t,t-1}$, was estimated by the Pearson correlation coefficient among values on matched units at time t and time $t - 1$. Temporal correlation in DRP values on different streams was high and ranged from 0.94 to 0.97 in our data. Intermediate computations necessary for the MVLUE, and the MVLUE estimates themselves appear in Table 2 and Figure 3. Average annual $\Delta\hat{\bar{y}}_t$ for the MVLUE was 0.305, while average annual $\Delta\hat{\bar{y}}_t$ estimated using Horvitz-Thompson estimates was slightly less at 0.292.

3.2 Parametric Analyses

By our definition, an analysis is parametric if it includes an assumption about the form of trend (e.g., linear trend, quadratic, exponential, etc.) or an assumption about the distribution of measurement errors, or both. We consider the family of parametric analyses to contain what others have called model-assisted (Sarndal et al., 1992) and model-based because analyses in this family rely partially or wholly on assumed models to provide a basis for replication and inference.

There are two main advantages of parametric analysis over design-based analyses, but those advantages come at a cost of reduced objectivity because (usually) un-testable assumptions about the system are made. The first advantage of parametric analyses is that assumptions inherent in the model add information to the system and can improve accuracy and precision of inferences to unsampled units, *provided the assumptions are correct*. Second, parametric analyses are extremely flexible and can generally be attempted on data collected in almost any manner (Olsen et al., 1998); however, this additional power and flexibility must be judiciously and carefully used. Edwards (1998) relates some famous miscalculations based on models applied to judgment or haphazard samples, and we wholeheartedly concur with his warnings against use of these unscientific sampling plans. A less important advantage of parametric analyses is that they closely parallel many infinite population procedures that can be performed using existing software programs.

In the family of parametric analyses, we include what we term unit slope analyses, as well as the simple linear, generalized linear, and mixed effects linear models reported by Smith and Rose (1991), Urquhart et al. (1998), Van Leuwnen et al. (1996), Van Stien et al. (1997), Lesica and Steele (1996), Urquhart and Kincaid (1999), Ver Hoef (2002), and Piepho and Ogutu (2002). We also include time series analysis in this family (Binder and Hidiroglou, 1988). We would include Baysian analyses in this class as well.

From this large class of analyses, we chose to illustrate just two general purpose procedures. The first we call a *unit slope* analysis (“unit” refers to sampling unit, not 1.0) because they estimate a slope statistic through time for every sampled unit. Collapsing the time series of observations collected on sampled units into a single number, and assuming sample units are independent, simplifies analysis by eliminating the need to consider temporal correlation in responses. The second analysis presented in this section is the mixed linear model analysis described by Piepho and Ogutu (2002), which is closely related to the analysis of Van Leeuwen et al. (1996). These linear models view the repeated measurement of responses at a site as a short time series and (potentially) include error structures to account for both spatial and temporal correlation. Other model-assisted analyses worth citing include the significance tests of Hirsch (1982), Elshaarawi (1992), Elshaarawi (1993), and Elshaarawi (1995). The methods of Elshaarawi (1995) assume measurement errors are normally distributed and derive both a parametric score and non-parametric test.

3.2.1 Unit Slope Analysis

The unit slope analysis assumes that trends in responses on individual sampling units have a particular form, they define a trend statistic computable on data from individual units that reflects the magnitude of those trends, and then summarize the distribution of that trend statistic across population units. This type of approach was described in Stevens and Olsen (1999) for analysis of trends under split panel designs with a [1-0,x-y] (McDonald, 2003) revisit structure, but the approach is in fact more general and is useful under more general revisit plans. While Stevens and Olsen (1999) suggested multi-phase regression (Sarndal et al., 1992) to compute a linear slope statistic for all units sampled at least twice, multi-phase regression does not need to be applied even though its use will probably improve precision when responses are cyclical or highly non-linear. We note that the trend statistic used in unit slope analyses can be practically anything, including the least-squares linear regression estimate of slope, the simple difference between time 1 and time *t* responses, or the Sen non-parametric slope estimate (Sen, 1968).

Assuming a sampled unit u has been visited at least twice, let β_u represent the trend statistic of interest computed using all data collected from u. Under many revisit designs, units will have different revisit frequencies or be visited during different sampling occasions, and the ability of β_u to accurately quantify trend for units with missed or differing sampling occasions needs to be considered. For example, if underlying trends are truly linear on all units, then least-squares linear β_u can be computed on all units with ≥2 visits regardless of when the revisits occurred. In this case, β_u is unbiased for the true linear slope on unit u even though the precision of β_u would be improved with more visits. The appropriateness and comparability of β_u across units when units are measured during different occasions depends on the particular form of β_u. For example, if responses are known or suspected to be cyclical, and if unit u was not visited during a period when responses were typically high but unit v was visited, β_u will probably not be comparable to β_v assuming both statistics are least-squares linear regression slopes. In this case, the multi-phase regression suggested by Stevens and Olsen (1999) attempts to use patterns observed on unit v and any others visited during the same periods to "correct" β_u so as to be more comparable to β_v. Ultimately, the appropriateness of β_u given the revisit plan at unit u, as well as the comparability of β_u and β_v given differing visitation schedules to u and v are judgments to be made by individual researchers.

To illustrate the method, we assume changes are linear and describe the straightforward unit slope analysis with the understanding that β_u may be calculated with different numbers of data points for every unit. In this case, least squares regression is used to estimate the coefficients α_u and β_u in the regression $\mathrm{E}[y_u] = \alpha_u + \beta_u t_u$, where y_u is the response measured on unit u at time t_u. Once slopes are estimated, the analysis summarizes the distribution of β_u across sampled units and makes inference to an underlying population of slopes. One reasonable way to summarize the distribution of trend statistics is to report the distribution's mean with accompanying confidence interval. It may also be reasonable to report the median and its confidence interval, or plot the distribution's histogram and cumulative frequency distribution. From these summaries, statements can be made about whether the average or median slope is significantly different from 0, as well as the proportion of units in the population experiencing trends greater (or less) than a specific some value.

Statements about the central tendency of the distribution of slope statistics can be made by testing the hypothesis H_0: $E[\beta_u / \pi_u] = 0$ versus the alternative H_A: $E[\beta_u / \pi_u] \neq 0$, where expected values are taken over all possible samples that could have been generated by the implemented design. Such testing is greatly simplified if all units in the population were sampled with the same probability (i.e., design was equi-probable) and if responses from different units can be considered independent. In fact, if the sample design is not equi-probable and inclusion probabilities π_u vary substantially among units, it is unclear what testing or confidence interval method should be implemented because the distribution of β_u / π_u is unknown and can be markedly non-normal. If an equi-probable design was used to sample units, π_u can be dropped from H_0 above and any appropriate 1-sample size α test or confidence interval can be implemented. If units are uncorrelated across space and time, a regular 1-sample *t* statistic or bootstrap method (Mace et al., 1996) can be used. If units close together in space give more similar responses than units far apart, or in other words if responses from different units are spatially or temporally correlated, slope estimates from different units will be correlated to some extent and a spatially adjusted 1-sample *t* statistic or block bootstrap method (Lahiri, 2003) can be used. Below, we conduct a regular 1-sample *t* test, a bootstrap *t* test, and a spatially adjusted *t* test of the 0-mean slope hypothesis.

We computed linear least-squares β_u's for all units in our example data set (Figure 4). The overall 1-sample *t*-ratio was defined as,

$$t_{obs} = \frac{\bar{\beta}}{se(\bar{\beta})}$$

where

$$\bar{\beta} = \frac{\sum_{u \in S} \beta_u}{n},$$

$$se(\bar{\beta}) = \sqrt{\frac{\sum_{u \in S} (\beta_u - \bar{\beta})^2}{n(n-1)}}$$

and S was the sample of units visited at least twice and n was the size of S (i.e, $n = |S|$). For the example DRP data, $n = 16$, $\bar{\beta} = 0.2888$, $se(\bar{\beta}) = 0.1098$, and $t_{obs} = 2.6301$. Assuming normality of unit slopes, Student's t distribution yields a significance of p=0.0189.

In many surveys, estimated slope statistics will be approximately normally distributed and the 1-sample *t* test will be adequate; however, in cases where the estimated slope statistics are not normally distributed, a bootstrap *t* test is more appropriate. The histogram of slopes in Figure 4 suggests the example data set might be such a case. Following Manly (1997, Section 3.10) and assuming sampled stream segments were independent, the 1-sample bootstrap procedure used to test H_0: $E[\beta_u] = 0$ first replaced the original slope observations with the deviations

$$\beta_u^* = \beta_u - \bar{\beta}.$$

This was done to assure that the null hypothesis was true for the collection of resampled (i.e.,β_u^*) values. To establish the distribution of t_{obs} under the null hypothesis, 4999 bootstrap samples of β_u^* values, each of size $n = 16$, were randomly drawn with replacement. For the i^{th} bootstrap sample, the bootstrap t-ratio was computed as,

$$t_i = \frac{\bar{\beta}^*}{se(\bar{\beta}^*)}$$

where

$$\bar{\beta}^* = \frac{\sum_{u \in S_i^*} \beta_u^*}{n},$$

$$se(\bar{\beta}^*) = \sqrt{\frac{\sum_{u \in S_i^*} (\beta_u^* - \bar{\beta}^*)^2}{n(n-1)}}$$

and S_i^* was the collection of β_u^* values in the i^{th} bootstrap sample. The 2-sided significance of t_{obs} was computed as,

$$p = \frac{1 + \sum_{i=1}^{4999} I(|t_i| \geq |t_{obs}|)}{5000}$$

where $I(|t_i| \geq |t_{obs}|)$ equaled 1 if $|t_i| \geq |t_{obs}|$ and 0 otherwise. In practice, n (and n-1 for that matter) in the denominator of $se(\bar{\beta})$ and $se(\bar{\beta}^*)$ can be dropped without affecting the computed p-value because n is a constant across all bootstrap samples (Manly, 1997, p. 64).

For our example data, the absolute value of t_i equaled or exceeded the absolute value of t_{obs} 412 times out of 4999, resulting in $p = 0.0826$. Assuming $\alpha = 0.05$, the bootstrap test arrived at a different conclusion (i.e., "fail to reject H_0") than the 1-sample t (i.e., "reject H_0"). In this case, disagreement between the tests was due to the limited number of streams (16) and marked non-normality caused by the large slope of 1.7 observed on stream 5. In practice, the 1-sample and bootstrap t tests will agree in most cases, and we recommend more inferential weight be given to the bootstrap test if normality of the slope statistic distribution is even slightly questionable, as here.

Often responses, and thus slope statistics, are spatially correlated, the regular 1-sample t and bootstrap tests are not appropriate because they assume independence of units. Two general methods for dealing with spatial correlation are the block bootstrap (Lahiri, 2003) and a spatial adjustment to a regular 1-sample *t* test. Both routines rely on estimation of a spatial variogram or correlogram and corresponding estimated variance function (Isaaks and Srivastava, 1989; Cressie, 1993). Block bootstrapping utilizes the range estimate from the variance function to estimate the maximum distance at which units are correlated. The algorithm then sets this distance as the "block size" and randomly resamples blocks of this size instead of individual units. Remaining steps of the algorithm are identical to the regular bootstrap. That is, 1-sample *t* statistics are computed on a large number of (block) bootstrap samples, and significance is the proportion of bootstrap *t* statistics greater in absolute value than the observed *t*. Resampling spatial blocks containing >1 unit effectively reduces sample size to the number of blocks.

Another alternative for cases with non-negligible spatial correlation is a spatially adjusted *t* test that involves estimating a covariance matrix from the variance function based on separation distances, and performing weighted least squares regression. In our example DRP data, river stretches were separated by large distances and resided in different drainages, leading us to suspect that spatial correlation was low; nonetheless, we conducted the spatially adjusted *t* test for illustration purposes. Like a regular *t* test, the spatially adjusted *t* test assumed individual slope statistics were normally distributed. If slope statistics are markedly non-normal and spatial correlation is present, the block bootstrap method mentioned in the previous paragraph should be used.

The spatially adjusted *t* test of the hypothesis H_0: $E[\beta_u] = 0$ estimated a spatial covariance matrix among units and used this estimate to adjust a regular 1-sample *t* test. The spatial covariance matrix was estimated by fitting a smoothed variance function to the observed spatial correlogram (Isaaks and Srivastava, 1989, Chapter 16). When fit to a correlogram, the variance function estimates correlation between units as a function of separation distance. Using example data, we fit the spherical covariance model

$$\rho(h) = \begin{cases} 1 - \frac{2}{\pi}\left(\frac{h}{\theta}\sqrt{1-\frac{h^2}{\theta^2}} + \sin^{-1}\frac{h}{\theta}\right) & \text{if } h \le \theta \\ 0 & \text{if } h > \theta \end{cases}$$

to observed slope statistics, where *h* was the straight-line 2-D distance between river reaches and θ was the (estimated) parameter that determined the range of distances over which observations were correlated. Parameter θ were estimated by least squares (see Isaaks and Srivastava, 1989), and the estimated covariance matrix ***V*** was computed from this estimated value and observed separation distances. In our example, the estimated spherical correlation model (Figure 5) indeed showed very little correlation (<0.05) among all units regardless of spatial separation. With very little correlation, the estimated correlation matrix was nearly equal to the identity matrix, and we would not expect much adjustment to the usual 1-sample *t* test.

Using the estimated ***V***, the estimate of mean slope was adjusted as

$$\begin{aligned} \bar{\beta}_s &= (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{\beta} \\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i v_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}} \end{aligned}$$

where, in the 1-sample case, ***X*** is a vector of 1's, $\boldsymbol{\beta}$ is a vector containing the estimated slope statistics for all units, and v_{ij} is the *ij*th element of the inverse of ***V***. It is clear from this equation that the adjusted estimate of $\bar{\beta}_s$ is a weighted average of the original slope statistics, where the weights are row (or column, because ***V*** symmetric) sums of the inverse of the estimated covariance matrix. The adjusted standard error of $\bar{\beta}_s$ was computed as

$$se(\bar{\beta}_s) = \sqrt{\hat{\sigma}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}} = \sqrt{\frac{\sum_{i=1}^{n}(\beta_i - \bar{\beta}_s)^2}{r_{df}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}}}$$

where r_{df} was residual degrees of freedom equal to the rank of $\boldsymbol{V}$.

In our example, the adjusted mean slope was $\bar{\beta}_s$ = 0.2888 with adjusted standard error $se(\bar{\beta}_s)$ = 0.1098, which are identical to their unadjusted counterparts to 4 decimal places. The test statistic $t = \bar{\beta}_s / se(\bar{\beta}_s)$ is at least approximately distributed the same as a deviate from the Student's t distribution with r_{df} degrees of freedom. The 2-tailed spatially adjusted p-value associated with t = 0.2888 / 0.1098 = 2.63 and r_{df} = 15 degrees of freedom p = 0.0189, indicating that the average slope on units in the population was significantly >0 at the α = 0.05 level.

3.2.2 Mixed Linear Models

Mixed linear model trend analyses, in general, are more complicated than unit slope analyses, but offer a high degree of flexibility and power if assumptions are satisfied. While most trend analyses involve some type of explicit or implicit linear model, we say a linear model is designed to detect trend if it involves random or mixed-effects that mimic classic repeated measures ANOVA techniques (Milliken and Johnson, 1984; Littell et al., 1996). These repeated-measures-like linear models view the sampled responses as short correlated time-series existing at each unit.

One desirable characteristic of mixed linear model trend analyses approach is their flexibility and the fact that they can be carried out on a wide variety of revisit designs. Missing values and unequal number of revisits to different units do not pose a problem for analysis assuming that the overall realized pattern of revisits is connected in the experimental design sense. Heuristically, being connected in the experimental design sense means that the model can estimate or infer a mean for every cell (here, cell = a sampling occasion for a single sample unit) in the design, whether there is data in the cell or not. For example, in order for the revisit plan to be experimentally connected, it is sufficient, but not necessary, for at least 1 unit to be visited every sampling occasion. A different revisit plan in which responses on every pair of sample units are measured together at least once is also connected. When the revisit plan is connected, all main effects of the linear model are estimable, but residual degrees of freedom are reduced over what they would have been if every unit had been visited every occasion.

Among the linear model trend analyses we found in the literature, we choose to describe and illustrate the analysis of Piepho and Ogutu (2002) because of its generality. Many other useful models are special cases of the Pielpho and Ogutu model. The linear model of Piepho and Ogutu is,

$$y_{ij} = \mu + w_j\beta + b_j + a_i + w_j t_i + c_{ij}$$

where y_{ij} is the response measured on unit *i* during sampling occasion *j*, μ is the mean of responses on all units over all time periods, w_j is a fixed covariate specifying the sample occasion (e.g., w_j = 1, 2, 3, …, or w_j = 1996, 1997, 1998, …), b_j is the random effect of the *j*th year on responses from all units, a_i is the random effect on the intercept of the *i*th sample unit, t_i is the random effect on the slope of the *i*th sample unit, and c_{ij} is the random effect of the *i*th sample unit during the *j*th occasion on the measured response. A picture of this model and its effects is given in Figure 6. This particular model assumes that responses were measured only once during a sampling occasion. If this is not the case, the model can be extended to include a random effect for multiple measurements on the same unit during the same occasion. Otherwise, the c_{ij} effects can be viewed as residual errors and used to construct an estimate of residual variation. Piepho and Ogutu (2002) assumed that all random effects were normally distributed with means of 0 and unknown variances.

Depending on the application, spatial or temporal correlation among the random (and fixed) effects in the linear model can be modeled in a variety of ways. When correlation among effects is modeled, the assumption of independence can be relaxed and p-values will be adjusted accordingly. To relax the normality assumption on residuals (i.e., on c_{ij}), regular or blocked randomization, or regular or blocked bootstrapping, could be employed; however, these re-sampling techniques are generally difficult to properly implement for large models involving many effects or in models containing either spatially or temporally correlated effects. The primary difficulty is ensuring that the null hypothesis is true during re-sampling, but this can be accomplished in many cases.

The objective of the analysis is to estimate the fixed overall slope, β, and assess the strength of evidence for H_0: $\beta = 0$ versus H_A: $\beta \neq 0$. To do this, Piepho and Ogutu (2002) propose both a “inter-site” (random site effect) and “intra-site” (fixed site effect) analysis; however, their subsequent simulations showed little practical difference between the two, except that the fixed “intra-site” analysis converged more often. Estimation can be carried out using SAS Proc Mixed (Littell et al., 1996; Piepho and Ogutu, 2002), or similar programs which use either restricted-maximum-likelihood (REML) or maximum-likelihood (ML) techniques. SAS Proc Mixed code for the “intra-site” and “inter-site” analyses is shown in the Appendix. Generalized linear mixed models (so-called GLIMIX models) can be used when the normality of responses is in question, although the regular mixed model is known to be reasonably robust to violations of normality.

Applied to our example data, the “intra-site” mixed linear model produced a slope estimate of $\hat{\beta}$ = 0.2914, with estimated $se(\hat{\beta}$ = 0.1249, and significance for testing the hypothesis H_0: $\beta = 0$ of p = 0.0349. The “inter-site” mixed linear model produced a slope estimate of $\hat{\beta}$ = 0.2828, with estimated $se(\hat{\beta})$ = 0.1255, and significance of p = 0.0404. Both these tests support the hypothesis of positive trend in DRP at the α = 0.05 level during the study period.

3.3 Nonparametric Analyses

Besides design-based and parametric analyses, the third major class of trend analysis we found in the literature was nonparametric analysis. We classify an analysis as nonparametric if it makes no assumptions about the *form* of trend in the population, or about the distribution of responses. Recall that design-based analyses also do not make assumptions about trend form or distribution of responses. The difference between design-based and nonparametric analyses is that design-based analyses summarize data across units first then look for consistent changes in those summaries through time, while nonparametric analyses summarize changes through time on individual units then look for consistent patterns across units. In this regard, the parametric analyses, and the unit slope analysis in particular, operates the same way as nonparametric analyses, except that the parametric analyses assume trends have a particular form or that responses follow a particular statistical distribution.

Among nonparametric trend analyses we encountered, most were derivatives or modifications of the well-known Mann-Kendal test (Manly, 2001). Here, we present the standard Mann-Kendal test unadjusted for seasonality and apply it to our example data set. In addition to Mann-Kendal, we present the CUSUM procedure (Manly, 2001, Sections 5.7 and 5.8) because we believe it is powerful and has general utility in a large number of applications.

3.3.1 Mann-Kendall Test

Assuming response y_{ij} was measured on unit i during occasion j¸ the Mann-Kendal test statistic for a particular site is,

$$C_i = \sum_{j=2}^{n_i}\sum_{k=1}^{j-1} sign(y_{ij} - y_{ik})$$

where n_i is the number of observations at site *i,* and the function *sign*(*x)* equals 1 if $x > 0$, 0 if $x = 0$, and -1 if $x < 0$. If interest lies in testing for trend at a particular site, and the time series of measurements at the site is short (generally <20), the significance of C_i can be assessed using tables (Hollander and Wolfe, 1999, Table A.21). If the time series at unit *i* is long (e.g., $n_i > 15$), the variance of C_i is approximately,

$$\text{var}(C_i) = n_i(n_i - 1)(2n_i + 5)/18 \quad ,$$

assuming the null hypothesis of no trend at site *i* is true. For $n_i > 15$, the distribution of

$$z_i = \frac{C_i}{\sqrt{\text{var}(C_i)}}$$

is well approximated by a standard normal. Significance levels of C_i in this case can be computed using software packages that contain routines for integrating under the standard normal distribution (e.g., Excel, R, SAS, etc.).

When multiple sites are involved, and interest lies in inference toward overall average trend in the population, one possibility is to test whether the mean Mann-Kendal C_i statistic is equal to 0. In this case,

$$\bar{C} = \frac{\sum_{i=1}^{n} C_i}{n} \quad .$$

If responses from different units are independent,

$$\text{var}(\bar{C}) = n^{-2}\sum_{i=1}^{n} \text{var}(C_i) \quad ,$$

and $z = \bar{C} / \sqrt{\mathrm{var}(\bar{C})}$ follows a standard normal distribution. If units are not independent, blocked bootstrapping or a spatially adjusted *z* test with estimated covariance matrix could be used.

Dropping the *i* subscript from y_{ij} to simplify notation, the Mann-Kendall statistic for river reach 1 was $C_1 = sign(y_{1991} – y_{1989}) + sign(y_{1993} – y_{1989}) + …+ sign(y_{1997} – y_{1995}) = sign(0.7 – 0.8) + sign(0.5 – 0.8) + … + sign(0.5 – 0.7)$ = –1 + –1 + –1 + –1 + 0 + 1 + –1 + –1 + 0 + –1 = -6. From Hollander and Wolfe (Hollander and Wolfe, 1999, Table A.21), the probability of a *C* statistic this large or larger in absolute value under the null hypothesis of no trend was $p = P(C \geq |C_1|) = 0.234$. The variance of C_1 under the null hypothesis of no trend was $\mathrm{var}(C_1) = 5(5 – 1)(2(5) + 5) / 18 = 16.67$, and the probability of observing a standard normal *Z* greater in absolute value than $z = -6 / 4.0825 = -1.47$ was 0.1416. The rest of the unit-specific Mann-Kendall statistics appear in Table 3.

The overall average Mann-Kendall statistic for our example data was $\bar{C}$ = 3.125 (Table 3) with estimated variance var($\bar{C}$) = 0.8984 and z = 3.48. The probability of observing a standard normal random variable more extreme than 3.48 is 0.001. This test again indicated a significant positive overall trend in DLP during the example study period.

3.3.2 CUSUM

Manly (1994), Manly and MacKenzie (2000; 2003), and Manly(2001, Sections 5.7 and 5.8) discuss use of control charts and CUSUM charts for detecting changes in a population mean. Control charts, which set upper and lower “warning” and “action” limits for the sample mean, have a long history in industrial process monitoring (Montgomery, 1991). We present the modification of standard CUSUM charts discussed by Manly (1994) and Manly and MacKenzie (2000)

Assume temporarily that *m* successive measurements are made on each of *n* units. Let $\bar{y}_{i} = m^{-1} \sum_t y_{it}$ be the average of all measurements on unit *i* through time. Further, order the units under consideration based on their averages so that $\bar{y}_1 \le \bar{y}_2 \le \ldots \le \bar{y}_n$. For each time *t*, the CUSUM method computes a series of *n* partial sums of differences of the form,

$$S_{it}=\sum_{k=1}^{i}(y_{kt}-\bar{y}_k),$$

for i = 1, 2, …, n. Collectively, the *n* partial sums S_{it} are called the CUSUM at time *t*, and when plotted against *i* for each *t* yield *m* CUSUM charts. The 5 CUSUM charts for our example data appear in Figure 7. Positive slope in the CUSUM (the –•– line in Figure 7) indicates that responses during that time period were, on average, higher than their long-term averages. Negative CUSUM slope indicates the opposite. A positive sloping then negative sloping CUSUM indicates that units with low long-term averages were higher than their respective means, while units with large long-term averages were lower than their respective means.

Assuming independence, Manly *et al*. (2000) use a single randomization procedure to perform 3 tasks related to the CUSUM. Randomization can be used to test whether a particular CUSUM for time *t* significantly departures from 0, or to test whether the *m* CUSUM charts as a group show significant overall trend through time, or to compute a 95% confidence envelope for the CUSUM at time *t* under the null hypothesis of no departure from the long-term mean at time *t*. The randomization method proposed for independent data permutes available observations on each unit through time a large number of times. Randomization in this manner is justified by noting that under the null hypothesis of no trend in the population mean, any of the *m* values observed at a site were equally likely to have occurred at any sampling occasion. The randomization effectively eliminates any trend in the data from a particular site, and assesses variation in the CUSUM under the null hypothesis. The CUSUM method is effected by serial correlation in the measurements at a given site; however, the randomization test can be adjusted to account for this (see Manly et al. (2000) and Manly et al. (2001)).

A 95% confidence envelope for the CUSUM at time *t* can be computed by discarding 2.5% of the smallest S_{it} and 2.5% of the largest S_{it} across randomizations and fixing the endpoints at the extremes of the remaining values. Specifically, if *R* random permutation iterations are performed, *R* values of S_{it} computed under the null hypothesis are available and the 95% confidence envelope for S_{it} has lower and upper endpoints at the 2.5th and 97.5th percentiles of these *R* values. If the observed value of S_{it} is outside this 95% confidence envelope there is significant departure from the long-term mean somewhere among units 1, 2, …, *i*.

To test the hypothesis of no difference between the population mean at time *t* and the long-term mean, an overall significance of the observed CUSUM at time *t* can be computed using the statistics,

$$Z_{it} = \frac{S_{it}}{\sqrt{\mathrm{var}(S_{it})}} \quad ,$$

which are computed each randomization for each value of *i* (i.e., each unit). The variance of S_{it} in the denominator of Z_{it} is,

$$\mathrm{var}(S_{it}) = \sum_{k=1}^{i} s_k^2$$

where

$$s_k^2 = \frac{\sum_{t=1}^{m} (y_{tk} - \bar{y}_k)^2}{m-1} \quad .$$

Note that var(S_{it}) is constant through time (i.e., varies across sites only), and because units maintain their order after permutation, var(S_{it}) can be computed using the original (un-permuted) data. A measure of the overall deviation of the *t*th CUSUM from 0 is

$$Z_{max,t} = \max(|Z_{1t}|, |Z_{2t}|, \ldots, |Z_{mt}|).$$

The significance of the t^{th} CUSUM is the proportion of $Z_{max,t}$ values obtained from permuted data that are larger than the $Z_{max,t}$ observed in the original (un-permuted) data.

A measure of the extent to which all m CUSUM charts collectively depart from the null hypothesis of no trend is the statistic,

$$Z_{max,T} = Z_{max,1} + Z_{max,2} + \ldots + Z_{max,m}$$

If there is no trend in the average population response through time, the value of $Z_{max,T}$ will be a "typical" value among those obtained by randomization. The significance level of the collective set of m CUSUMs is the proportion of randomized $Z_{max,T}$ that are larger than the observed $Z_{max,T}$.

Missing values, or non-sampled occasions, do not cause difficulties other than to add the subscript i to m (as in m_i) to reflect the fact that sample size varies across units. When data from some occasions are missing, the long-term means for each unit are calculated from the available data, and the randomization scheme permutes the available data among the occasions that were observed. That is, randomizations leave missing value in their respective locations and permute the remaining (non-missing) data. For example, the vector of data from reach 6 of the example data set is [*NA*, 0.8, 0.5, *NA*, 0.8], where *NA* represents a missing value. The long-term mean for site 6 would be computed as (0.8 + 0.5 + 0.8) / 3 = 0.7, and 2 typical random permutations of site 6's data are [*NA*, 0.5, 0.8, *NA*, 0.8] and [*NA*, 0.8, 0.8, *NA*, 0.5].

The 95% confidence envelopes for the 5 CUSUMs computed on example data (Figure 7) show one value of S_{it} below the confidence envelope in 1989, and 8 of 16 values above the confidence envelope in 1997. This indicates some departure from the long-term mean in 1989, and pretty strong evidence for departure from the long-term mean in 1997. Inspection of the 1997 CUSUM chart shows only 1 negatively sloped segment, indicating that all river segments except 1 produced

DRP responses above their respective long-run averages. The $Z_{max,t}$ test results were similar in that they showed a marginal departure from the long-term mean in 1989 and clear departure from the mean in 1997; in addition however, the $Z_{max,t}$ test for 1993 shows a significant downward shift in the mean that year. The observed $Z_{max,t}$ statistics for the 5 sampled years between 1989 and 1997 were 2.09, 1.47, 2.30, 1.31, and 3.69 respectively. The significance levels associated with each $Z_{max,t}$ statistic were 0.085 for 1989, 0.511 for 1991, 0.044 for 1993, 0.615 for 1995, and 0.001 for 1997. The test for significant trend in the collective set of CUSUM charts had test statistic $Z_{max,T}$ = 10.86 with significance level of 0.001. This final test gives clear evidence of a trend in DRP during the sampled years.

4. Discussion

All tests run on example DRP data are summarized in Table 4. Estimates of average annual change in DRP ranged from 0.2828 to 0.3048 mg / liter / year among those analyses that produced a trend estimate. Among those analyses that produced a significance level, *p*-values for the null hypothesis of no trend ranged from 0.001 produced by the non-parametric analyses to 0.083 for the unit slope analysis that used bootstrapping. These results provide reasonably clear evidence of significant positive trend in DRP during the study period. If this were an article on DLP in the streams of New Zealand, these results would likely be followed up with (bootstrap) confidence intervals on slope estimates, and a more complete description of rivers that seem to be contributing most to the trend.

Given the multitude of tests available for trend detection, it is important to keep in mind that running more tests on the same data set increases the probability of obtaining at least one significant result regardless of the true underlying trend. If, for example, responses were truly not experiencing trend, yet we ran 7 trend detection analyses, the probability of at least 1 significant result is 30% (=1 - $(1-0.05)^7$) when we run each test at a $\alpha = 0.05$ level. It is clearly not reasonable to run multiple trend detection analyses on the same data set, declare trend to be significant when 1 or more of these tests are significant, and expect the overall Type I error rate to remain below the level of the individual tests. When possible, we recommend deciding *a priori* which trend analysis to apply and sticking to it. If assumptions of the analysis selected *a priori* turn out to be violated, and another analysis is chosen, authors are duty bound to report this fact and reasons for the switch. When multiple tests are desired, or unavoidable perhaps due to questionable assumptions in all candidate analyses, we recommend correcting (reducing) the level at which individual tests are performed in order to control the experiment's overall Type I error rate. This is a well known problem in statistical hypothesis testing, and the standard Bonferroni correction in which individual tests are conducted at $\alpha_B = \alpha / k$, where k is the number of tests, is one relatively simple solution.

5. Conclusions

There exist a wide variety of analyses for detection of trend, and undoubtedly more will appear in the near future as large-scale and long-term surveys become more popular. Most analyses we found were specialized to the data at hand and are what we would classify as parametric. Design-based analyses are applied to probability samples and detect trend with a minimum of assumptions. Non-parametric analyses also make few assumptions, but are restricted to equi-probable samples and go about estimation differently than design-based analyses. Design-based analyses generally summarize responses across units each sample occasion, then estimate trends based on changes in those summaries. Nonparametric analyses generally summarize changes in responses through time for each sampled unit, then estimate trend by summarizing those changes across units in the population. Parametric analyses vary wildly in their assumptions and, like nonparametrics, require an equi-probable sampling design. Parametric analyses are powerful and can be applied to almost any data set with enough assumptions. Common sense and the plausibility of assumptions should guide researchers toward one of these classes of analyses. Model assumptions should always be stated and checked because invalid assumptions always introduce at least some bias.

6. Appendix

This appendix contains SAS code to perform both the intra-site and inter-site analyses proposed by Piepho and Ogutu (2002). The inter-site analysis can be performed with the following code:

```
proc mixed data=rivers method=REML nobound;
        class year site;
        model drp1 = wyear /ddfm=satterth s;
        random int/sub=year;
        random int wyear/sub=site type=UN;
```

For analysis of the DRP data, variable ‘wyear’ equaled 1989, 1991, 1993, 1995, and 1997. Variable ‘year’ was identical to ‘wyear’, but was treated as a discrete (class) variable. Alternative covariance structures (rather than the unstructured form) can be fitted by changing the ‘type=’ option on the last line. ‘type=CS’ or ‘type=AR(1)’ are reasonable alternative choices in many cases.

The intra-site analysis can be fitted by the following code:

```
proc mixed data=rivers method=REML nobound;
        class year site;
        model drp1 = wyear /ddfm=satterth s;
        random int wyear/sub=site type=UN;
```

7. Acknowledgments

Majority funding for this project was provided by U. S. Geological Survey Biological Research Division, Alaska Science Center, Anchorage, AK. Minority funding was provided by the Fish, Farm, and Forest Commission (FFFC), San Francisco, CA, U. S. National Marine Fisheries Service, Santa Cruz, CA, and Western EcoSystems Technology, Inc., Cheyenne, WY. The authors wish to thank Ms. Karen Oakley (USGS-BRD) for her encouragement, and Ms. Nadine Wilson (West, Inc.) for her assistance composing the manuscript.

Table 1: Yearly mean dissolved reactive phosphorus (DRP) concentration in 16 river reaches on the south island of New Zealand. X and Y are distances (km) of the river's sample reach east and north from a fixed origin. Missing data designated by NA.

River			DRP (mg/litre)				
Reach	X	Y	1989	1991	1993	1995	1997
1	43	134	0.8	0.7	0.5	0.7	0.5
2	55	97	3.6	1.9	2.7	1.4	4.1
3	106	141	2.6	2.2	2.9	2.6	4.1
4	124	160	4.3	5.2	5.8	5.6	8.6
5	140	68	7.8	11.8	16.4	18.1	21.8
6	153	225	*NA*	0.8	0.5	*NA*	0.8
7	193	147	*NA*	*NA*	0.7	*NA*	0.8
8	220	84	0.8	1.8	2.5	1.9	5.1
9	259	125	6.2	9.5	8.5	8.8	13.1
10	260	193	13.1	14.4	11.8	15	16.8
11	284	255	4.3	4.1	3.2	4.7	2.6
12	310	312	3.1	3.2	3.4	5.6	5.9
13	328	226	1.4	0.9	0.6	2.5	*NA*
14	345	299	3.2	3.3	4.2	4.7	4.4
15	444	462	1.3	1.1	0.7	1	1.1
16	488	451	2.9	2.2	2.3	2.2	4.8

Table 2: Intermediate calculations for the MVLUE and Horvitz-Thompson estimates of average DRP each occasion of the example study. In this case, the Horvitz-Thompson estimates are the sample averages. The MVLUE assumed constant variation through time and estimated correlation with the Pearson correlation coefficient.

											Horvitz-	
									MVLUE		Thompson	
									Estimates		Estimates	
Year	ψ_t	n_{tu}	n_{tm}	$\rho_{t,t-1}$	β_t	$\bar{Y}_{tu}$	$\bar{Y}_{tm}$	$\bar{Y}_{tm1,}$	$\bar{Y}_t$	$\Delta\bar{Y}_t$	$\bar{Y}_t^{HT}$	$\Delta\bar{Y}_t^{HT}$
1989	1	14	0	-	-	3.96	-	-	3.96	-	3.96	
1991	0.0667	1	14	0.9539	0.9539	0.8	4.45	3.96	4.21	0.25	4.21	0.25
1993	0.0625	1	15	0.9429	0.9429	0.7	4.40	4.21	4.17	-0.038	4.17	-0.038
1995	0	0	14	0.9772	0.9772	-	5.34	4.68	4.84	0.68	5.34	1.17
1997	0.0109	2	14	0.9605	0.9605	0.8	7.15	5.56	6.40	1.55	6.30	0.96
Average									4.71	0.305[a]	4.80	0.292[a]

[a] 2-year change was divided by 2 to estimate annual change.

Table 3: Unit-specific and mean Mann-Kendell statistics for the example DRP data contained in Table 1.

Site	C_i	var(C_i)	z_i	P(Z_i>|z_i|)	P(C_i>|c_i|)[a]
1	-6	16.67	-1.47	0.1416	0.234
2	0	16.67	0	1	1
3	5	16.67	1.22	0.2207	0.359
4	8	16.67	1.96	0.05	0.084
5	10	16.67	2.45	0.0143	0.016
6	0	3.67	0	1	1[b]
7	1	1	1	0.3173	1[b]
8	8	16.67	1.96	0.05	0.084
9	6	16.67	1.47	0.1416	0.234
10	6	16.67	1.47	0.1416	0.234
11	-4	16.67	-0.98	0.3272	0.484
12	10	16.67	2.45	0.0143	0.016
13	0	8.67	0	1	1
14	8	16.67	1.96	0.05	0.084
15	-3	16.67	-0.73	0.4624	0.65
16	1	16.67	0.24	0.8065	1
$\bar{C}$:	3.125				
var($\bar{C}$):	0.8986				

[a] From Hollander and Wolfe (1999) Table A.21.

[b] Hollander and Wolfe (1999) Table A.21 reports p for n > 3. This value computed by the authors.

Table 4: Summary of trend analyses applied to the DRP data in Table 1.

Procedure	Trend Estimate	P-value
Horvitz-Thompson	0.2929	na
MVLUE	0.3048	na
Unit slope, *t*	0.2888	0.0189
Unit slope, spatial *t*	0.2888	0.0189
Unit slope, bootstrap *t*	0.2888	0.0826
Linear model: inter-site	0.2828	0.0404
Linear model: intra-site	0.2914	0.0349
Mann-Kendall	na	0.0010
CUSUM	na	0.0010

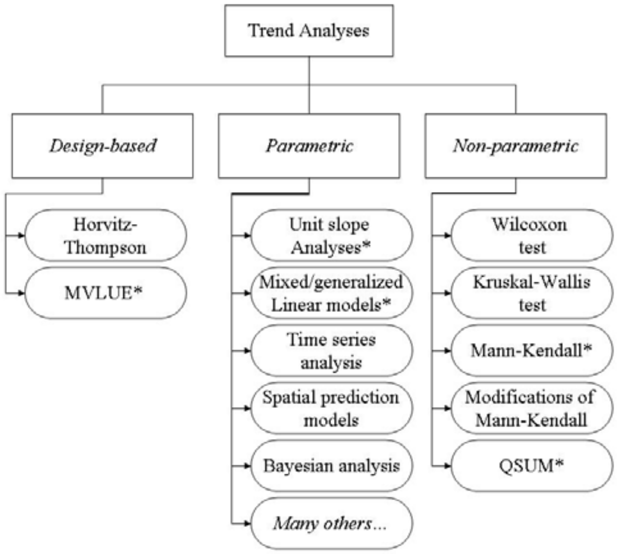

Figure 1: Classes of trend analyses (squares*)* found in the literature, and a few examples of analyses in each class (ovals). Analyses with stars are described in the text and applied to the example DRP data.

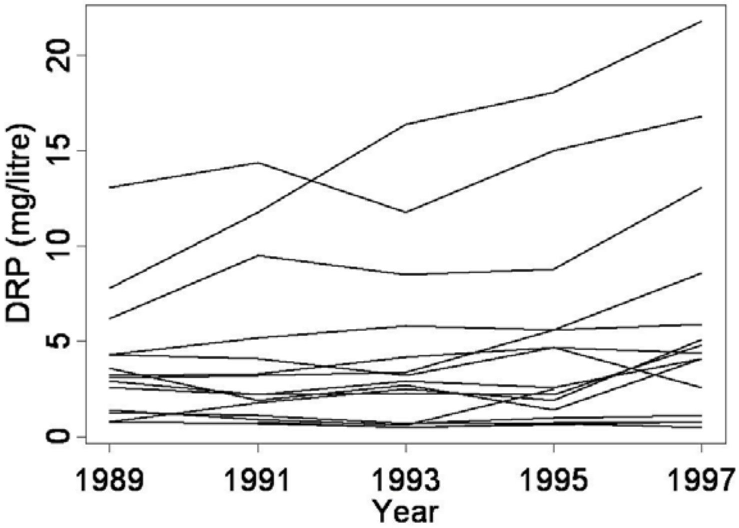

Figure 2: Graph of yearly mean dissolved reactive phosphorus (DRP) values displayed in Table 1. DRP values were collected in 16 river reaches on the south island of New Zealand.

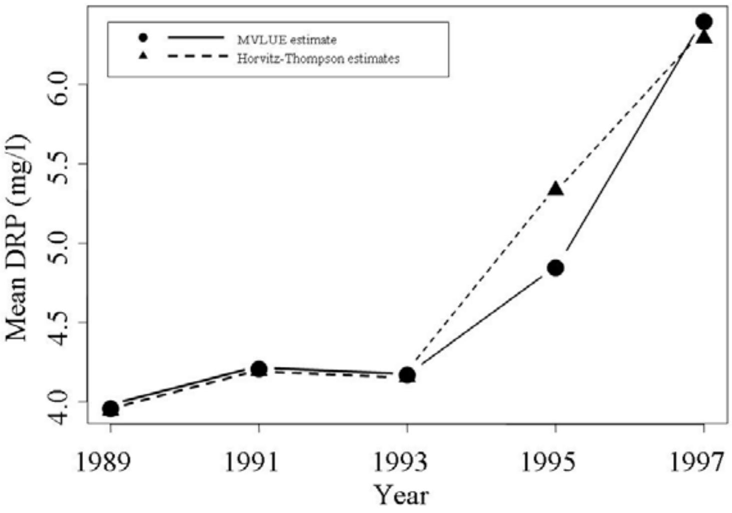

Figure 3: MVLUE and Horvitz-Thompson estimates of mean DRP each year of the example study.

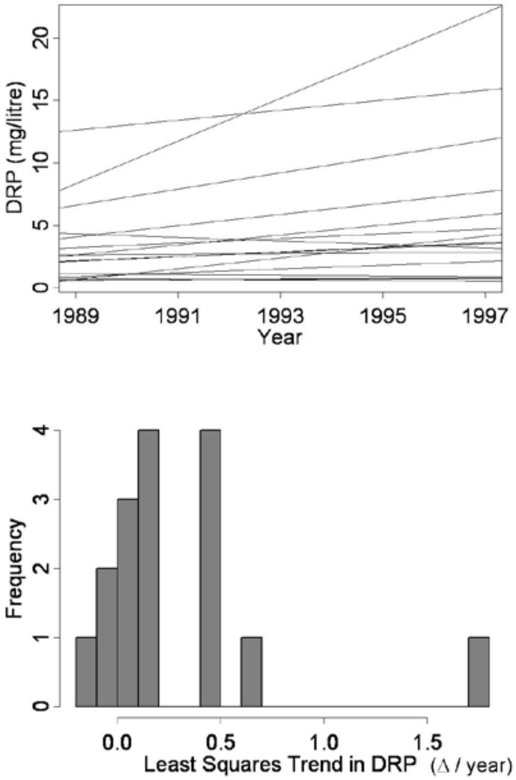

Figure 4: Least square regressions fit to all units in the example data set (top), and a histogram of slope values from all regressions (bottom). The unit slope analysis tested whether mean slope was >0.

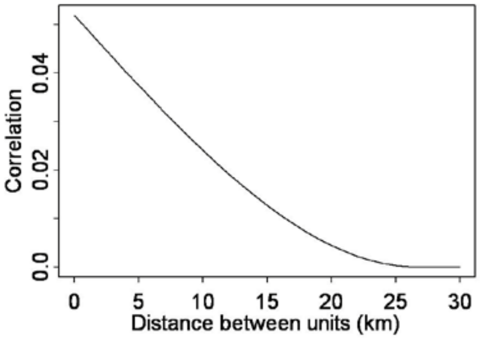

Figure 5: The estimated correlation function for DRP in units at various separation distances.

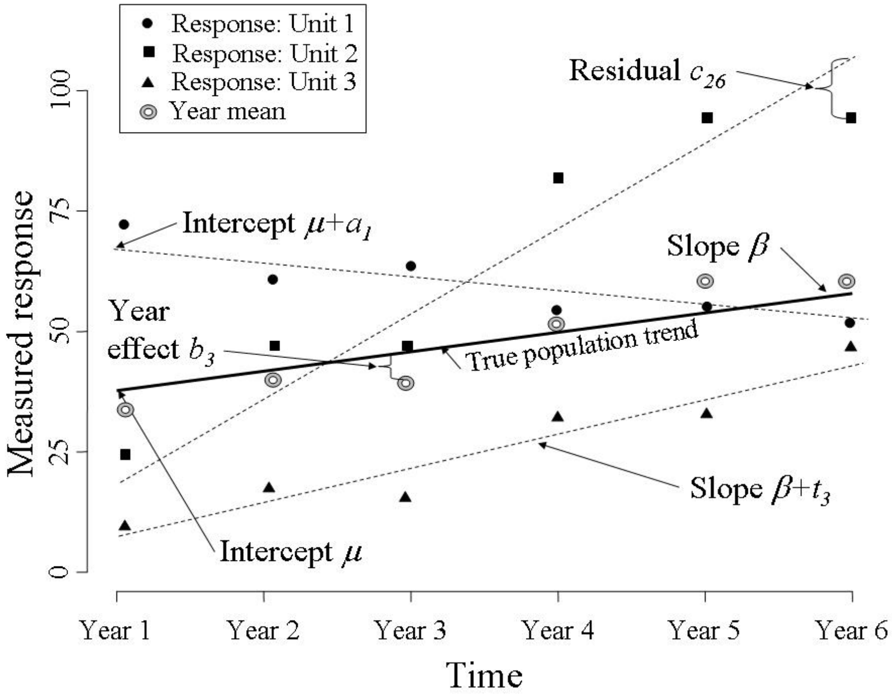

Figure 6: Pictorial representation of the mixed linear model proposed by Piepho and Ogutu (2002) to detect trends. Random effects are a_i (unit intercept) b_j (year), c_{ij} (residual), and t_i (unit slope). Fixed effects are μ and β. As proposed, trend is detected if $\beta \neq 0$. Lines are "eyeballed" estimates for illustration only.

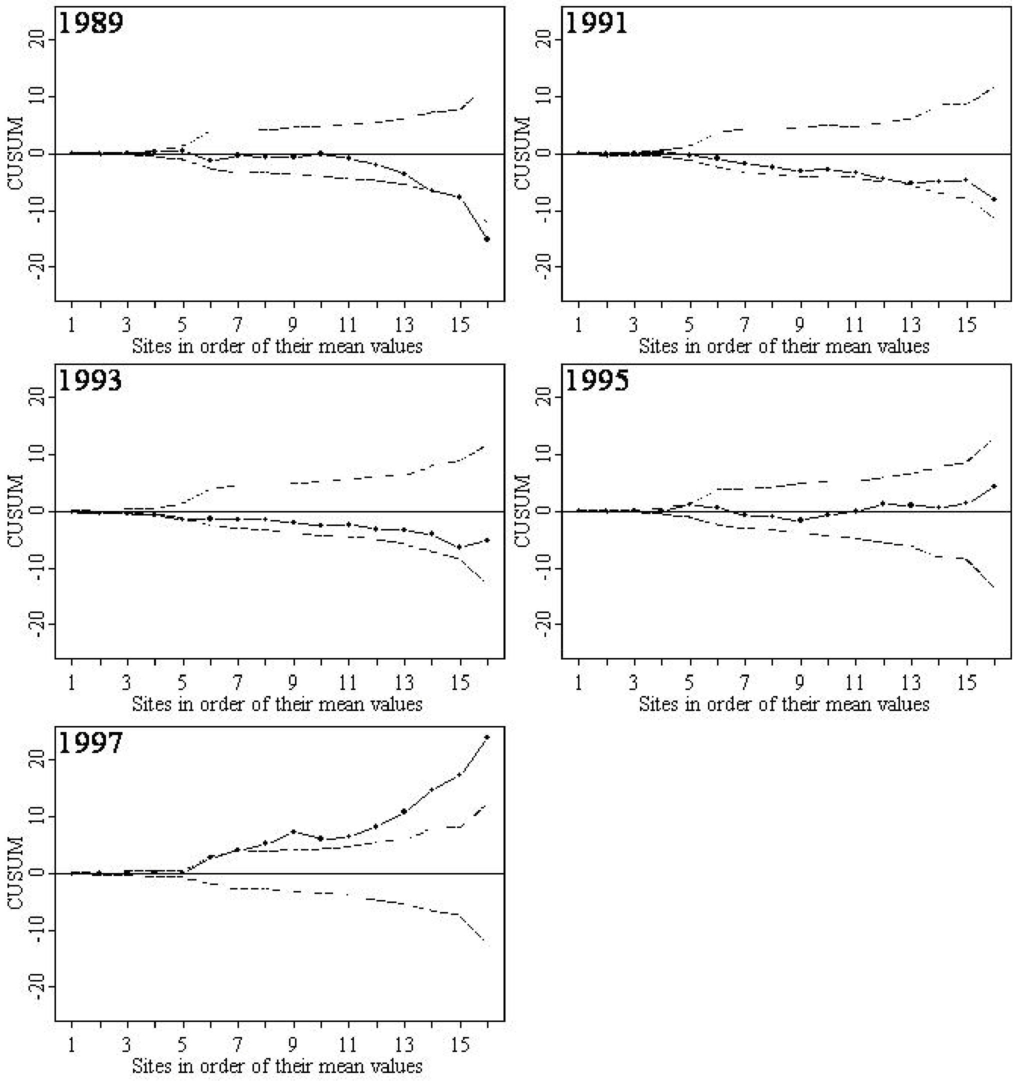

Figure 7: CUSUM charts for trend in the DLP during each year of the example study. The CUSUM (–•–) is considered to provide significant evidence against the null hypothesis of no trend if it exceeds the 95% confidence envelope lines (– –) at some point. The 95% confidence envelopes were determined by permuting values at each site over time.

Literature Cited

Agarwal, C. L. & Tikkiwal, B. D. (1975) Two-stage sampling on successive occasions. Proceedings of the 62nd session of Indian science (pp. 31).

Berryman, D., Bobee, B., Chuis, D., & Haemmerli, J. (1988). Nonparametric tests for trend detection in water quality time series. Water Resources Bulletin, 24, 545-556.

Binder, D. A. & Hidiroglou, M. A. (1988) Sampling in time. (In P.R. Krishnaiah & C.R. Rao (Eds.), Handbook of Statistics (pp. 187-211). Amsterdam: North Holland.)

Cabilio, P. & Tilley, J. (1999). Power calculations for tests of trend with missing observations. Environmetrics, 10, 803-816.

Cassel, C.M., Sarndal, C.E. & Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. (New York, NY: Wiley)

Cochran, W.G. (1977). Sampling Techniques. (New York, NY: Wiley)

Conover, W.J. (1980). Practical nonparametric statistics. (New York, NY: John Wiley & Sons)

Cressie, N.A. (1993). Statistics for Spatial Data. (New York, NY: Wiley)

Eager, C., Miller-Weeks, M., Gillespie, A.J.R. & Burkman, W. (1991). Summary report: forest health monitoring: New England/ Mid-Atlantic. (Radnor, PA: US Department of Agriculture, Forest Service)

Edwards, D. (1998). Issues and themes for natural resources trend and change detection. Ecological Applications, 8, 323-325.

El-Shaarawi, A. H. (1995). Trend detection and estimation with environmental applications. Mathematics and Computers in Simulation, 39, 441-447.

El-Shaarawi, A. H. & Niculescu, S. P. (1992). On Kendall's tau as a test of trend in time series data. Environmetrics, 3, 389-411.

El-Shaarawi, A. H. & Niculescu, S. P. (1993). A simple test for detecting non-linear trend. Environmetics, 4, 233-242.

Ericson, W. A. (1988) Bayesian inference in finite populations. (In P.R. Krishnaiah & C.R. Rao (Eds.), Handbook of Statistics (pp. 213-246). Amsterdam: North Holland.)

Ernst, T. L., Leibowitz, N. C., Roose, D., Stehman, S., & Urquhart, N. S. (1995). Evaluation of USEPA environmental monitoring and assessment program's (EMAP) - wetlands sampling design and classification. Environmental Management, 19, 99-113.

Esterby, S. R. (1993). Trend analysis methods for environmental data. Environmetrics, 4, 459-481.

Guttorp. P, Meiring, W., & Sampson, P. D. (1994). A space-time analysis of ground-level ozone data. Environmetrics, 5, 241-254.

Hirsch, R. M., Alexander, R. B., & Smith, R. A. (1991). Selection of methods for the detection and estimation of trends in water quality. Water Resources Research, 27, 803-813.

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

Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of trend analysis for monthly quality data. Water Resources Research, 18, 107-121.

Hollander, M. & Wolfe, D.A. (1999). Nonparametric statistical methods. (New York, NY: John Wiley & Sons)

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.

Huang, H. C. & Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kaman filter. Computational Statistics and Data Analysis, 22, 159-175.

Isaaks, E.H. & Srivastava, R.M. (1989). An Introduction to Applied Geostatistics. (New York, NY: Oxford university Press)

Kendall, M.G. (1975). Rank correlation methods. (London, England: Charles Griffin)

Kish, L. (1965). Survey Sampling. (New York, NY: Wiley)

Krishnaiah, P.R. & Rao, C.R. (1988). Handbook of Statistics. (New York, NY: North-Holland)

Lahiri, S.N. (2003). Resampling methods for dependent data. (New York, NY: Springer)

Leatherberry, E.C., Spencer, J.S., Schmidt, T.L. & Carroll, M.R. (1995). An analysis of Minnesota's fifth forest resources inventory, 1990. (St. Paul, MN: US Department of Agriculture, Forest Service, North Central Forest Experiment Station)

Lesica, P. & Steele, B. M. (1996). A method for monitoring long-term population trends: an example using rare arctic-alpine plants. Ecological Applications, 6, 879-887.

Lettenmaier, D. O. (1988). Multivariate nonparametric tests for trend in water quality. Water Resource Bulletin, 24, 505-512.

Link, W. A. & Sauer, J. R. (1998). Estimating population range from count data: application to the North American breeding bird survey. Ecological Applications, 8, 258-268.

Link, W. A. & Sauer, J. R. (2002). A hierarchical analysis of population change with application to cerulean warblers. Ecology, 83, 2832-2840.

Link, W. A. & Doherty, P. F. Jr. (2002). Scaling in Sensitivity Analysis. Ecology, 83, 3299-3305.

Littell, R.C., Milliken, G.A., Stroup, W.W. & Wolfinger, R.D. (1996). SAS System for Mixed Models. (Cary, NC: SAS Institute Inc.)

Lohr, S.L. (1999). Sampling: design and analysis. (New York, NY: Duxbury Press)

Mace, R. D., Waller, J. S., Manly, T. L., Lyon, L. J., & Zuuring, H. (1996). Relationships among grizzly bears, roads and habitat in the Swan Mountains, Montana. Journal of Applied Ecology, 33, 1395-1404.

MacNally, R. & Hart, B. T. (1997). Use of CUSUM methods for water quality monitoring in storages. Environmental Science and Technology, 31, 2114-2119.

Manly, B.F.J. (1997). Randomization, Bootstrap and Monte Carlo Methods in Biology. (London, England: Chapman and Hall)

Manly, B.F.J. (2001). Statistics for Environmental Science and Management. (Boca Raton, FL: Chapman and Hall/CRC)

Manly, B. F. J. & Mackenzie, D. (2000). A cumulative sum type of method for environmental monitoring. Environmetrics, 11, 151-166.

Manly, B. F. J. & Mackenzie, D. I. (2003). CUSUM environmental monitoring in time and space. Environmental and Ecological Statistics, 10, 231-247.

Manly, B. F. J. (1994) CUSUM methods for detecting changes in monitored environmental variables. (In Fletcher, D. J. & Manly, B. F. J. (Eds.), Statistics in ecology and environmental monitoring (pp. 225-238). Dunedin, New Zealand: University of Otago.)

Mann, H. B. (1945). nonparametric tests against trend. Econometrica, 13, 245-259.

McDonald, T.L. (1996) Analysis of finite population surveys: sample size and testing considerations. Dissertation, Oregon State University

McDonald, T. L. (2003). Review of environmental monitoring methods: survey designs. Environmental Monitoring and Assessment, 85, 277-292.

McRoberts, R. E. & Hansen, M. (1999). Annual forest inventories for the north central region of the united states. Journal of Agricultural, Biological, and Environmental Statistics, 4, 361-371.

Milliken, G.A. & Johnson, D.E. (1984). Analysis of messy data. (New York, NY: Van Hostrand Reinhold Co. Inc.)

Moisen, G. G. & Edwards, T. C. (1999). Use of generalized linear models and digital data in a forest inventory of northern utah. Journal of Agricultural, Biological, and Environmental Statistics, 4, 372-390.

Montgomery, D.C. (1991). Introduction to Statatistical Quality Control. (New York: Wiley)

Neter, J., Wasserman, W. & Kutner, M.H. (1985). Applied Linear Statistical Models, Regression, Analysis of Variance, and Experimental Designs. (Homewood, IL: Richard D. Irwin, Inc.)

Nijman, T., Verbeek, M., & Van Soest, A. (1990). The efficiency of rotating-panel designs in an analysis-of-variance model. Journal of Econometrics, 49, 373-399.

Nusser, S. M., Breidt, F. J., & Fuller, W. A. (1998). Design and estimation for investigating the dynamics of natural resources. Ecological Applications, 8, 234-245.

Nusser, S. M. & Goebel, J. J. (1997). The national resources inventory: a multi-resource monitoring program. Ecological and Environmental Statistics, 4, 181-204.

Olsen, A. R., Sedransk, J., Gotway, C., Liggett, W., Rathbun, S., Reckhow, K., Young, L., & Edwards, D. (1998). Statistical issues for monitoring ecological and natural resources in the united states. Environmental Monitoring and Assessment.

Olsen, A. R. & Smith, E. P. (1999). Introduction to the special issue of surveys over time. Journal of Agricultural, Biological, and Environmental Statistics, 4, 328-330.

Overton, W. S., White, D., and Stevens, D. L. Design report for emap, evironmental monitoring and assessment program. 1990. Washington, DC, US Environmental Protection Agency.

Pankratz, A. (1983). Rorecasting with Univariate Box-Jenkins Models. (New York, NY: John Wiley and Sons)

Piepho, H. P. & Ogutu, J. O. (2002). A simple mixed model for trend analysis in wildlife populations. Journal of Agricultural, Biological, and Environmental Statistics, 7, 350-360.

Reams, G. A. & Deusen, P. V. (1999). The southern ammual forest inventory system. Journal of Agricultural, Biological, and Environmental Statistics, 4, 346-360.

Sarndal, C.E., Swensson, B. & Wretman, J.H. (1992). Model Assisted Survery Sampling. (New York, NY: Springer-Verlag)

Scheaffer, R.L., Mendenhall, W. & Ott, L. (1986). Elementary Survey Sampling. (Boston: PWS-Kent)

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probability. Journal of the Indian Society of Agricultural Statistics, 5, 119-127.

Sen, P. k. (1968). Estimates of regression coefficeint based on kendall's tau. Journal of the American Statistical Association, 63, 1379-1389.

Singh, A. C., Kennedy, B., & Wu, S. (2001). Regression composite estimation for the Canadian Labour Force survey with a rotating panel design. Survey Methodology, 27, 33-44.

Smith, E. P. & Rose, K. A. (1991). Trend detection in the presence of covariates: stagewise versus multiple regression. Environmetrics, 2, 153-168.

Stevens, D. L. & Olsen, A. R. (1999). Spatially restricted surveys over time for aquatic resources. Journal of Agricultural, Biological, and Environmental Statistics, 4, 415-428.

Stevens, D. L. & Olsen, A. R. (2002). Variance estimation for spatially balanced samples of environmental resources. Environmetrics, 14, 593-610.

Thompson, S.K. (1992). Sampling. (New York, NY: Wiley)

Urquhart, N. S. & Kincaid, T. M. (1999). Designs for detecting trend from repeated surveys of ecological resources. Journal of Agricultural, Biological, and Environmental Statistics, 4, 404-414.

Urquhart, N. S., Paulsen, S. G., & Larsen, D. P. (1998). Monitoring for policy-relevant regional trends over time. Ecological Applications, 8, 246-257.

Van Strien, A. J., Van de Pavert, R., Moss, D., Yates, T. J., Van Swaay, C. A. M., & Vos, P. (1997). The statistical power of two butterfly monitoring schemes to detect trends. Journal of Applied Ecology, 34, 817-828.

VanLeeuwen, D. M., Murray, L. W., & Urquhart, N. S. (1996). A mixed model with both fixed and random trend components across time. Journal of Agricultural, Biological, and Environmental Statistics, 1, 435-453.

Ver Hoef, J. (2002). Sampling and geostatistics for spatial data. Ecoscience, 9, 152-161.

Wikle, C. K. & Royle, J. A. (1999). Space-time dynamic design of environmental monitoring networks. Journal of Agricultural, Biological, and Environmental Statistics, 4, 489-508.

Woodard, R., Sun, D., He, Z., & Sheriff, S. L. (1999). Estimating hunting success rates via bayesian generalized linear models. Journal of Agricultural, Biological, and Environmental Statistics, 4, 456-472.

Yates, F. & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. Journal of the Royal Statistical Society, B, 15, 235-261.