Trends Examples

Dave Lorenz

January 21, 2016

Abstract

These examples demonstrate some functions in the `smwrStats` package that facilitate or perform trend analysis. The first example uses the ConecuhFlows dataset from Appendix C2 in Helsel and Hirsch (2002) to illustrate trend analysis of annual series data. The second example uses the KlamathTP data set in the smwrData package to illustrate seasonally adjusted trend analysis. This example uses the data modified for censoring in uncensored techniques. All section references in these examples are from Helsel and Hirsch (2002). The critical alpha value for all tests is 0.05.

Contents

1 Introduction

These examples use data from the smwrData package. The data are retrieved in the following code.

```
> # Load the stats, smwrData, and smwrStats packages
> library(stats)
> library(smwrData)
> library(smwrStats)
> # Get the datasets
> data(ConecuhFlows)
> data(KlamathTP)
```

2 The Mann-Kendall Trend Test

Helsel and Hirsch (2002) stress plotting the data to help understand the data and find an appropriate statistical technique. For this analysis, a simple graph of the annual mean flow by year shows a downward trend—more observations are above the median line in the first ten years and more below in the last ten years. The `kensen.test` function is used to compute the trend test.

```
> setSweave("trend01", 5, 5)
> with(ConecuhFlows, timePlot(Year, Flow, Plot=list(what="both")))
> with(ConecuhFlows, refLine(horizontal=median(Flow)))
> # Required call to close PDF output graphics
> graphics.off()
```

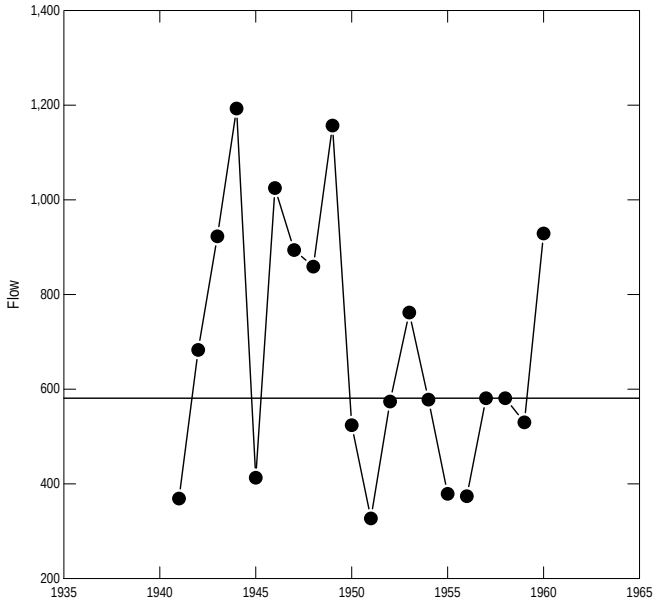

Figure 1. The time-series plot of annual mean flow in the Conecuh River at Brantley, AL.

The `kensen.test` function has three arguments—`y`, the response variable to test for trend; `t`, the time corresponding to each value in `y`; and `n.min`, the minimum number of observations required for adjusting for serial correlation. The default value for `n.min` is set at 10, in keeping with the 10-year minimum

requirement in the seasonal Kendall test (`seaken`), but a more practical limit seems to be 25, which is used in this example.

```
> # The Mann-Kendall trend test
> with(ConecuhFlows, kensen.test(Flow, Year, 25))

        Kendall's tau with the Sen slope estimator

data:  Flow and Year
tau = -0.12137, p-value = 0.4751
alternative hypothesis: true tau is not equal to 0
sample estimates:
      slope median.data median.time           n
     -8.875     581.000    1950.500      20.000
```

The attained p-value is greater than 0.05, so the null hypothesis of no trend is not rejected. There is no evidence of a significant downward trend in the annual mean flows from 1941 through 1960 based on these data.

The `serial.test` function can also be used to assess any effect from serial correlation. It may be more appropriate for shorter time periods that the adjustment in `kensen.test`. There are two options for the method used in `serial.test`. Both are shown below. Refer to the documentation for details about the methods.

```
> # The default method, Wilcoxon rank-sum serial test
> with(ConecuhFlows, serial.test(Flow))

        Wilcoxon test for serial dependence

data:  Flow
W = 54, p-value = 0.4967
alternative hypothesis: true lag-1 serial dependence is not equal to 0

> # The runs test method
> with(ConecuhFlows, serial.test(Flow, method="runs"))

        Runs test for serial dependence

data:  Flow
Runs = 12, p-value = 0.4698
alternative hypothesis: true lag-1 serial dependence is not equal to 0
```

3 Klamath River Data.

Helsel and Hirsch (2002) stress plotting the data to help understand the data and find an appropriate statistical technique. For these data, two graphs are created, one that displays the concentrations over time, and the second that displays the relation between flow and concentration.

```
> # setSweave is a specialized function that sets up the graphics page for
> # Sweave scripts. For interactive use, it should be removed and the
> # default setting for set.up can be used.
> setSweave("trend02", 5, 5)
> with(KlamathTP, timePlot(sample_dt, TP_ss, Plot=list(what="points")))
> # Required call to close PDF output graphics
> graphics.off()
```

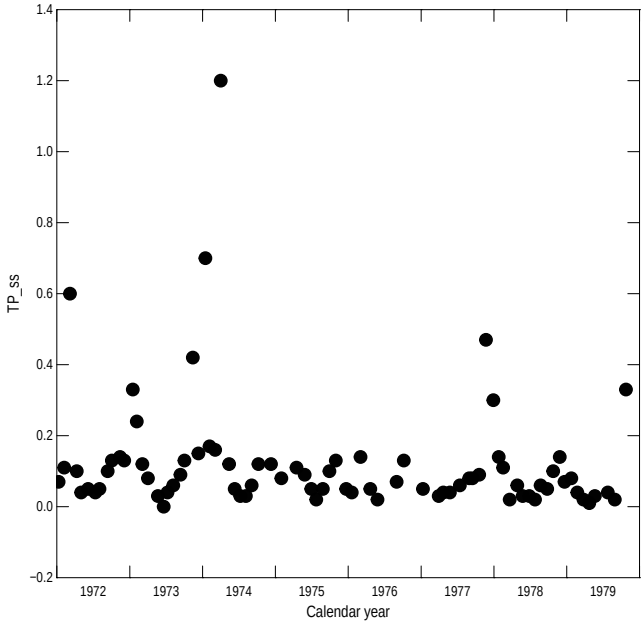

Figure 2. The time-series plot of total phosphorus concentration. Note the strong seasonal pattern.

```
> setSweave("trend03", 5, 5)
> with(KlamathTP, xyPlot(Flow, TP_ss))
```

```
> # Required call to close PDF output graphics
> graphics.off()
```

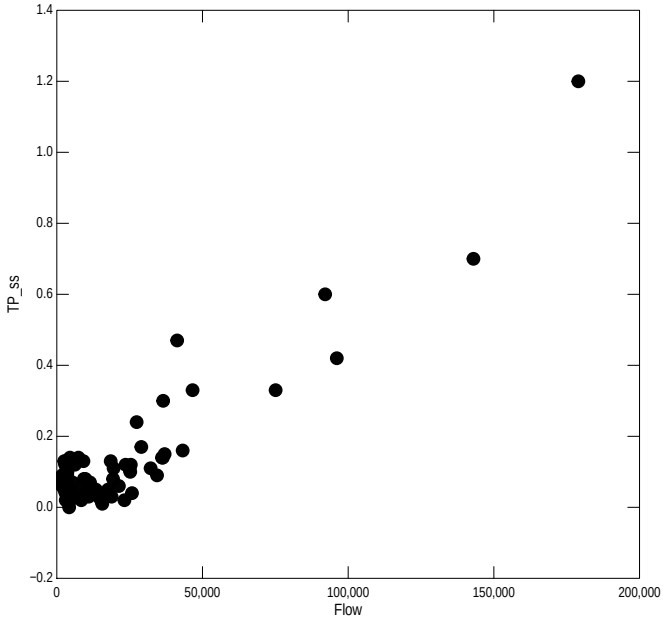

Figure 3. The relation between flow and total phosphorus concentration.

4 Seasonal Kendall Trend Tests

The `seaken` function performs the seasonal Kendall test described in section 12.4.1. It requires a complete, regular set of data for each season, with a missing value (NA) if no sample was taken during a season. That regular set can be created using the `regularSeries` function.

The first trend test is a simple test of the trend in concentration with no adjustment for flow. The first sample was taken in January 1972 and the last sample in the dataset was in October 1979. The analysis period will be from January 1, 1972 through December 31, 1979. Because of the way the computers store and interpret dates, the end date must be January 1, 1980.

```
> # Construct the regular series: 96 monthly observations
> KlamathTP.RS <- with(KlamathTP, regularSeries(TP_ss, sample_dt,
+                                                                                    begin="19
> # Do the analysis: Value is the name of the column in KlamathTP.RS
> # that contains the TP data
> with(KlamathTP.RS, seaken(Value, 12))

        Seasonal Kendall with correlation correction

data:  Value (8 years and 12 seasons)
tau = -0.2605, p-value = 0.005624
alternative hypothesis: true slope is not equal to 0
sample estimates:
       slope  median.data  median.time
-0.004999999  0.075000003  4.000000000
```

The attained p-value for the test is 0.005624, which is less than the critical alpha level set for the test, so we reject the null hypothesis of no trend and accept the alternate hypothesis of a trend, which is downward at a rate of about 0.005 mg/L/y. An estimate of the average trend in percent can be computed as 100 percent times the trend divided by the median concentration (median.data in the report): 100 * -0.005 / 0.075 = -6.67 percent per year.

Smith and others (1982) used linear regression to get the flow-adjusted concentrations. Helsel and Hirsch (2002) refer to this as a Mixed method for dealing with seasonality in table 12.3. The residuals from the regression analysis are used as the flow-adjusted concentrations.

```
> # Compute the flow-adjusted concentrations. Figure 2 serves as justification
> # for the linear fit for these data.
> KlamathTP$FAC <- residuals(lm(TP_ss ~ Flow, data=KlamathTP,
+         na.action=na.exclude)) # required to preserve any missing values
```

```
> # Construct the regular series: 96 monthly observations
> KlamathTP.RS2 <- with(KlamathTP, regularSeries(FAC, sample_dt,
+                                                                                    begin="19
> # Do the analysis: Value is the name of the column in KlamathTP.RS
> # that contains the FAC data
> KTP <- with(KlamathTP.RS2, seaken(Value, 12))
> print(KTP)

        Seasonal Kendall with correlation correction

data:  Value (8 years and 12 seasons)
tau = -0.071429, p-value = 0.4899
alternative hypothesis: true slope is not equal to 0
sample estimates:
       slope  median.data  median.time
-0.001932378 -0.005219474  4.000000000
```

The attained p-value for the test is 0.4899, which is greater than the critical alpha level set for the test, so we do not reject the null hypothesis of no trend. For these data, there is also a fair trend in decreasing flow, which can result in a less significant trend in the FAC than in the raw concentrations. The trend is about -0.0019 mg/L/y and the rate in percent would use the median concentration from the simple trend test in (0.075) and would be -2.6 percent per year.

5 Parametric Trend Tests

Section 12.4.3 in Helsel and Hirsch (2002) discusses the construction of a linear regression model with periodic functions. This example will demonstrate only the flow- and seasonally-adjusted trends. The regression for flow adjustment only requires a linear relation, but the parametric seasonal trend model requires that all assumptions of linear regression are met—linearity and the homoscedasticity, independence and normality of the residuals.

To maintain those assumptions of linear regression, the log transformation of concentration and flow are almost always required. The total phosphorus data are left censored, so Tobit regression should be used. However, simple substitution and ordinary least squares have been historically used for trend analyses for low rates of censoring. For example Schertz and others (1991) accepted about 5 percent censoring before using Tobit regression. Modern user interfaces to censored data make those techniques easy to use, so there is little reason not to use them for any censored data, but for this example of a single censored value, simple substitution will be used.

Smith and others (1982) used a value of 0 as the substitute value, but for a log transform, a different, positive, value must be used. This example will use one-half the detection limit. The first executable statement in the code below computes the substitute value and converts the sample date to decimal time format, which is easier to interpret as a annual rate. The linear regression trend test uses the `fourier` function (in package `smwrBase`) to construct the sine and cosine variables. This example uses the natural log transform because it is easier to express the result as a trend in percent per year than the common logarithm.

```
> # Simple substitution for one left-censored value
> KlamathTP <- transform(KlamathTP, TP_ss2 = ifelse(TP_rmk == "<", TP/2, TP),
+                                                                          D
> # The trend analysis, residual plot review indicates that
> # Flow must second-order
> KTP.lm <- lm(log(TP_ss2) ~ Dectime + quadratic(log(Flow)) + fourier(Dectime),
+                                                  data=KlamathTP)
> summary(KTP.lm)


Call:
lm(formula = log(TP_ss2) ~ Dectime + quadratic(log(Flow)) + fourier(Dectime),
    data = KlamathTP)

Residuals:
     Min       1Q   Median       3Q      Max
-1.89213 -0.17547  0.09404  0.31029  0.84872
```

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     123.18306   47.10131   2.615 0.010853 *
Dectime                          -0.06377    0.02384  -2.675 0.009237 **
quadratic(log(Flow))(9.44088)1   0.45121    0.07476   6.035 6.26e-08 ***
quadratic(log(Flow))(9.44088)2   0.25316    0.04342   5.831 1.45e-07 ***
fourier(Dectime)sin(k=1)        -0.42164    0.11274  -3.740 0.000367 ***
fourier(Dectime)cos(k=1)         0.47389    0.09289   5.101 2.63e-06 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 0.4869 on 72 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7557,        Adjusted R-squared:  0.7388
F-statistic: 44.55 on 5 and 72 DF,  p-value: < 2.2e-16
```

The attained p-value for the trend is about 0.0092, which is less than the critical alpha level set for the test. The value of the coefficient (-0.06377) can be used to estimate the trend in percent per year: 100 * (exp(-0.06377) - 1) = -6.18. That value, with the median concentration from the previous nonparametric analysis can be used to estimate the trend in mg/L: (-6.18 / 100) * 0.075 = -0.0046, approximately.

6 Regional Kendall Trend Tests

The seasonal Kendall trend test is easily modified to perform a regional Kendall trend test by substituting sites for seasons (Helsel and others, 2006). This example replicates example RK3b from Helsel and others (2006). The required format for the data for the `regken` is a matrix with rows representing the annual series of data and the columns representing the sites. This format facilitates managing the data and assures a complete record. The `regken` function has an additional argument `correct` that controls whether the adjustment for cross correlation is used or not.

```
> # Get the data from the downloads folder for the report
> RK3b <- read.delim("http://pubs.usgs.gov/sir/2005/5275/downloads/RK3b.txt",
+   header=FALSE, skip=1)
> names(RK3b) <- c("Year", "Site", "NH4")
> # Reformat to a wide data frame group2row is in smwrBase and make the matrix
> RK3b <- group2row(RK3b, "Year", "Site", "NH4")
> RK3b.mat <- data.matrix(RK3b[, -1]) # column 1 is Year
> # Preform the trend test
> regken(RK3b.mat)

        Regional Kendall trend test with correlation correction

data:  RK3b.mat (12 years and 25 sites)
tau = 0.28066, p-value = 5.766e-06
alternative hypothesis: true slope is not equal to 0
sample estimates:
      slope median.data median.time
       0.16        4.30        6.00

> regken(RK3b.mat, correct=FALSE)

        Regional Kendall trend test

data:  RK3b.mat (12 years and 25 sites)
tau = 0.28066, p-value < 2.2e-16
alternative hypothesis: true slope is not equal to 0
sample estimates:
      slope median.data median.time
       0.16        4.30        6.00
```

The p-values can be affected by cross correlation among the sites. The corrected and uncorrected p-values from this example are very much smaller than 0.05, so spatial correlation does not affect the conclusions from this

analysis, when the p-values are closer to the preselected criterion, the conclusions may be affected. The code below demonstrates a simple way to view the overall correlation structure of any data matrix. For these data, almost every site is correlated with many other sites at greater than the Spearman's rho of 0.5.

```
> # Pretty print the Spearman correlation
> printCor(cor(RK3b.mat, method="spearman", use="pair"), 0.5)
```

```
           XXXXXXXXXXXXXXXXXXXX
       XXXX11111122223333334445
       15671345891469012367914 91
       ........................
       NNNNNNNNNNNNNNNNNNNNNNNN
       HHHHHHHHHHHHHHHHHHHHHHHH
       44444444444444444444444444
 X1.NH4\+ ++ +     +++   +    +
 X5.NH4+\++  + + +  + ++ ++  + +
 X6.NH4 +\+ ++++ +  +++++++  +++
 X7.NH4+++\+ + + + ++++ +++ ++++
X11.NH4+  +\       + +  +     +
X13.NH4  +  \++ +     +   +   +
X14.NH4++++ +\++ ++ + ++ ++  + +
X15.NH4  +  ++\ ++ ++++++++  ++
X18.NH4 +++  + \ +  +  +     + +
X19.NH4     + + \ ++ ++ + +   +
X21.NH4 +++  +++ \  + +++++  + +
X24.NH4      +  + \    +
X26.NH4+  ++  + +  \ ++ ++    ++
X29.NH4++++  +++ +  \++++++  +++
X30.NH4+ +++  + +  ++\+ +++  +++
X31.NH4 +++ +++ ++ +++\ +++  +++
X32.NH4 ++   +++ ++ +  \ ++  + +
X33.NH4  +++  + ++ ++++ \++  +++
X36.NH4++++  ++  + ++++++\+  +++
X37.NH4 +++ +++ ++  ++++++\  + +
X39.NH4                    \+
X41.NH4   +                +\  +
X44.NH4 +++  +++ +  +++++++  \++
X49.NH4+ ++++ + +  ++++ ++   +\+
X51.NH4 +++  + + + ++++++++ +++\
```

References

[1] Helsel, D.R. and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p.

[2] Helsel, D.R., Mueller, D.K., and Slack, J.R., 2006, Computer program for the Kendall family of trend tests: U.S. Geological Survey Scientific Investigations Report 2005–5275, 4 p.

[3] Schertz, T.L., Alexander, R.B., and Ohe, D.J., 1991, The computer program estimate trend (ESTREND), a system for the detection of trends in water-quality data: U.S. Geological Survey Water-Resources Investigations Report 91-4040, 63 p.

[4] Smith, R.A., Hirsch, R.M, and Slack, J.R., 1982, A study of trends in total phosphorus measurements at NASQAN stations: U.S. Geological Circular 2190, 34 p. Available online at "pubs.usgs.gov/wsp/2190/report.pdf".