Seasonal Trend Analysis of Monthly Water Quality Data

Jie Tian and George C.J. Fernandez,
Department of Applied Economics & Statistics, University of Nevada, Reno

ABSTRACT

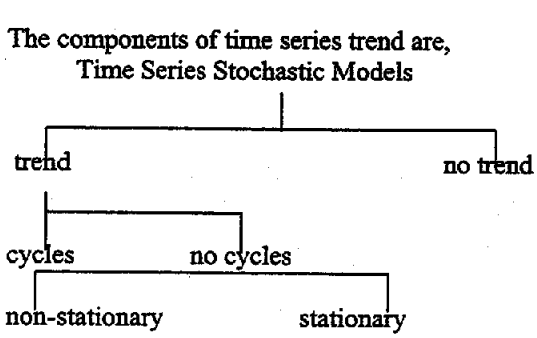
This paper presents user-friendly SAS macro and SAS programs to check the three parts of a time series: Trend, Cyclical patterns, and Stationarity. The SAS macro KENDALL performs the trend analysis. Seasonal Kendall Trend analysis, including summery statistics; overall Tau and the P-value of the test for trend; monthly Tau values, the corresponding P-values for each month; the Seasonal Kendall estimate with confidence intervals; 3 exploratory plots by months comparative boxplot, comparative histogram, and trend plot against year. Spectral analysis program calculates Periodogram ordinates and Spectral density estimates, Periodogram plots against frequency, plots of Spectral density estimates against frequency and period, 2 tests for White Noise. Dickey-Fuller program performs a test for Stationarity using the test of the unit-root hypothesis. An example is given using 1 water-quality data from Truckee River, Nevada. The water-quality variable being used is total nitrogen (NTO, mg/l).

(*key words*: Seasonal Kendall Trend Analysis, Spectral Analysis, Stationarity)

INTRODUCTION

The issue of surface-water quality is of vital importance. Various pollution sources related to natural-resource industries, primarily agriculture and mining, are having a strong effect on rivers and streams. Therefore, trend analysis of water quality fluctuations is essential for short and long term policy making.

The components of time series trend are,



SEASONAL KENDALL TREND ANALYSIS

Tau: Mann (1945) first suggested using the test for significance of Kendall's Tau as a test for trend where the X variable is time (T). The Mann-Kendall test can be stated most generally as a test for whether Y values tend to increase or decrease with T (monotonic change).

H_0: $\tau=0$, no correlation between x and y
H_a: $\tau\neq0$, x and y are correlated

To perform the test, Kendall's S statistic is computed form the Y,T data pairs:

$$S=P-M$$

where P= the number of $Y_i<Y_j$ for all i<j
M= the number of $Y_i>Y_j$ for i<j
A table of exact critical values is found in table B8.(Helsel & Hirsch 1992)

$$\tau=\frac{S}{n(n-1)/2}$$

For n>10

$$\sigma_s=\sqrt{(n/18)*(n-1)*(2n+5)}$$

$$Z_s=\begin{cases}\dfrac{S-1}{\sigma_s} & if S>0\\ 0 & if S=0\\ \dfrac{S+1}{\sigma_s} & if S<0\end{cases}$$

The null hypothesis is rejected at significance level α if $|Z_s|> Z_{\alpha/2}$ Where $Z_{\alpha/2}$ is the critical value from standard normal distribution.
Because of its robust statistical properties, Kendall's Tau is used very frequently. It is valid for data that are non-normal or cyclical. And it has the useful property of other non-parametric tests in that it is invariant to (monotonic) power transformations (e.g. logarithm).

Dealing With Seasonality : There are many instances where changes between different seasons of the year are a major source of variation in the Y variable. Therefore, we may be interested in modeling the seasonality to allow different predictions of Y for differing seasons.

The seasonal Kendall test (Hirsch et al., 1982) accounts for seasonality by computing the Mann-Kendall test on each of m seasons separately, and then combining the results. *Overall Tau*, is a weighted average of twelve rank-correlation coefficients, one for each month of the year. For each variable and site, Overall Tau and the twelve Monthly Tau values are tabulated in Appendix A. The P-value accompanying each Tau value is the result of a test of significance for each month. The test of significance of overall Tau is the test for trend. The Overall Tau is much more sensitive than the monthly tests for the simple reason that the overall test is based on a much larger sample (approximately 12 times larger). Thus, it is possible to have an Overall Tau that is significantly different from zero while none of the monthly Taus upon which it is based is significant. (See Appendix A)

Slope: The rate of change of each water-quality is quantified by the seasonal Kendall slope estimator (see Hirsch, et al.,1982, and Gilbert, 1987, pp.227-228). It has the same robust statistical properties as Tau. Moreover, the test of significance of Tau can also be interpreted as a test of significance of the hypothesis that the median change per year (slope) is different from zero. The Overall Seasonal Kendall Estimate, along with 90% and 95% confidence limits of the change per year, is tabulated in Appendix A. There are 3 kinds of estimates: (1) change per year (slope), (2) percent change per year (percentage of the median), and (3) relative change per year (%).

Relative Change per Year: The Seasonal Kendall Estimate for relative change per year is defined by Sungsue Rheem as the median of all values of

$$\frac{(Y_{ij} - Y_{ik}) / Y_{ik}}{j - k}$$

for i=1, ...,12, and $1 \le k < j \le n$, where Y_{ij} is the water-quality value for month i and year j, year k is the year other than j, and n is the number of total years.

SPECTRAL ANALYSIS

Uneven long-term variation in the series, other than trend, is called cycles. Spectral Analysis is a statistical approach to detect regular cyclical pattern, or periodicities. In Spectral Analysis the data are transformed with a finite Fourier transformation and decomposed into waves of different frequencies. Thus the time series model is expressed in terms of sine and cosine components:

$$Y_t = \sum_{k=1}^{m} (A_k \cos(\omega_k t) + B_k \sin(\omega_k t)) + e_t$$

where

* Y_t is the original time series variable with n observations.
* m=n/2, if n is even; m=(n-1)/2, if n is odd.
* A_k specifies cosine coefficients representing the amplitude, or height, of the cosine component.
* B_k specifies sine coefficients representing the amplitude, or height, of the sine component.
* ω_k specifies the frequencies, $2\pi k/n$, where k=1,2,...,m and $0 \le \omega_k \le \pi$.
* e_t is a random error term.

The time series is regressed on these components and the sum of squares are calculated for each regression component: the periodogram ordinates can be written as $n/2(A^2+B^2)$.

The significant components are then detected. The estimates are displayed as either Periodograms or Spectral Density Estimates. The Periodogram is an estimate of a theoretical quantity called a spectrum. The weighted moving averages of periodogram ordinates are called Spectral Density Estimates.

If the plot has high ordinates at high frequencies and low ordinates at low frequencies, this is indicative of short cycles as with negatively correlated series. If the plot has high ordinates at low frequencies and low ordinates at high frequencies, this pattern is indicative of long sinusoidal cycles as expected with positively correlated series. The spectrum for a white-noise process is simply a horizontal line. Repeated series and long cycles in the data can also be detected from these plots.

For the example, the time series is positively correlated. Each value of the variable PERIOD represents one month. The 3rd plot in Appendix B has several peaks before 50th month. Strong seasonal components appear before the date June 1981.

Testing for White Noise: PROC SPECTRA can be used to test the null hypothesis that a time series is white noise. There are two tests for white noise.

Fisher-Kappa test:
This is the ratio of the largest periodogram ordinate to the average of all the ordinates. The test is designed to detect one sinusoidal component buried in white noise. The large first ordinate is excluded from this analysis.

Bartlett Kolmogorov Smirnov (BKS) test:
For each frequency, ω_k, the sum of periodogram ordinates from ω_1 to ω_k is divided by the sum of all periodogram ordinates. This test has more power than the

Fisher-Kappa test to detect departures from the white-noise hypothesis over the whole range of frequencies.

The computed statistic for the Fisher-Kappa test can be compared to critical values in Appendix 3 (SAS/ETS Software: Applications Guide 1, 1991). For the BKS test, if there are 31 or more periodogram ordinates (m-1>30), then the critical statistics can be approximated with the following formulas:
5% critical point: $1.36(m-1)^{-1/2}$
1% critical point: $1.63(m-1)^{-1/2}$

For the example, both tests give the same results: the null hypothesis of white noise is rejected.

TEST FOR STATIONARITY

Stationarity is a mathematical and statistical property of time series data. Much of the probability theory is based on the assumption that time series are stationary. With the condition, values of the error term, ε, from the distance past, have a little influence on current values of the time series variable.

A series is stationary if $\mu t = \mu$ for all t, and $\sigma^2_t=\sigma^2$ for all t.

A formal statistical test for the existence of stationarity, known as the test of the unit-root hypothesis, was developed by Dickey and Fuller (1979). It involves a (p+1)th order autoregressive equation:
$Y_t = \alpha_1 Y_{t-1}+\alpha_2 Y_{t-2}+\ldots+\alpha_{p+1} Y_{t-p-1}+\varepsilon_t$
If all the characteristic roots of the equation are less than 1 in absolute value, Y_t is stationary.

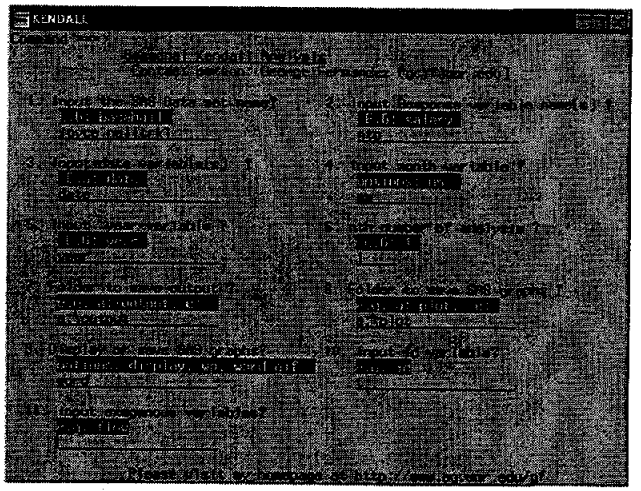
The calculated statistics have a special distribution. But they don't have to be compared to tables provided by Fuller(Shown in Appendix 2, SAS/ETS Software: Applications Guide 1, 1991), because the PROBDF function (SAS/ETS) computes the probability of observing a test statistic under the assumption that the null hypothesis is true. When PROBDF returns a small (significant) probability value, the unit root hypothesis is rejected.

IMPLEMENTATION

Seasonal Kendall Analysis can be performed by using the SAS MACRO KENDALL and the user-friendly macro-call window. The macro KENDALL was developed by modifying the SAS codes of Sungsue Rheem and Golde I. Holtzman and by incorporating additional codes for generating exploratory plots. The users can input the macro parameters, name of the input data set, the names of response variables (water-quality variable), the name of date variable, etc in the macro-call window (Figure 1). SAS users can download this macro-call file and run this KENDALL macro directly by accessing the Internet and visiting George Fernandez's home page at http://www.ag.unr.edu/gf and by following the instruction in downloading.

The other statistical procedure Spectral Analysis, and the test for Stationarity---has been the automated developing SAS programming code. SAS/STAT, SAS/BASE, SAS/GRAPH , and SAS/ETS are used. Programs and edited output are presented in Appendix A, B, C for 3 parts respectively.

Fig. 1. Kendall macro-call window:.



The last option of inputting exogenous variables is not required. They are the possible factors which account for a change in a process bringing the substance to the river. For detection of an, indeed, genuine trend, the adjustment is accomplished by applying the Seasonal Kendall Analysis to the residuals of a regression of the water-quality variable.

REFERENCE:

Bowerman, B.L.& O'Connell R.T. Forecasting and Time Series, an Applied Approach, Ohio: 3rd.edition, Wadsworth, 1993.

Helsel, D.R.& Hirsch, R.M., Statistical Methods in Water Resources, Amsterdam: Elsevier Science B.V., 1992.

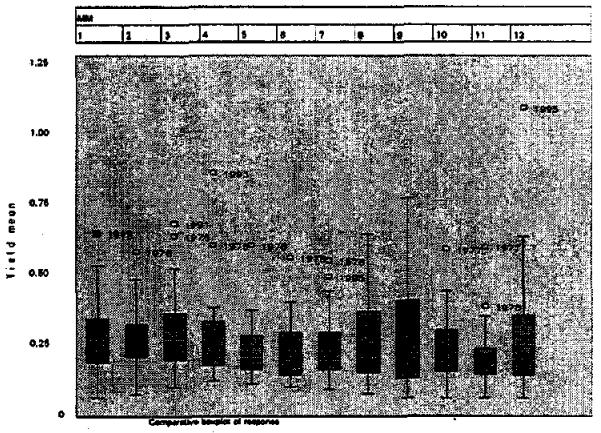
Ragavan, A.J. "Trend Analysis of Monthly Water Quality Data", M. S. Thesis in Hydrology, Univeresity of Nevada, Reno, 1996.

Rheem, S & Holtzman, G.I. , A SAS® Program for Seasonal Kendall Trend Analysis of Monthly Water Quality Data.

SAS/ETS Software: Applications Guide 1, Version 6, 1st ed. Cary, N.C.: SAS Institute, 1991.

Contact for authors at,
Jie (Joyce) Tian
(775) 784 4897
jietian9@yahoo.com

George C.J. fernandez
(775) 784 4206
gcjf@unr.edu



Appendix A:
Output of Seasonal Kendall Analysis by SAS

MACRO KENDALL

TREND PLOT BY MONTHS

STARTING DATE	ENDING DATE	# OF YEARS	# OF OBS.	MEDIAN	MEAN	STD.DEV.	MINIMUM	MAXIMUM
77-05-01	97-06-01	7337	242	0.21034	0.25083	0.15742	0.05	1.08

Kendall overall tau

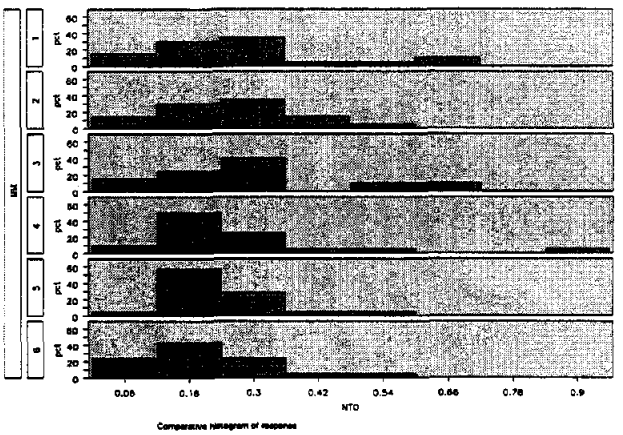
S	VAR(S) UNDER HYPOTHESIS OF NO TREND	Z	OVERALL TAU	P-VALUE OF TEST FOR TREND
424	11691.33	3.91209	0.18284	.000091502

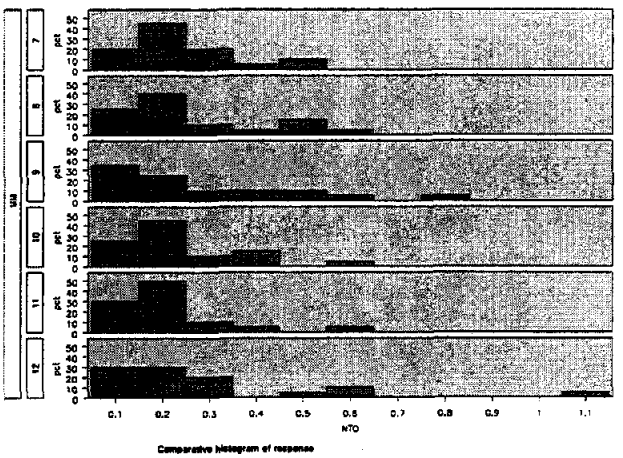
Kendall monthly tau

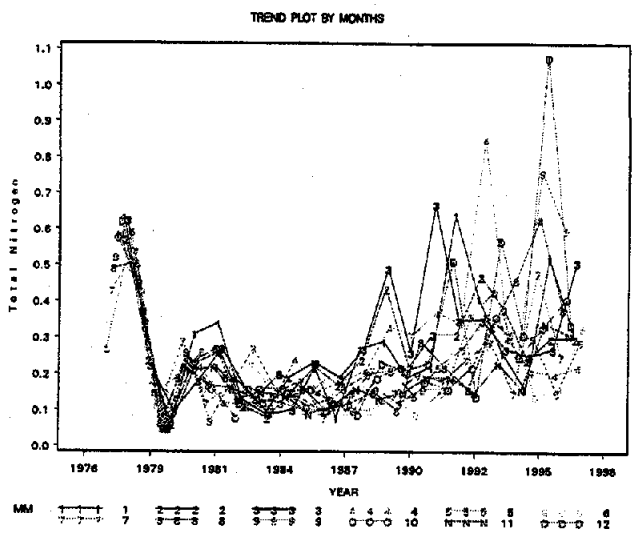
MONTH	TAU	p-value	# OF YEARS
1	0.13684	0.41730	20
2	0.30526	0.06441	20
3	0.33684	0.04095	20
4	0.25263	0.12729	20
5	0.14354	0.38075	21
6	-0.06667	0.69464	21
7	0.07368	0.67319	20
8	0.14737	0.38103	20
9	0.26316	0.11189	20
10	0.25263	0.12729	20
11	0.13684	0.41730	20
12	0.24211	0.14429	20

Kendall seasonal slope

DESCR	CHANGE PER YEAR (SLOPE)	CHANGE PER YEAR (PERCENTAGE OF MEDIAN)	RELATIVE CHANGE PER YEAR (%)	
SEASONAL KENDALL ESTIMATE	0.0000	0.01	0.01	Mid value
0.0000 0.01	0.01			
LOWER LIMIT OF 95% C.I.	0.0000	0.00	0.00	
UPPER LIMIT OF 95% C.I.	0.0000	0.00	0.00	
LOWER LIMIT OF 90% C.I.	0.0000	0.00	0.01	
UPPER LIMIT OF 90% C.I.	0.0000	0.00	0.00	







Appendix b
Spectral Analysis SAS Program Statements

```
proc spectra data=joyce.nalltrk1 out=joyce.ntosea1 p s
          adjmean whitetest;
    var nto;
    weights  1 2 3 4 3 2 1;
run;
proc print data=joyce.ntosea1(obs=10);
run;
```

```
*SAS program for Periodogram ;
goptions colors=(BLACK, RED, BLUE,YELLOW, GREEN, MAGENTA, CYAN)
reset=all norotate cells hpos=0 vpos=0 nosymbol noprompt
noborder gsfmode=replace
ctext=black  DEVICE=cgmmwwc
gaccess='sasgastd>c:\plot\spectral\periodogram.cgm';

proc gplot data=joyce.ntosea1;
     plot p_01*freq;
symbol1 i=splines v=dot;
title 'Periodogram of Total Nitrogen';
run;
```

```
*SAS program for Spectral density ;
goptions colors=(BLACK, RED, BLUE,YELLOW, GREEN, MAGENTA, CYAN)
reset=all norotate cells hpos=0 vpos=0 nosymbol noprompt
noborder gsfmode=replace
ctext=black  DEVICE=cgmmwwc
gaccess="sasgastd>c:\plot\spectral\density.cgm" ;

proc gplot data=joyce.ntosea1;
     plot s_01*freq;
symbol1 i=splines v=dot;
title 'Spectral density of Total Nitrogen';
run;
```

```
*SAS program Spectral density of Total Nitrogen
against period;
goptions colors=(BLACK, RED, BLUE,YELLOW, GREEN, MAGENTA, CYAN)
reset=all norotate cells hpos=0 vpos=0 nosymbol noprompt
noborder gsfmode=replace
ctext=black  DEVICE=cgmmwwc
gaccess='sasgastd>c:\plot\spectral\period.cgm"  ;
```

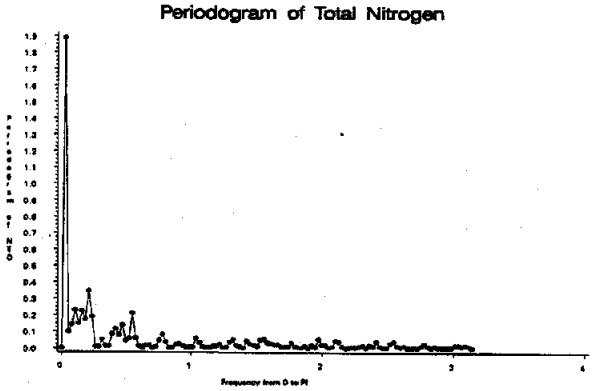
```
proc gplot data=joyce.ntosea1;
     plot s_01*period;
symbol1 i=splines v=dot;
title 'Spectral density of Total Nitrogen against
       period';
run;
```

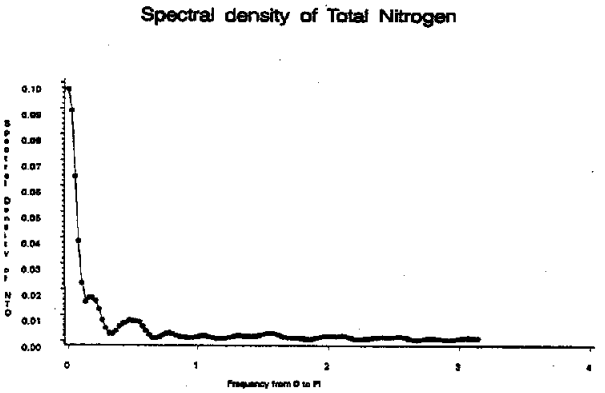
SAS output and graphs:

```
                    SPECTRA Procedure
      ----- Test for White Noise for variable NTO -----

          Fisher's Kappa: (M-1)*MAX(P(*))/SUM(P(*))
          Parameters:     M-1             =       120
                          MAX(P(*)) =     1.889
                          SUM(P(*)) =     5.997
Test Statistic: Kappa     =    37.7974

Bartlett's Kolmogorov-Smirnov Statistic:
Test Statistic =    0.5458
```

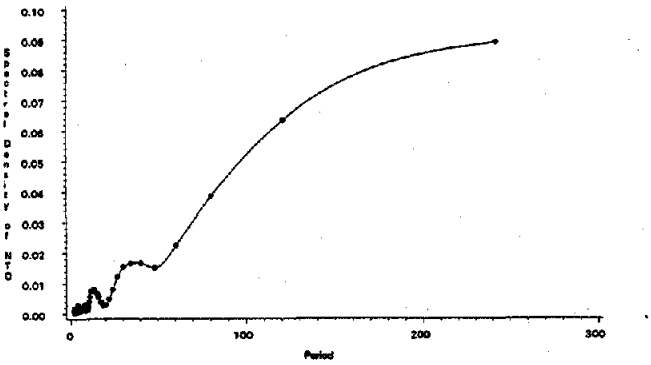
The SAS System

OBS	FREQ	PERIOD	P_01	S_01
1	0.00000	.	0.00000	0.097337
2	0.02596	242.000	1.88882	0.089082
3	0.05193	121.000	0.09982	0.063482
4	0.07789	80.667	0.14161	0.038500
5	0.10385	60.500	0.22906	0.022370
6	0.12982	48.400	0.14877	0.015076
7	0.15578	40.333	0.22377	0.016637
8	0.18175	34.571	0.17438	0.016505
9	0.20771	30.250	0.34593	0.015370
10	0.23367	26.889	0.19892	0.012188
...	...	...	...	...





Spectral density of Total Nitrogen against period



Appendix C
SAS Program for Testing for Stationary

```
data dfuller;
   set joyce.nalltrk1;
   y1=lag(nto);
   yd=dif(nto);
   yd1=lag1(nto); yd2=lag2(nto);
   yd3=lag3(nto); yd4=lag4(nto);
run;

proc reg data=dfuller outest=alpha covout;
   model yd=y1 yd1-yd4/noprint;
run;

data Pval;
  set alpha;   retain a;
  if _type_='PARMS' then a=y1-1;
  if _type_='COV' & _name_='Y1' then do;
    x=a/sqrt(y1);
    p=probdf(x, 99, 1, 'SSM');
 end;  run;

proc print data=pval; title 'P-value';
run;
```