Regional Trend Analysis of Kjeldahl Nitrogen

Dave Lorenz

July 27, 2017

Abstract

This example illustrates the data manipulations for the regional seasonal Kendall analysis. The example uses kjeldahl nitrogen data. The regional seasonal Kendall analysis requires a common time frame for all of the trend tests.

Contents

1 Introduction

The data used in this application are a small subset of the data used by Schertz and others (1991). The data are samples taken from water year 1969 (October, 1968) through water year 1989 (September, 1989). Nineteen stations were selected and kjeldahl nitrogen selected from the original data.

```
> # Load restrend and other packages and the data
> library(restrend)
> library(smwrBase)
> library(smwrQW)
> data(EstrendSub)
> head(EstrendSub)

     STAID       DATES    QI QD RN.organic PN.organic RAmmonia
1 07227500 1968-10-01   7.6 NA                              NA
2 07227500 1968-10-03   5.3 NA                              NA
3 07227500 1968-10-16 532.0 NA                              NA
4 07227500 1968-10-19  17.0 NA                              NA
5 07227500 1968-11-01  17.0 NA                              NA
6 07227500 1968-12-01   6.6 NA                              NA
  PAmmonia RKjeldahl PKjeldahl RTotal.P PTotal.P RCopper PCopper
1       NA                  NA              NA              NA
2       NA                  NA              NA              NA
3       NA                  NA              NA              NA
4       NA                  NA              NA              NA
5       NA                  NA              NA              NA
6       NA                  NA              NA              NA
  RIron PIron Calcium Chloride
1          NA      95      280
2          NA      NA       NA
3          NA      42      106
4          NA     121      435
5          NA     150      512
6          NA     138      510
```

2 Summarize the Sample Data

In general, it is desirable, but not necessary, to subset the data before proceeding with the analysis of a subset of the constituents. Nutrient concentrations can vary widely in natural waters and can range from completely uncensored to highly censored. No flow adjustment will be done for these data becuase of the potential high variability in censoring–for these data, about 10 percent of ammonia samples are censored, but the other constituents have much lower percentages.

```
> # Convert to class qw
> EstrendSub.qw <- convert2qw(EstrendSub)
> # Create the subset
> KN <- subset(EstrendSub.qw, select=c("STAID", "DATES",
+   "PKjeldahl"))
> # Rename to remove leading P, not required--just pretty
> names(KN)[3] <- "Kjeldahl"
> # The sampling for nutrients started later, so remove the samples that
> # have no data.
> KN <- dropAllMissing(KN, "Kjeldahl")
```

The `sampReport` function creates a simple PDF file that contains a report of the sample date ranges and graph of samples for each station. It can be used to help define the starting and ending date ranges for the trend tests as well as identifying sample gaps and other sampling issues.

```
> # Create the report
> sampReport(KN, DATES="DATES", STAID="STAID", file="KNSampling")
```

```
    STAID  FirstSamp   LastSamp NumSamp
1  07227500 1974-05-01 1989-08-15      92
2  07228000 1974-04-30 1989-08-16     143
3  07297910 1974-10-24 1989-08-15     121
4  07299540 1978-01-11 1986-07-22      67
5  07300000 1974-10-25 1989-08-16      96
6  07308500 1974-10-15 1989-08-18     138
7  07316000 1976-10-19 1986-09-09      74
8  07331600 1974-10-15 1986-09-09     125
9  07336820 1974-04-17 1989-06-20     106
10 07342500 1979-12-06 1989-06-19      61
11 07343200 1974-04-17 1989-06-21      95
12 07343500 1982-11-17 1985-09-04      18
13 07344500 1982-11-18 1985-08-29      18
14 07346070 1974-04-17 1989-08-10      90
```
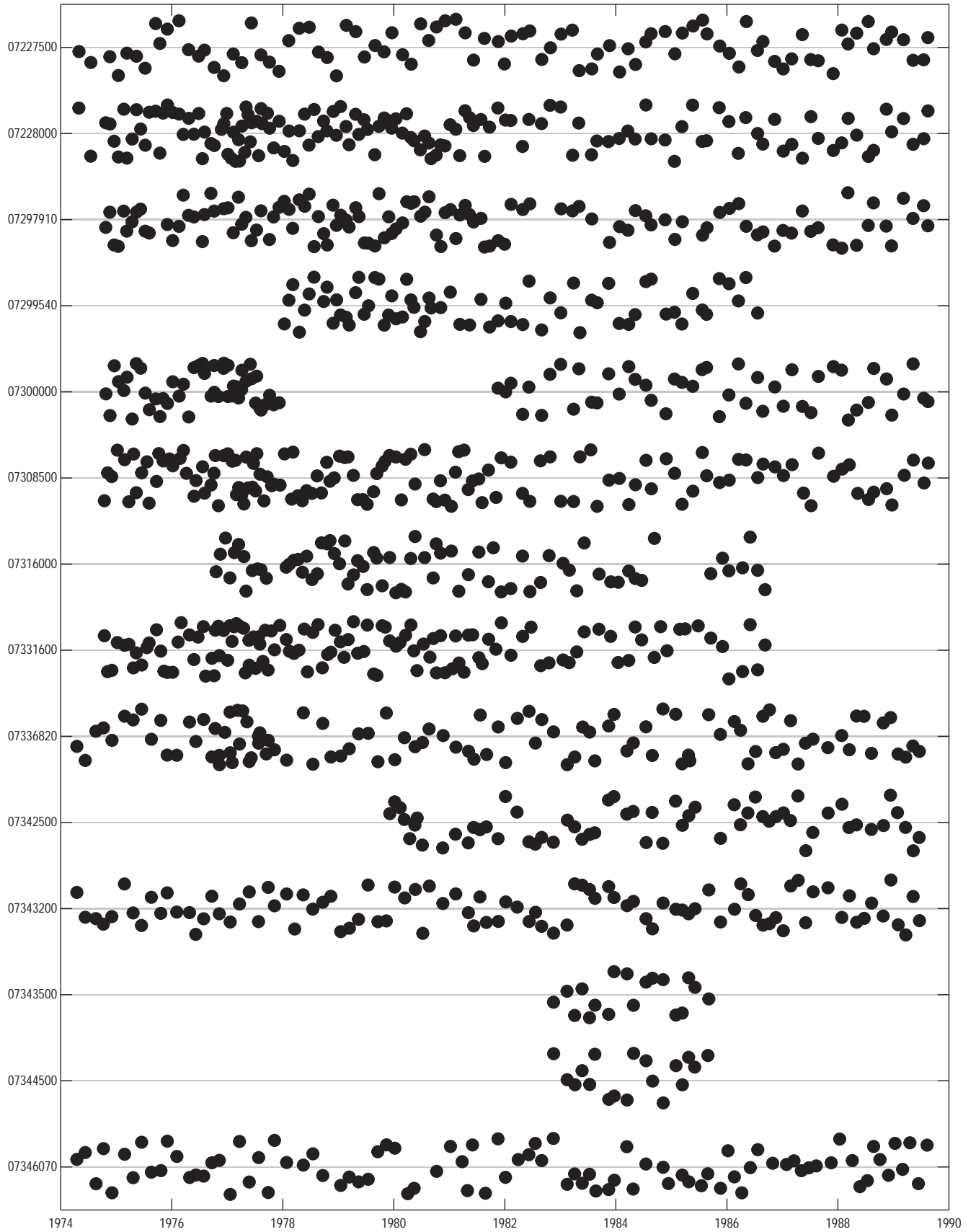
The call to `sampReport` returns the file name invisibly (KNSampling.pdf). Because it is a full-size portrait PDF file, it is inserted here with compressed pages. The report gives the actual begin and end dates for sampling and the graph shows the sampling dates for each station. It is easy to see that the pattern of sampling is very irregular from station to station. Note that this report only shows when a sample for Kjeldahl nitrogen was taken.

3   Set up the Project

The user must balance the need to include as many stations as possible and the targeted time frame for the trend estimation. For these data, most Kjeldahl nitrogen sampling started in late 1974 and extended through late 1989.

The (setProj) function sets up the trend estimation project. There are many arguments to (setProj), see the documentation for details. The constituent names or response variable names are referred to as `Snames` in keeping with the names used in the original ESTREND.

After projects have been set up, the user can get a list of the projects by using `lsProj` or can specify a project to use with `useProj`. The function `useProj` must be used to continue working on a project after the user quits from the R session.

```
> # Set up the project
> setProj("kn", KN, STAID="STAID", DATES="DATES",
+                                  Snames="Kjeldahl",
+                                  type="seasonal", Start="1974-10-01", End="1989-10-01")

[1] "kn"
```

It is useful to verify which stations and snames will be analyzed and what the seasonal definitions are. The user need only enter the name of the R data object in the console. For these data, the seasonal definition is 0 in all cases where the status is not "OK." The percentage of censoring for all stations is also displayed from the code following this paragraph. The censoring levels from station 07297910 is shown becuase that station has greater than 5 percent censoring.

```
> # Which are OK?
> estrend.st


         snames
stations  Kjeldahl
  07227500 "OK"
  07228000 "OK"
  07297910 "OK"
  07299540 "short record"
  07300000 "OK"
  07308500 "OK"
  07316000 "short record"
  07331600 "short record"
  07336820 "OK"
  07342500 "short record"
  07343200 "OK"
  07343500 "too few data"
  07344500 "too few data"
  07346070 "OK"

> # What seasonal definition?
> estrend.ss
```

```
          snames
stations   Kjeldahl
  07227500        6
  07228000        6
  07297910        6
  07299540        0
  07300000        6
  07308500        6
  07316000        0
  07331600        0
  07336820        6
  07342500        0
  07343200        6
  07343500        0
  07344500        0
  07346070        6

> # What about censoring percentage?
> estrend.cp


          snames
stations   Kjeldahl
  07227500 1.111111
  07228000 1.418440
  07297910 8.264463
  07299540 2.985075
  07300000 3.125000
  07308500 0.000000
  07316000 1.351351
  07331600 0.800000
  07336820 1.941748
  07342500 0.000000
  07343200 0.000000
  07343500 0.000000
  07344500 0.000000
  07346070 2.298851
```

4 Seasonal Kendall Trend Test

Eight stations "07227500," "07228000," "07297910," "07300000," "07308500," "07336820," "07343200," and "07346070" meet the criteria for analysis and each of those have the record for a 6-period seasonal Kendall analysis. But, station "07297910" has greater than 5 percent censoring and will be analyzed using the censored seasonal Kendall by default. There are 2 censoring levels for the data for that station, which complicates the analysis. If there were a single censoring level, then one could raise the percentage criterion for the uncensored test to 10 without substantially affecting the results(Helsel ref?), but that rule does not apply when there is more than one censoring level. The approach will be to accept the default of 5 percent censoring and rerun that station with the 10 percent criterion and compare the result. If they are similar, then accept the larger percent criterion and proceed. If they are substantially different, then that station will be dropped from the regional analysis.

The function SKTrends executes the trend test on all valid combinations of stations and snames. It can also execute the test on subsets if some changes need to be made, as shown in the code immediately following this paragraph. The SKTrends function creates a PDF file that contains the result of the analysis and a series graph on each page. The results for

```
> # Trend tests, accepting default seasons (6) and percent censoring (5)
> SKTrends()

[1] "kn_sk.pdf"


> # repeat for the more highly censored stations
> # The use of the log transform must be turned off to make it comparable
> SKTrends(Stations="07297910", max.cens=10, use.logs=FALSE,
+   report="kn_07297910")

[1] "kn_07297910.pdf"
```

07297910 Kjeldahl

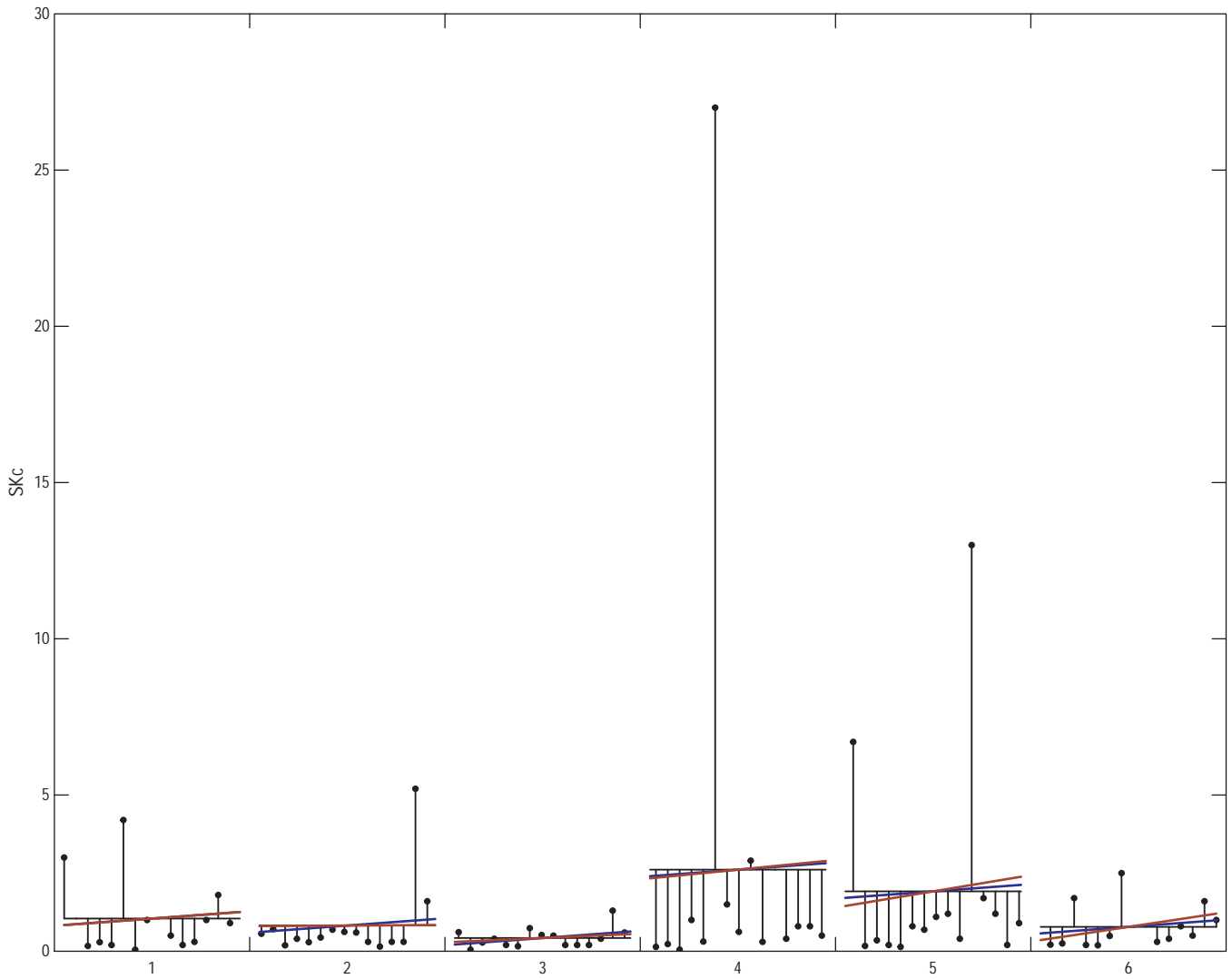 Seasonal Kendall's tau with the Turnbull slope estimator

data: Kjeldahl (15 years and 6 seasons)
tau = 0.20641, p-value = 0.07129
alternative hypothesis: true slope is not equal to 0
sample estimates:
      slope median data median time
      0.025   0.500   7.500

```
07297910  Kjeldahl

 Seasonal Kendall with correlation correction

data:  Kjeldahl (15 years and 6 seasons)
tau = 0.20463, p-value = 0.07337
alternative hypothesis: true slope is not equal to 0
sample estimates:
      slope median.data median.time
 0.02535714  0.50000000  7.50000000
```
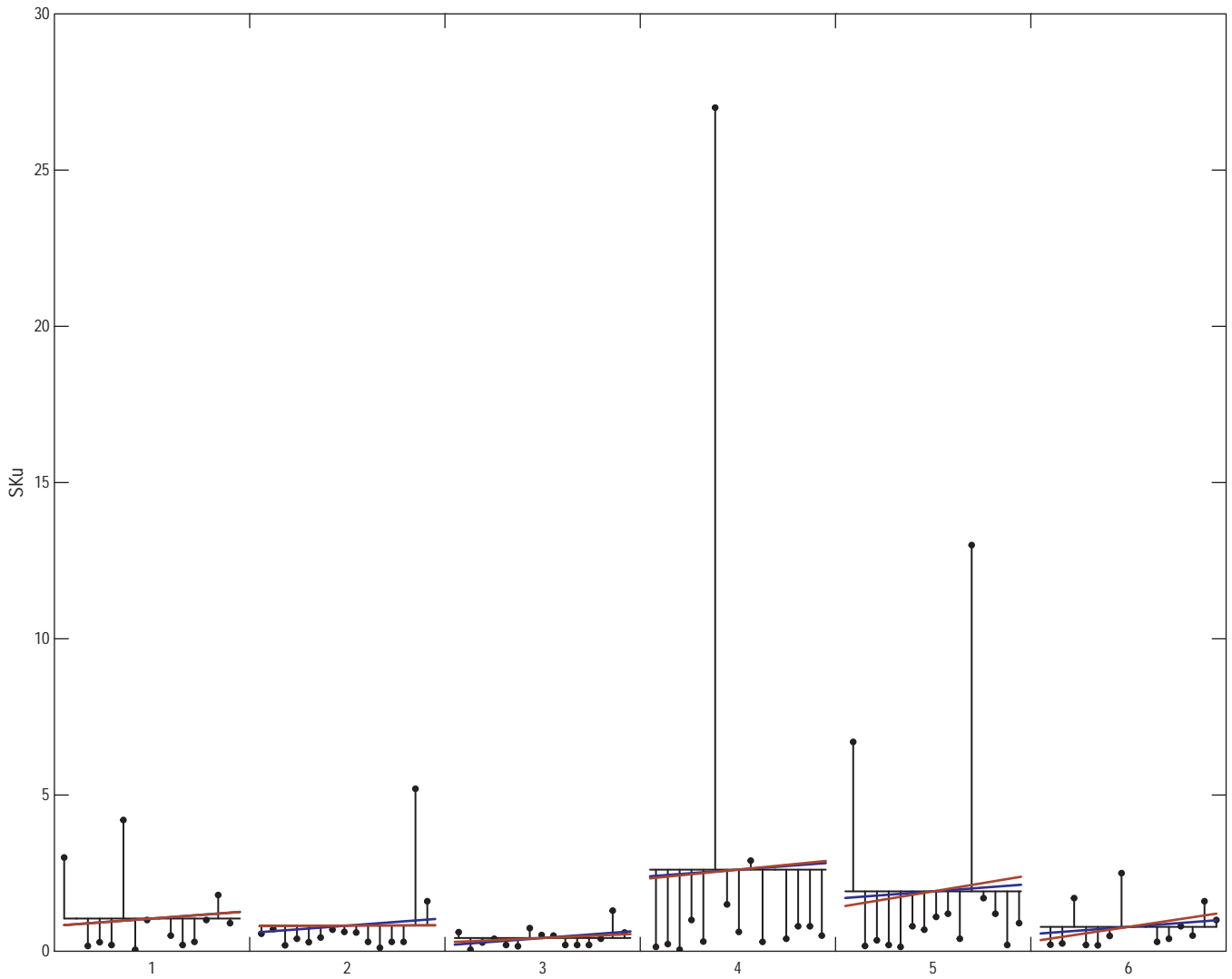
The results are very similar—the value for tau changed from 0.20641 to 0.20463 and the slope changed from 0.025 to 0.0254. Proceed with the analyis, but use the log transform to be consistent with the other stations.

```
> # The log transform is turned on to make it comparable to the other stations
> SKTrends(Stations="07297910", max.cens=10, use.logs=TRUE,
+   report="kn_07297910_log")


[1] "kn_07297910_log.pdf"
```

5   Regional Trend

When completed, or to check on intermediate results, the estimated trends can be extracted using the `getTrends` function. By default, all stations and snames are extracted. The output dataset is explained in the documentation for `getTrends`. The user has the option to set a significance level to determine whether there is a significant trend, the default level is 0.05. The code immediately following this paragraph illustrates the use of `getTrends`. For these Kjeldahl nitrogen stations, there are no significant trends at the alpha level of 0.05, but the trend is downward at most stations.

```
> # get the trends
> kn.tnd <- getTrends()
> print(kn.tnd)


   Station Response                         Type NumYears NumSeas
1 07227500 Kjeldahl uncensored seasonal Kendall       15       6
2 07228000 Kjeldahl uncensored seasonal Kendall       15       6
3 07297910 Kjeldahl uncensored seasonal Kendall       15       6
4 07300000 Kjeldahl uncensored seasonal Kendall       15       6
5 07308500 Kjeldahl uncensored seasonal Kendall       15       6
6 07336820 Kjeldahl uncensored seasonal Kendall       15       6
7 07343200 Kjeldahl uncensored seasonal Kendall       15       6
8 07346070 Kjeldahl uncensored seasonal Kendall       15       6
  Nobs RepValue        Trend  Trend.pct    P.value Trend.dir
1   89 1.100000 -0.080524358 -7.3203961 0.18233918      none
2   89 0.800000 -0.029281242 -3.6601553 0.06440882      none
3   85 0.500000  0.036219028  7.2438057 0.07336938      none
4   67 0.600000  0.000000000  0.0000000 0.87113285         *
5   89 1.300000 -0.028866231 -2.2204793 0.05688475      none
6   82 0.800000 -0.013989759 -1.7487198 0.25667536      none
7   84 0.964987  0.000000000  0.0000000 0.97513056         *
8   80 0.770000  0.005133306  0.6666631 0.53326058      none
```

The `RSKTrends` function can be used to assess any regional trend. By default, all stations are used, but subsets can be specified for specific regions within the larger study. The `RSKTrends` function peforms the analysis on a single constituent (Sname), which must be specified in the call. Its use is illustrated in the code immediately following this paragraph. Four p-values are reported—the raw p-value, the p-value corrected for serial correlation, and two p-values corrected for serial and

spatial correlation. The raw p-value is based on the data without any correction for serial or spatial correlation. The second p-value is corrected only for serial correlation, similar to the seasonal Kendall test. The corrected p-value for serial and spatial correlation is based on Dietz and Killeen (1981) and the alternative p-value is based on the methods described by Douglas and others (2000). Two methods for correcting the p-value for spatial correlation are provided because the correction based on Dietz and Killeen (1981) appears excessive when there are a few highly correlated stations.

```
> # Any regional trend?
> RSKTrends(kn.tnd$Station, "Kjeldahl")


        Regional Seasonal Kendall Test for Trend

data:  Region (8 sites, 15 years, and 6 seasons)

The value of S is -249.
The tau correlation coefficient is -0.05765.

The attained p-values:
                                            Raw = 0.048
               Corrected for serial correlation = 0.1672
   Corrected for serial and spatial correlation = 0.251
 Alternative for serial and spatial correlation = 0.2122

The median of the median slope at each site = -0.008821
```

For these data, the raw p-value is less than the 0.05 level used for the seaonal Kendall tests, but the p-values corrected for serial correlation and for both serial and spatial correlation are greater than 0.05, indicating that the null hypothesis of no regional trend cannot be rejected.

6 Further Remarks

Because trend analysis is not necessarily a straightforward process, but requires user assessments at several points in the process, it is not necessarily a good idea to simply create scripts and run them without any user review and interaction. To overcome recording the steps in a script, the functions in restrend record all changes to the projects database in a list called `estrend.cl`. It can be viewed at any time simply by entering estrend.cl in the console window. It can be saved with the data to ensure that the trend analysis is reproducible. Note that the use of the `getTrends` and `RSKTrends` functions is not recorded because they does not change anything recorded in the project database.

```
> # get the history
> estrend.cl

[[1]]
setProj(project = "kn", data = KN, STAID = "STAID", DATES = "DATES",
    Snames = "Kjeldahl", type = "seasonal", Start = "1974-10-01",
    End = "1989-10-01")

[[2]]
SKTrends()

[[3]]
SKTrends(Stations = "07297910", use.logs = FALSE, max.cens = 10,
    report = "kn_07297910")

[[4]]
SKTrends(Stations = "07297910", use.logs = TRUE, max.cens = 10,
    report = "kn_07297910_log")
```

References

[1] Dietz, E. J., and Killeen, T.J., 1981, A nonparametric multivariate test for monotone trend with pharmaceutical applications: Journal of the American Statistical Association, v. 76, p 169–174.

[2] Douglas, E.M., Vogel, R.M., and Kroll, C.N., 2000, Trends in floods and low flows in the United States: impact of spatial correlation: Journal of Hydrology, v. 240, p. 90–105.

[3] Lorenz, D.L., in preparation, restrend: an R package for EStimate TRENDs: U.S. Geological Survey Open File Report, ? p.

[4] 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, 72 p.