Trends in Lake Water Chemistry

Comparing Analyses in R Statistical Software and WQStat Plus Software

Julian A. Scott
Hydrologist, National Stream and Aquatic Ecology Center
April 2020

BACKGROUND

As recently as 2019, the US Forest Service has used the software WQStat Plus for analyzing lake water chemistry data. However, this software has recently become more expensive and difficult to acquire. The open-source software R Statistical Package can conduct many of the same analyses previously done in WQStat. To evaluate whether analyses conducted in R produce similar results to those done in WQStat, we compare the results of lake water chemistry trend tests in R to the results from WQStat. Specifically, we:

(1) use R to analyze lake water chemistry time series data from Emerald Lake and Florence Lake on the Bighorn National Forest and
(2) compare these R-derived results to the results of a WQStat analysis of same data, found in the “Long-Term Lakes Studies - Water Chemistry Review: Bighorn National Forest” report (hereafter, Bighorn Report) by the Forest Service National Groundwater Program (Gurrieri and Gazzetti 2019).

The statistical tests that are evaluated include:

- The Kruskal Wallis rank sum test to determine whether seasonal patterns exist (R function `kruskal.test` from the `stats` package (R 2019)).
- The Kendall Trend test, a nonparametric test for monotonic trend (R function `kendallTrendTest` in the `EnvStats` package (Millard 2013)).
- The Kendall Seasonal Trend test, a nonparametric test for monotonic trend within seasons (R function: `kendallSeasonalTrendTest` in the `EnvStats` package (Millard 2013)).

METHODS

Lake water chemistry data from Florence Lake and Emerald Lake were read into R as comma-delimited text files. For each lake, sampling occurred at least twice a year, at least once in June or July and at least once in August, September, October, or November (Table 1).

The Bighorn Report identified two seasons for which they carried out the Kendall Seasonal Trend test: pre-August 1st samples and post-August 1st samples. Thus, for the analysis in R, for each dataset a variable called 'Season' was created by reclassifying the sample date into two classes: Pre-August 1st and Post-August 1st (Table 2). In addition, a 'sample.date' variable was created by, for each sample date (e.g., 1994-08-10 and 2002-07-09), adding together the full year and the quotient equal to the day of year divided by 365 (e.g. 1994.608 and 2002.521). This was important for ensuring that trend values are in units/year, which matches the reported trends in Bighorn Report. In the R analysis, the value of any analyte with a reported value of 0 was set to 0.001.

Caveats

Note that in the absence of details about the exact methods used in the WQStat software and the fact that some statistical test values were not reported in the Bighorn Report, there are limits to the comparisons that can be made. For example, the Kendall Tau values are not reported in the Bighorn Report.

Table 1. For the R analysis, year and month of samples collected at Florence Lake (F; n = 57) and Emerald Lake (E; n = 53).

		Pre Aug 1	Pre Aug 1	Post Aug 1	Post Aug 1	Post Aug 1	Post Aug 1
Year #	Year	Jun	Jul	Aug	Sep	Oct	Nov
1	1993				F E		
2	1994		F E		F		
3	1995		F E		E		
4	1996	E	F	F E	F E		
5	1997	F	E	F E	F E		
6	1998	E	F	F E	F E		
7	1999	E	F	F E	E		
8	2000	F E		F E	E		
9	2001	F E	F		F E		
10	2002	F E	F E		F		
11	2003		F E	F	F E		
12	2004		F E	F E	F E		
13	2005		F E	F E	F E		
14	2006		F E		E		
15	2007	E	F	F E	F E		
16	2008		F E	F E	F E		
17	2009		F	F E	F E		
18	2010		F E	F E	F E		
19	2016		F E	F E	F		E
20	2017		F E	F E	E	F	
21	2018		F E	F E	F E		

Table 2. The specific R functions and function arguments used in this analysis.

1	`kruskal.test(formula = Analyte ~ Season, data = df)` The Kruskal Wallis rank sum test to determine whether seasonal patterns exist, where Analyte is the value of the analyte, Season is a categorical value indicating whether the sample was collected Pre or Post August 1st, and the data frame is the table containing the variables used in the formula.
2	`kendallTrendTest(formula = Analyte ~ sample.date, data = df)` The Kendall Trend test for monotonic trend, where sample.date is the sum of the year and the quotient equal to the day of year divided by 365.
3	`kendallSeasonalTrendTest(formula = Analyte ~ Season + Year, data = df)` The Kendall Seasonal Trend test for monotonic trend within each season, where Year is the year of the sample.date.

RESULTS

The packages and functions used in R appear to produce nearly identical results to the WQStat software. To facilitate the comparison, Tables 2 and 3 in the Bighorn Report are reproduced here as Table 3 and 4, respectively. Results from the R analysis are added to these tables in the same format for comparison.

From these tables, we see that the seasonality test (Kruskal Wallis) for both software is in perfect agreement for both lakes (Tables 3 and 4). The Kendall Seasonal Trend test also agree for the two software, except for Florence Lake Cl and K and Emerald Lake NO3. For Florence Lake, in the Bighorn Report, trend directions for Cl and K are reported as significant at the 85% level (Table 3). Its inferred that at the 95% level the trends are not significant. In the R analysis, significance is restricted to the 95% level and neither analyte has a significant trend at this level (Table 3). For Emerald Lake, the Bighorn Report notes that while NO3 has a significant downward trend (95% CI) it has a slope of 0, calling into question the WQStat trend (Table 4). The R analysis supports this interpretation and does not find a significant trend. For the Kendall Trend Tests, the two software agree for both lakes (Tables 3 and 4).

Table 3. Florence Lake trend analysis summary. Blue shading represents significant upward trends and orange shading represents significant downward trends (significance of at least the 95% confidence level). See Table 5 for reported test statistics used in determining shading. This table is a reproduction of Table 2 in the Bighorn NF Report (WQStat Analysis), with the same format applied to R analysis results.

	WQStat Analysis						R Analysis					
Measure	Seasonality (??% CI)	Trend Kendall Seasonal (95% CI)	Slope (units/year) Kendall Seasonal (95% CI)	Trend Kendall (99% CI)	Slope (units/year) Kendall (99% CI)	Comments	Seasonality (99% CI)	Trend Kendall Seasonal (95% CI)	Slope (units/year) Kendall Seasonal (95% CI)	Kendall Trend (99% CI)	Slope (units/year) Kendall (99% CI)	Comments
ANC	yes	up	0.545	up	0.680		yes	up	0.544	up	0.681	
pH	yes	up	0.027	up	0.029		yes	up	0.027	up	0.029	
Cond	no	up	0.147	up	0.153		no	up	0.145	up	0.150	
CaDis	no	up	0.019	up	0.020		no	up	0.019	up	0.020	
ClDis	no	down^	-0.0004^	no			no	no		no		
KDis	no	up^	0.0011^	no		Non monotonic	no	no		no		
MgDis	no	up	0.002	up	0.002		no	up	0.002	up	0.002	
NaDis	no	up	0.004	up	0.005		no	up	0.004	up	0.005	
NH4	no	no		**			no	no		no		
NO3	yes	no		**			yes	no		no		
SO4	no	up	0.009	no			no	up	0.009	no		

** Values not reported, ^ Significant at the 85% confidence level.

Table 4. Emerald Lake trend analysis summary. Blue shading represents significant upward trends and orange shading represents significant downward trends (significance of at least the 95% confidence level). See Table 5 for reported test statistics used in determining shading. This table is a reproduction of Table 3 in the Bighorn NF Report (WQStat Analysis), with the same format applied to R analysis results.

	WQStat Analysis						R Analysis					
Measure	Seasonality (??% CI)	Trend Kendall Seasonal (95% CI)	Slope (units/year) Kendall Seasonal (95% CI)	Trend Kendall (99% CI)	Slope (units/year) Kendall (99% CI)	Comments	Seasonality (99% CI)	Trend Kendall Seasonal (95% CI)	Slope (units/year) Kendall Seasonal (95% CI)	Kendall Trend (99% CI)	Slope (units/year) Kendall (99% CI)	Comments
ANC	yes	up	0.503	no			yes	up	0.500	no		
pH	no	up	0.027	up	0.027		no	up	0.027	up	0.027	
Cond	no	up	0.116	up	0.125		no	up	0.115	up	0.123	
CaDis	no	up	0.016	up	0.016		no	up	0.016	up	0.016	
ClDis	no	no		**			no	no		no		
KDis	no	up	0.002	no		Non monotonic	no	up	0.002	no		
MgDis	no	up	0.002	up	0.002		no	up	0.002	up	0.002	
NaDis	yes	no		**			yes	no		no		
NH4	yes	no		**			yes	no		no		
NO3	no	down	0.000	no		Slope of 0 suggests trend may not exist	no	no		no		Result confirms WQStat comment
SO4	no	no		**			no	no		no		

**Values not reported

In addition to reproducing the two result tables from the Bighorn Report, we also tabulate test statistic data for the Kendall Seasonal Trend and Kendall Trend tests that are found in Appendices A and B of the report. Specifically, the Z test-statistic and slope values for the Kendall Seasonal Trend and Kendall Trend tests for both lakes from the WQStat analysis for all analytes are provided in Table 5 and 6.

These tables also contain the corresponding test results from the R analysis, plus the Chi-Sq value and p-value for the Kruskal Wallis tests and Kendall Tau and p-values. These additional values are not found in the Bighorn report. For the four test statistics that are present for both software, a one-to-one comparison is made for all analytes in Figures 1 and 2. For these scatter plots, the points represent the given trend statistic for each analyte. These figures demonstrate that there is good agreement in the test results between the two software, as the points lie on the 1:1 line, with few exceptions.

Table 5. Florence Lake test statistics reported in the Bighorn Report (WQStat Analysis) compared to the R Analysis. Note, the Kendall Z column for the WQStat Analysis was listed as the 'Mann Kendall normal approx' in the Bighorn Report (e.g., see Appendix A Sen's Slope Estimator scatter plots) - this was assumed to be the Z value for the test.

	WQStat Analysis										R Analysis									
Measure	Kruskall Wallis Chi-Sq	Kruskall Wallis p-value	Kendall Seasonal Z	Slope (units/year) Kendall Seasonal (95% CI)	Kendall Seasonal Z p-value	Kendall Seasonal Tau	Kendall Z	Slope (units/year) Kendall (99% CI)	Kendall Z p-value	Kendall Tau	Kruskall Wallis Chi-Sq	Kruskall Wallis p-value	Kendall Seasonal Z	Slope (units/year) Kendall Seasonal (95% CI)	Kendall Seasonal Z p-value	Kendall Seasonal Tau	Kendall Z	Kendall Slope (99% CI)	Kendall Z p-value	Kendall Tau
ANC	**	**	3.198	0.545	**	**	3.490	0.680	**	**	8.668	0.003	3.198	0.544	0.001	0.304	3.490	0.681	0.000	0.318
pH	**	**	4.912	0.027	**	**	4.966	0.029	**	**	6.362	0.012	4.912	0.027	0.000	0.452	4.966	0.029	0.000	0.452
Cond	**	**	4.390	0.147	**	**	4.537	0.153	**	**	0.921	0.337	4.338	0.145	0.000	0.403	4.482	0.150	0.000	0.409
CaDis	**	**	4.157	0.019	**	**	4.572	0.020	**	**	1.736	0.188	4.157	0.019	0.000	0.400	4.579	0.020	0.000	0.417
ClDis	**	**	-1.622	-0.0004^	**	**	-1.842	-0.001	**	**	3.691	0.055	-1.763	-0.001	0.078	-0.155	-1.916	-0.001	0.055	-0.175
KDis	**	**	1.388	0.0011^	**	**	1.606	0.001	**	**	0.288	0.591	1.270	0.001	0.204	0.114	1.509	0.001	0.131	0.138
MgDis	**	**	3.894	0.002	**	**	4.272	0.002	**	**	1.828	0.176	3.879	0.002	0.000	0.367	4.200	0.002	0.000	0.382
NaDis	**	**	2.747	0.004	**	**	2.941	0.005	**	**	2.414	0.120	2.772	0.004	0.006	0.256	2.982	0.005	0.003	0.272
NH4	**	**	1.053	0.000	**	**	1.334	0.000	**	**	1.267	0.260	1.356	0.000	0.175	0.145	1.684	0.000	0.092	0.147
NO3	**	**	0.816	0.005	**	**	0.220	0.002	**	**	12.865	0.000	0.842	0.004	0.400	0.061	0.234	0.002	0.815	0.022
SO4	**	**	2.033	0.009	**	**	2.306	0.010	**	**	1.224	0.269	2.020	0.009	0.043	0.186	2.299	0.010	0.021	0.210
H.*	**	**	-4.860	-0.013	**	**	-4.931	-0.014	**	**	6.401	0.011	-4.860	-0.013	0.000	-0.448	-4.931	-0.014	0.000	-0.449

*The H analyte for WQStat does not have shading because it was not included in Table 2 in the Bighorn Report.
**Values not reported

Table 6. Emerald Lake test statistics reported in the Bighorn Report (WQStat Analysis) compared to the R Analysis. Note, the Kendall Z column for the WQStat Analysis was listed as the 'Mann Kendall normal approx' in the Bighorn Report (e.g., see Appendix A Sen's Slope Estimator scatter plots) - this was assumed to be the Z value for the test.

	WQStat Analysis										R Analysis									
Measure	Kruskall Wallis Chi-Sq	Kruskall Wallis p-value	Kendall Seasonal Z	Slope (units/year) Kendall Seasonal (95% CI)	Kendall Seasonal Z p-value	Kendall Seasonal Tau	Kendall Z	Slope (units/year) Kendall (99% CI)	Kendall Z p-value	Kendall Tau	Kruskall Wallis Chi-Sq	Kruskall Wallis p-value	Kendall Seasonal Z	Slope (units/year) Kendall Seasonal (95% CI)	Kendall Seasonal Z p-value	Kendall Seasonal Tau	Kendall Z	Kendall Slope (99% CI)	Kendall Z p-value	Kendall Tau
ANC	**	**	2.381	0.503	**	**	2.432	0.558	**	**	13.876	0.000	2.381	0.500	0.017	0.251	2.432	0.558	0.015	0.231
pH	**	**	4.483	0.027	**	**	4.670	0.027	**	**	1.213	0.271	4.483	0.027	0.000	0.454	4.674	0.027	0.000	0.443
Cond	**	**	3.894	0.116	**	**	4.435	0.125	**	**	3.334	0.068	3.922	0.115	0.000	0.399	4.419	0.123	0.000	0.419
CaDis	**	**	4.134	0.016	**	**	4.459	0.016	**	**	0.066	0.797	4.134	0.016	0.000	0.398	4.459	0.016	0.000	0.422
ClDis	**	**	0.379	0.000	**	**	0.123	0	**	**	0.655	0.418	0.407	0.000	0.684	0.032	0.154	0.000	0.878	0.015
KDis	**	**	2.453	0.002	**	**	2.049	0.002	**	**	0.526	0.468	2.565	0.002	0.010	0.235	2.103	0.002	0.035	0.200
MgDis	**	**	2.610	0.002	**	**	2.796	0.002	**	**	0.115	0.734	2.580	0.002	0.010	0.235	2.772	0.002	0.006	0.263
NaDis	**	**	0.827	0.001	**	**	1.451	0.002	**	**	13.412	0.000	0.827	0.001	0.408	0.084	1.466	0.002	0.143	0.139
NH4	**	**	0.144	0	**	**	0.892	0	**	**	6.533	0.011	0.346	0.000	0.729	0.076	1.340	0.000	0.180	0.125
NO3	**	**	-1.923	0	**	**	-1.985	0	**	**	1.162	0.281	-1.804	0.000	0.071	-0.173	-1.863	0.000	0.063	-0.167
SO4	**	**	0.756	0.000	**	**	1.059	0.003	**	**	1.423	0.233	0.742	0.002	0.458	0.050	1.059	0.003	0.290	0.101
H.*	**	**	-4.428	-0.010	**	**	-4.650	-0.010	**	**	1.274	0.259	-4.456	-0.010	0.000	-0.453	-4.666	-0.010	0.000	-0.442

*The H analyte for WQStat does not have shading because it was not included in Table 3 in the Bighorn Report.
**Values not reported

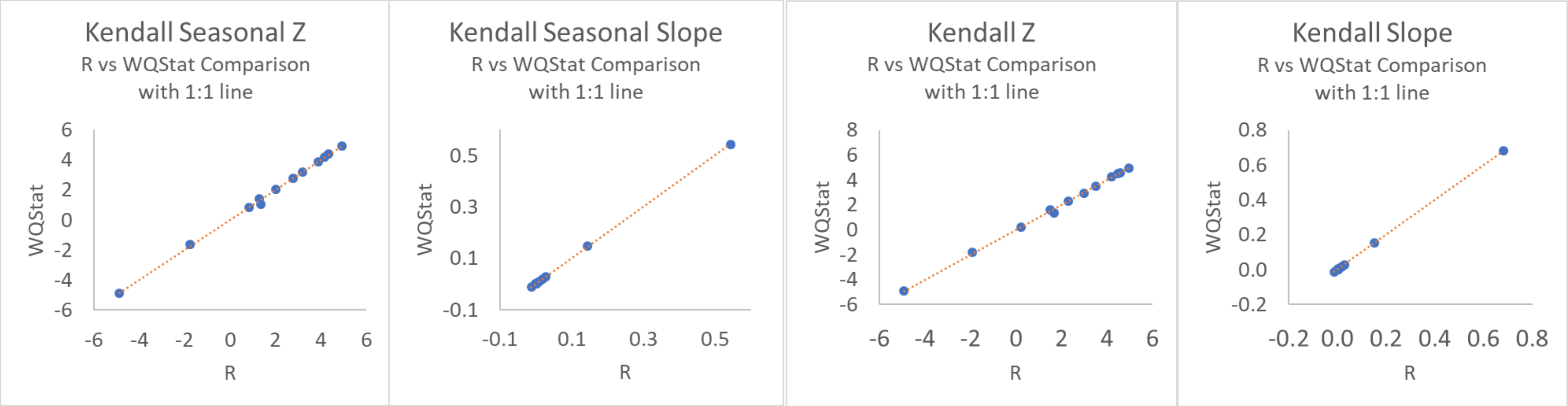

Figure 1. Comparison of four trend statistics for Florence Lake from analyses done in R (x axis) and WQStat (y axis). Points on the 1:1 line indicate perfect agreement in the test results between the two software. Data can be found in Table 3 and Table 5.

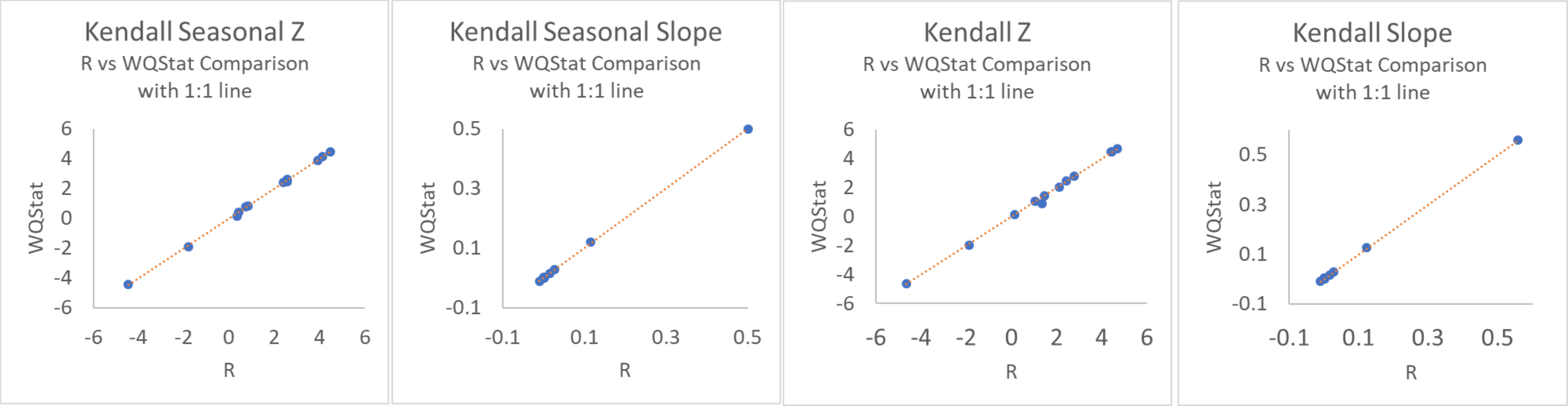

Figure 2. Comparison of four trend statistics for Emerald Lake from analyses done in R (x axis) and WQStat (y axis). Points on the 1:1 line indicate perfect agreement in the test results between the two software. Data can be found in Table 4 and Table 6.

CONCLUSION

Carrying out trend analysis in R is an excellent alternative to the WQStat software. This work shows that analysis done on the same datasets produce nearly identical results. Because the exact settings and arguments for the statistical test done in WQStat are not provided in the Bighorn Report, it is assumed that the minor differences between results are likely due to small, unknown differences in methodology.

The R packages and functions used in this statistical analysis are well documented and the results are highly reproduceable. R has extensive options for plotting trend results. In addition, other sophisticated trend analysis options are widely available in R, including Generalized Additive Models.

Please contact the author or the National Stream and Aquatic Ecology Center for additional details on this analysis, including the R code and datasets.

REFERENCES

Gurrieri, J. and E. Gazzetti. 2019. Long-Term Lakes Studies - Water Chemistry Review Bighorn National Forest. Forest Service, National Groundwater Program.

Millard SP. 2013. EnvStats: An R Package for Environmental Statistics. Springer, New York. ISBN 978-1-4614-8455-4

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.