A Nonparametric Trend Test for Seasonal Data With Serial Dependence

ROBERT M. HIRSCH AND JAMES R. SLACK

U.S. Geological Survey

INTRODUCTION

One of the problems in detecting and evaluating trends in hydrologic data is the confounding effect of serial dependence. When a data set shows a drift towards higher values (or lower values) over the period of record, one needs to ask the following question: Is this drift an indication of an underlying change or is it an indication of long-term persistence? Whether one is examining a data set by eye or doing a formal test, this question will arise. One part of the answer to the question may come from an analysis of the generating mechanism for the data. Perhaps the data are dependent on some process which is serially correlated. In this case, working with residuals may eliminate or reduce the persistence in the data. Where this is not possible or not appropriate, then one may need to consider serial dependence in the formal trend test. Parametric methods for doing this are well developed and documented [see Box and Jenkins, 1970; Box and Tiao, 1975; D'Astous and Hipel, 1979]. However, with some hydrologic data there may be compelling reasons for using a nonparametric approach to trend detection. Hirsch et al. [1982] and Lettenmaier et al. [1982] discuss the reasons for using nonparametric procedures for water quality data. Lettenmaier [1979] discusses network design implications of serial dependence in conjunction with nonparametric testing but does not offer an operational scheme for adjusting trend tests for dependence. (Lettenmaier assumed the correlation structure to be known.) Sen [1963, 1965] proposed some extensions of nonparametric tests to data sets with certain types of dependence and showed that the test statistics were asymptotically normal. Lettenmaier [1976] found that for sample sizes and correlations encountered in practice, the normal approximations were unacceptable. Hirsch et al. [1982] propose a modified version of the Mann-Kendall test, the Seasonal Kendall test, but note that it is not robust against serial dependence. That is, when serial dependence exists, the actual significance level of the test exceeds the nominal significance level. In this paper we propose a modification of the original Seasonal Kendall test which is robust against serial dependence and, like the original, is based entirely on ranks. Missing values or censoring present no obstacles to its application. By a Monte Carlo experiment we demonstrate that this modified test is robust against serial dependence (in terms of type I error) except when the data have very strong long-term persistence or when sample sizes are small (e.g., 5 years of monthly data).

REASONS FOR USING NONPARAMETRIC TESTS

This paper will not consider in detail the reasons for using nonparametric rather than parametric tests, and comparisons of power between the two types of tests will not be made [see Bradley, 1968; Hirsch et al., 1982]. The test described is intended for use with seasonal data which are suspected of being serially correlated and where one or more of the following conditions exist in the data set.

1. The data are nonnormal. Many types of hydrologic data are distinctly nonnormal (usually positively skewed), in particular, discharge, and water quality variables related to washoff phenomena (sediment and nutrients attached to sediment) or biological indicators (biomass, bacterial counts, and chlorophyll). Dissolved constituents concentrations are distinctly nonnormal in some cases but not in others. Among all the commonly measured variables, only temperature, pH, and dissolved oxygen can be considered to be typically normal or near normal. When data sets are small, as is often the case with water quality data, the tests for normality will only reveal the most extreme violations. Using a test that relies on an assumption of normality, even when the hypothesis of normality cannot be rejected, should probably be done only with considerable caution by checking for undue influence of extreme values on the outcome of the test.

2. There are missing values in the data. The parametric procedures for trend detection, used when serial correlation exists [Box and Tiao, 1975; Hipel et al., 1975], depend on uniform sampling. Techniques exist to deal with a few isolated data gaps [Lettenmaier, 1976; D'Astous and Hipel, 1979] by estimating values for the missing data. However, if there are a lot of missing values, or one or more long gaps exist, the effect of data fill in on the identification of the stochastic process and the ultimate trend testing becomes very problematic. Harned et al. [1981] have employed various methods of aggregating seasonal data into annual summary values. This has the advantage that such annual series typically have only minimal serial dependence, and thus testing for trends can be carried out in straightforward fashion (either parametrically or nonparametrically). However, in the presence of missing values (or any irregular sampling schedule) and seasonality,
these annual summary values will be biased and trends may be detected which are simply artifacts of the year-to-year variations in the sampling schedule.

3. The data are censored. Censored data are those observations reported as being “less than” or “greater than” some specific value. Typical examples include concentration values for metals, or organic compounds which fall below the limit of detection (LD) of the analytical procedure and then are reported as “less than LD.” Censoring may also exist in flood data when long historical records are used. But this case would generally involve annual series data rather than seasonal data. Where “less than LD” observations arise in a data set, parametric methods require substituting some numerical value for the “less than LD” observations. Whatever numerical value is used, it will make the parametric test inexact and will severely violate the assumption of normality. Provided that the LD has not changed over the period of record, nonparametric tests such as the one described here may be used with no difficulty. All “less than LD” values are considered tied with each other and are considered to be lower than any numerical values at or above LD. If LD has changed over the record from LD1 to LD2, then all data indicated as “less than LD,” as well as any numerical values less than LD1, must be recoded to “less than LD,” and then the test may be run as described above.

THE ORIGINAL SEASONAL KENDALL TEST

We first describe the univariate test for trend described by Mann [1945]. Let \( X_1, X_2, \ldots, X_n \) be a sequence of observations ordered by time. We wish to test the null hypothesis \( H_0 \) that the observations are randomly ordered versus the alternative of monotone trend over time. Let
\[
\text{sgn}(x) = \begin{cases} 
+1 & x > 0 \\
0 & x = 0 \\
-1 & x < 0 
\end{cases}
\]
Then under \( H_0 \) the test statistic
\[
S = \sum_{i<j} \text{sgn}(X_j - X_i)
\]
has mean 0 and variance \( \sigma^2 = n(n-1)(2n+5)/18 \) and is asymptotically normal [Kendall, 1975].

Hirsch et al. [1982] defined a multivariate extension of this test, designed for seasonal data. We describe it initially here for the case where there are complete records for all \( n \) years and no ties. Let the matrix
\[
X = \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1p} \\
X_{21} & X_{22} & \cdots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{np}
\end{pmatrix}
\]
denote a sequence of observations taken over \( p \) seasons for \( n \) years. The null hypothesis \( H_0 \) is that for each of the \( p \) seasons the \( n \) observations are randomly ordered, versus the alternative of a monotone trend in one or more seasons. Let the matrix
\[
R = \begin{pmatrix}
R_{11} & R_{12} & \cdots & R_{1p} \\
R_{21} & R_{22} & \cdots & R_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
R_{n1} & R_{n2} & \cdots & R_{np}
\end{pmatrix}
\]
be the matrix of ranks corresponding to the observations in \( X \), where the \( n \) observations for each season are ranked among themselves. Specifically,
\[
R_{ij} = \left[ n + 1 + \sum_{k=1}^{n} \text{sgn}(X_{ij} - X_{ik}) \right]/2
\]
Thus each column of \( R \) is a permutation of \( (1, 2, \ldots, n) \). The Mann-Kendall test statistic for each season is
\[
S_g = \sum_{i<j} \text{sgn}(X_{ij} - X_{ig}) \quad g = 1, 2, \ldots, p
\]
The Seasonal Kendall test statistic is
\[
S' = \sum_{g=1}^{p} S_g
\]
and it is asymptotically normal with mean 0 and variance
\[
\text{var}[S'] = \sum_{g} \sigma_g^2 + \sum_{g \neq h} \sigma_{gh}
\]
Where \( \sigma_g^2 = \text{var}[S_g] \) and \( \sigma_{gh} = \text{cov}(S_g, S_h) \). Hirsch et al. [1982] assume that the data are independent and thus all of the covariance terms equal zero. They also demonstrate that the normal approximation is quite accurate even for sample sizes as small as \( n = 2, p = 12 \).

THE ESTIMATE OF THE COVARIANCE

Dietz and Killeen [1981], in defining a related multivariate distribution-free test, develop a consistent estimator for \( \sigma_{gh} \):
\[
\hat{\sigma}_{gh} = \frac{K_{gh}}{3} + \frac{(n^3 - n)r_{gh}}{9}
\]
where
\[
K_{gh} = \sum_{i<j} \text{sgn}[(X_{ij} - X_{ih})(X_{jh} - X_{gh})]
\]
\[
r_{gh} = \frac{3}{n^3 - n} \sum_{i,j,k} \text{sgn}(X_{ij} - X_{ih})(X_{jk} - X_{kh})
\]
If there are no ties and no missing values, \( r_{gh} \) is Spearman's correlation coefficient for seasons \( g \) and \( h \) [Conover, 1980; Lehman, 1975]. If there are no missing values, (6) reduces to
\[
\hat{\sigma}_{gh} = \frac{K_{gh} + 4 \sum_{i=1}^{n} R_{ig} R_{ih} - n(n+1)^2}{3}
\]
Using these estimates of \( \sigma_{gh} \) in the computation of the variance of \( S' \), we have a test that does not rely on an assumption of independence.

EMPIRICAL SIGNIFICANCE LEVEL

FOR SMALL SAMPLES

The trend test was performed on samples of size \( n = 5, 10, 20 \), and \( p = 12 \) from an autoregressive moving average (ARMA) process [Box and Jenkins, 1970]; in particular,
\[
X_{i,g} = \phi X_{i,g-1} + U_{i,g} - \theta U_{i,g-1} \quad g = 2, 3, \ldots, 12 \quad i = 1, 2, \ldots, n
\]
where \( U_{i,g}/\sigma_u \) are independently and identically distributed according to the normal distribution with mean zero and variance one \( \{N(0, 1)\} \), \( \sigma_u^2 = (1 - \phi^2)/(1 - 2\phi\theta + \theta^2) \) and \( X_{0,12} \) (the starting value) is \( N(0, 1) \). The process described here is an
TABLE 1. Empirical Significance Level for the Modified Seasonal Kendall Test

<table>
<thead>
<tr>
<th>n</th>
<th>ϕ</th>
<th>ρ₁</th>
<th>0.01</th>
<th>0.02</th>
<th>0.05</th>
<th>0.10</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.0</td>
<td>0.0*</td>
<td>0.000</td>
<td>0.000</td>
<td>0.009</td>
<td>0.060</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.2*</td>
<td>0.000</td>
<td>0.000</td>
<td>0.010</td>
<td>0.077</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.2*</td>
<td>0.000</td>
<td>0.000</td>
<td>0.012</td>
<td>0.080</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.4*</td>
<td>0.000</td>
<td>0.000</td>
<td>0.014</td>
<td>0.081</td>
<td>0.228</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.4*</td>
<td>0.000</td>
<td>0.000</td>
<td>0.012</td>
<td>0.088</td>
<td>0.240</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.4*</td>
<td>0.000</td>
<td>0.000</td>
<td>0.014</td>
<td>0.102</td>
<td>0.240</td>
</tr>
<tr>
<td>10</td>
<td>0.0</td>
<td>0.0*</td>
<td>0.003</td>
<td>0.010</td>
<td>0.041</td>
<td>0.094</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.2*</td>
<td>0.002</td>
<td>0.009</td>
<td>0.044</td>
<td>0.101</td>
<td>0.205</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.2*</td>
<td>0.006</td>
<td>0.006</td>
<td>0.045</td>
<td>0.109</td>
<td>0.205</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.2*</td>
<td>0.014</td>
<td>0.047</td>
<td>0.112</td>
<td>0.219</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.2*</td>
<td>0.016</td>
<td>0.056</td>
<td>0.114</td>
<td>0.220</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.2*</td>
<td>0.004</td>
<td>0.014</td>
<td>0.057</td>
<td>0.124</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.005</td>
<td>0.018</td>
<td>0.056</td>
<td>0.125</td>
<td>0.247</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>0.0*</td>
<td>0.008</td>
<td>0.017</td>
<td>0.047</td>
<td>0.102</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.2*</td>
<td>0.008</td>
<td>0.017</td>
<td>0.048</td>
<td>0.106</td>
<td>0.196</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.2*</td>
<td>0.010</td>
<td>0.020</td>
<td>0.052</td>
<td>0.110</td>
<td>0.204</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.4*</td>
<td>0.010</td>
<td>0.024</td>
<td>0.054</td>
<td>0.110</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.6*</td>
<td>0.011</td>
<td>0.026</td>
<td>0.057</td>
<td>0.118</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.2*</td>
<td>0.026</td>
<td>0.064</td>
<td>0.124</td>
<td>0.224</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.013</td>
<td>0.029</td>
<td>0.066</td>
<td>0.125</td>
<td>0.240</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.048</td>
<td>0.082</td>
<td>0.160</td>
<td>0.250</td>
<td>0.378</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.060</td>
<td>0.093</td>
<td>0.168</td>
<td>0.262</td>
<td>0.379</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.060</td>
<td>0.094</td>
<td>0.174</td>
<td>0.263</td>
<td>0.384</td>
<td></td>
</tr>
</tbody>
</table>

Control limits α ± 2[(1 - α)2/2000]1/2 for the empirical level; for a nominal level α: α = 0.01, 0.006-0.014; α = 0.02, 0.014-0.026; α = 0.05, 0.040-0.060; α = 0.10, 0.087-0.113; α = 0.20, 0.182-0.218.

*Process is independent.
†Process is AR(1).

ARMA (1, 1) process with mean zero and variance one. For purposes of description, the process is parameterized not by (ϕ, θ) but by (ϕ, ρ₁), where

\[
ρ₁ = \frac{(1 - ϕθ)(ϕ - θ)}{1 - 2ϕθ + ϑ^2}
\]

and is the lag one correlation coefficient. Note that when ϕ = ρ₁, the process is AR(1), and when ϕ = ρ₁ = 0.0, the process is independent.

Table 1 lists the empirical level of the test where empirical level is the ratio of number of rejections of H₀ to number of trials (2000) for a given nominal significance level. The nominal levels considered are α = 0.01, 0.02, 0.05, 0.10, 0.20. Table 1 shows that for n = 5 (5 years x 12 months = 60 observations) the asymptotic normal distribution yields very conservative results especially where α is low (0.01, 0.02, 0.05). Where the data are generated from a process with very high persistence (particularly ϕ = 0.9), the test is liberal at α = 0.1 and 0.2. For n = 10, the test performs much better. It is conservative at α = 0.01 and 0.02 for all processes except those with ϕ = 0.9. For α = 0.05, 0.1, and 0.2 and ϕ < 0.9, the empirical and nominal levels generally match closely. For ϕ = 0.9, it is again quite liberal, with empirical levels exceeding nominal levels by a factor of 2 or more. For n = 20 (a total of 240 observations), the empirical and nominal levels agree well except where ϕ = 0.9. In all cases (n = 5, 10, 20 for all ϕ and α) the empirical level is affected only slightly by ρ₁. This can be explained by the fact that ρ₁ describes the short-term (month-to-month) correlation, and the covariance terms in (4) adjust for much of this correlation. What they do not adjust for is the correlation between values a year (or multiples of a year) apart in time. That is, (4) is based on an assumption of independence between Xₖ₋ₙ and Xₖ₊ₙ, and where ϕ is high (e.g., 0.9), this independence is severely violated. That this test should be rather inexact when ϕ = 0.9 is not surprising considering the fact that by almost any measure or technique it is very hard to distinguish strong persistence from trend. For example, a sample autocorrelation function (ACF) which does not decay to zero at high lags is one diagnostic indicator of trend [Nelson, 1973, p. 75], and yet this behavior in an ACF is precisely what one finds in stationary ARMA (1, 1) processes with high ϕ values. Our examination of a large number of sample ACF’s for deseasonalized water quality and flow data indicates that the vast majority of cases have characteristics of AR(1) processes with 0.0 ≤ ρ₁ ≤ 0.6, and only a few show indications of ARMA (1, 1) behavior with ϕ > 0.6, and many of these may be explained by the presence of man-induced trend.

Figure 1 summarizes the results in Table 1 for α = 0.05 and n = 10 and compares them with the empirical significance levels for the test described previously [Hirsch et al., 1982] where all δᵦ values are set to zero on the basis of the independence assumption. A figure for n = 20 would look very nearly identical.

Based on these results, it appears that using (6) for estimating δᵦ rather than setting δᵦ to 0 results in a far more accurate test provided that n is about 10 or larger. However, for n = 5, the approximation is poor.

POWER OF THE MODIFIED TEST

This improvement in the robustness of the test is not without "cost." What one gives up by using the modified form of the test versus the original test is power. If the data being considered were a serially independent process added to a linear trend, then the probability of rejecting H₀ for a given α.

Fig. 1. Empirical level of the trend tests for ARMA(1, 1) data as a function of the autoregressive parameter ϕ. Monthly data (p = 12) for 10 years (n = 10). Nominal significance level for test (α) is 0.05. Based on 2000 repetitions.
Empirical Evaluation of Significance with Missing Values

Data were generated as described above, but a specified fraction of the data were deleted from the record. The deleted values were selected randomly with each observation having an equal probability of deletion. Table 2 gives the results for missing value frequencies of 0, 10, 30, and 50% for independent series and AR(1) with $\phi = 0.4$ for $n = 10$, and $n = 20$, $p = 12$. The number of repetitions was 2000.

The results show no clear pattern of differences among the various amounts of missing data. Of the 60 results for a non-zero amount of missing data, only three show empirical levels which differ significantly ($z = 0.05$) from the no missing data case. These significant differences were evaluated by the chi-square test for difference in probability [Conover, 1980, p. 144-146]. Note that in 60 results, the expected number of significant difference is 3 ($0.05 \times 60$). These results indicate that the significance level of the modified test is not substantially affected by missing data at least up to a level of 50% missing.

Modification to Accommodate Censoring and Ties

When data are reported as "less than" a limit of detection, they may be arbitrarily set to some constant value which is less than the limit of detection for purposes of nonparametric trend testing. This is because the nonparametric tests are based on ranks rather than magnitudes; all censored values may be viewed as sharing the same rank, and this rank is less than the rank of any noncensored value. Thus the problem of censoring reduces to a problem of dealing with ties. For purposes of this discussion we will assume that there are no missing values. When ties and missing values are both present, one must combine the modifications described in the last section with those described in this one.

### TABLE 2. Empirical Level for the Modified Seasonal Kendall Test

<table>
<thead>
<tr>
<th>Nominal $\alpha$</th>
<th>Record Length in Years</th>
<th>Percent of Data Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>0.01</td>
<td>0.0030</td>
<td>0.0030</td>
</tr>
<tr>
<td></td>
<td>0.0085</td>
<td>0.0060</td>
</tr>
<tr>
<td>0.05</td>
<td>0.0410</td>
<td>0.0420</td>
</tr>
<tr>
<td>0.10</td>
<td>0.0945</td>
<td>0.1010</td>
</tr>
<tr>
<td></td>
<td>0.0470</td>
<td>0.0530</td>
</tr>
<tr>
<td>0.20</td>
<td>0.2190</td>
<td>0.1815*</td>
</tr>
<tr>
<td></td>
<td>0.1980</td>
<td>0.2165</td>
</tr>
<tr>
<td>0.40</td>
<td>0.2140</td>
<td>0.1990</td>
</tr>
</tbody>
</table>

2000 Monte Carlo trends, 10 or 20 years of monthly data, with 0, 10, 30, or 50% missing data.

*Indicates that empirical level with missing values differs significantly ($z = 0.05$) from empirical level with nonmissing values.

Modifications to Accommodate Missing Values

To accommodate missing values, we extended the definition of the sgn function given in (1) to handle missing values. Define $\text{sgn} (X_{ig} - X_{i+})$ to be zero if either $X_{ig}$ or $X_{i+}$ is missing.

In essence, we say that since we cannot tell whether a missing value is greater or less than any actual value, it is neither. In light of this, (3) becomes

$$ R_{ig} = \left[ n_g + 1 + \sum_{i=1}^{n_g} \text{sgn} (X_{ig} - X_{i+}) \right] / 2 $$

where $n_g$ is the number of nonmissing observations for season $g$. Now the ranks of the nonmissing observations are unchanged and each missing value is assigned the average or midrank ($n_g + 1$)/2. The Mann-Kendall test statistic $S_g$ is unchanged, and its variance remains the same, namely,

$$ \sigma^2 = n_g(n_g - 1)(2n_g + 5)/18 $$

Within (6) for $\delta_{gph}$, $K_{gph}$ (equation (7)) remains unchanged, but $r_{gph}$ (equation (8)) takes on a new value to give a revised (9) of:

$$ \delta_{gph} = \left[ K_{gph} + 4 \sum_{i=1}^{n_g} R_{ig} r_{gph} - n(n_g + 1)(n_g + 1) \right] / 3 $$

Fig. 2. Power of the trend tests. Percentage of trials in which trend was detected (500 repetitions) as a function of trend slope in percentage of the noise standard deviation per year. For independent monthly ($p = 12$) series of length 10 years ($n = 10$) for $\alpha = 0.05$. The closed symbol is for the test with $\delta_{gph} = 0$; the open symbol is for the test with $\delta_{gph}$ determined from (6).

would be higher using the original formulation with all $\delta_{gph} = 0$ than with the modified version given here. Figure 2 shows an empirical evaluation of power for the two formulations of the test. The records are 10 years long and serially independent. There were 500 repetitions at each of nine amounts of added linear trend, including zero trend. Trend slopes are expressed as a percentage of the trend-free standard deviation of the process. Expressed as a ratio, the most extreme difference in power occurs at a trend of 8% per year, where the power of the original test is 1.49 times the power of the modified test.

Thus choosing between the two tests involves a trade-off. The original test is more powerful, but the significance level can be seriously in error if there is serial correlation. The modified test requires some sacrifice of power but offers a more nearly exact statement of significance for a wide variety of cases.
The test statistics $S_g$, $g = 1, 2, \cdots, p$ are computed as in (4), and $S'$ is the sum of these $S_g$ values. Equation (5), giving the variance of $S_g$ becomes

$$
s_g^2 = \left[ \frac{n(n-1)(2n-5) + \sum_{j=1}^{m} t_j(t_j - 1)(2t_j + 5)}{18} \right]
$$

where $m$ is the number of tied groups among the $X_{ij}$ and $t_j$ is the size of the $j$th tied group [Kendall, 1975]. The formula for $d_{ag}$ remains the same except that midranks are used in assigning the values of $R_{ij}$ for (9). Thus, if there are $t_j$ censored values, they all have rank $t_j(t_j - 1)/2$.

**Empirical Evaluation of Significance With Censoring**

Rather than consider the general case of ties, we have limited our consideration here to the case of censoring. Data were generated as described in (10), but those values below a given value (LD) were assigned a value equal to LD. LD values were chosen for the simulation to achieve a certain percentage of censoring on the average. The following cases were considered: $n = 10$, independent, and AR(1) with $\rho_1 = 0.4$, with 10, 30, and 50% censored, and $n = 20$, independent, and AR(1) with $\rho_1 = 0.4$ with 50% censored. Two thousand replicates were used in all cases. At $\alpha$ levels of 0.01, 0.02, 0.05, 0.10, and 0.20, there were no instances where the empirical level of the test differed significantly (at the 5% level) from the empirical level that was found when no censoring occurred. Significant differences were evaluated by the chi-square test for difference in probability [Conover, 1980, p. 144–146].

**Comparison with a Related Test**

Dietz and Killean [1981] propose a multivariate nonparametric test for monotone trend which is based on Kendall's tau. Their test statistic is the weighted sum of squares of the $S_g$ values, where the matrix of weights is the inverse of the covariance matrix (the $d_{ag}$ and $d_{ab}$ terms). The test statistic is asymptotically $\chi^2$ on $p$, degrees of freedom. Dietz and Killean examined the accuracy of the $\chi^2$ approximation for small samples. They found that the empirical level increased with increasing $n$ and decreased with increasing $p$. They examined the empirical level of their test for $p = 12$ (the 12 variables), for $n = 10$ in cases where trend is close for $n > 10$ for $p = 12$, for $n > 30$ if $p = 4$, and for $n > 20$ if $p = 2$. In contrast, the modified test proposed in this paper detected trend in exactly 10 cases out of 200 trials. The expected number of detections was 10 (0.05 × 200). The modified seasonal Kendall test proposed in this paper detected trend in exactly 10 cases in the same set of Monte Carlo trials. For smaller values of $n$ the conservativeness of their test was even more severe. For an $n$ of 40, the empirical level rose, and there were six detections in 200 trials. Limited experiments with the power of their test show that for $n = 10$ in cases where trend is sufficiently large that trend is detected in our test ($\alpha = 0.05$) with a power of about 0.9, the Dietz and Killean test has a power of about 0.01. As $n$ is increased to 30 and beyond, the powers of the two tests approach each other more closely. Thus, based on Monte Carlo experiments by Dietz and Killean and ourselves, we see that the $\chi^2$ approximation becomes reasonably close for $n > 40$ if $p = 12$, for $n > 30$ if $p = 4$, and for $n > 20$ if $p = 2$. In contrast, for our modified seasonal Kendall test, the normal approximation is close for $n > 10$ for $p = 12$.

The kind of situation Dietz and Killean envision for applying their test is a case where the several variables were different in kind, not just seasonal values of the same variable. The example they present is of several measures of blood chemistry. They were implicitly concerned with the possibilities that some of these measures may show upward trends while others showed downward trends. Our test would be inappropriate for such a case. If upward trends in one or more seasons are counterbalanced by downward trends in an equal number of seasons, then the power of our test would equal $\alpha$ no matter how large the trends. This is because our test statistic is a sum of Kendall $S_g$ statistics (which would tend to cancel each other out) and theirs is a weighted sum of squares of the $S_g$ which would grow as the amount of trend grew.

**Conclusion**

The Seasonal Kendall test as originally presented by Hirsch et al. [1982] is robust against seasonality, departures from normality, and may be used in situations where there is censoring or many missing values. It is not, however, robust against serial dependence. That is, when the data arise from a stationary ARMA(1, 1) process, even one with monthly lag 1 serial correlations as low as 0.2, the probability that significant trend will be detected (at the level $\alpha$) is substantially higher than $\alpha$. The modification described here is to estimate the covariance between the Seasonal Kendall ($S_g$) statistics from the data, rather than setting it to zero. This estimate of the covariance was developed by Dietz and Killean [1981]. When the modified test is applied to data that arise from a stationary ARMA(1, 1) process, with $\alpha$ parameter $\phi < 0.6$ and record length at least 10 years of 12 months each, the probability of detecting significant trend at (level $\alpha$) is close to $\alpha$. The modified test is not robust against highly persistent processes ($\phi > 0.6$), but these may be atypical of hydrologic time series.

The modified test does not work well at small sample sizes less than 10 years and is less powerful than the original test when data are, in fact, independent. The original test is a useful screening device (and computationally much less demanding) but is inexact. The modified test is a more exact (conservative) and expensive test, useful for long seasonal time series. The test proposed by Dietz and Killean is probably only applicable for data sets of greater than 40 years of monthly data but has the advantage of sensitivity to opposing trends in different seasons, which is true of neither the original or modified seasonal Kendall test.

**References**


Hirsch, R. M., J. R. Slack, and R. A. Smith, *Techniques of trend


(Received November 28, 1983; revised February 28, 1984; accepted February 29, 1984.)