Donatto Surratt, Ecologist, NPS
Michael Waldon, Hydrologist, USFWS

ARM Loxahatchee National Wildlife Refuge Total Phosphorus Outlier Analysis and Proposed Alternative Screening Criterion

February 22, 2012

Background

During the November 2011 Technical Oversight Committee (TOC) meeting under the topic of Data Usability, there was a discussion of the State's proposed method for classifying outliers. Under this proposal, all total phosphorus (TP) data that exceed three standard deviations (three-sigma) above the mean would be flagged as outliers. Mike Waldon speculated that this proposal would classify an inappropriately large amount of data as outliers because of the skewed frequency distribution of TP concentrations. State TOC representatives requested an example of how their proposed outlier analysis would result in identifying more data as outliers than expected based on a normally distributed dataset. Further, they requested that, if their method did prove to be problematic, an alternative proposal be presented.

Here, we apply the State-proposed approach to outlier classification using TP data collected at the 14 interior water quality monitoring stations used to assess Consent Decree water quality compliance. Detection limits for TP preclude classification of exceptionally small TP concentration values as outliers. Therefore, the three-sigma method is only applicable to classification of high values as outliers.

When applying the three-sigma method, data used in the evaluation must display an approximately normal distribution, particularly in the tails of the frequency distribution. Here, we demonstrate that normal distribution is not the case for non-transformed TP data.

Environmental monitoring data are often found to have an approximate log-normal distribution. As an alternative to testing the non-transformed data, we applied a logarithmic transformation to the data prior to use of the three-sigma method. Based on these results, we propose an alternative approach to identifying outliers in Refuge TP data using an adjusted sigma criterion applied to log-transformed values applied to each site independent of data from other Refuge sites.

Methods

Data source. DBHYDRO was accessed to download TP data for the 14 water quality monitoring stations. The complete period of record (September 1993 through October 2011) was downloaded and used for this analysis. Flagged data from May and June 2005 were included in this analysis.

Statistical approaches. Applying the three-sigma method for a normally distributed dataset should result in no more than 1.35 samples in 1,000 (0.135%) being identified as outliers in a one-tail test. That is, the cumulative standard normal distribution evaluated with a z-score of three has a value of 0.99865, or 1 – 0.00135. We test the three-sigma method on non-transformed and log-transformed data to determine if this assumption is upheld when applied to Refuge TP data. In this assessment, we apply the three-sigma method to individual sampling stations, and then aggregate the total number of samples and total number of classified outliers from all stations to provide an estimate of the fraction of samples that are identified as outliers under this approach. Further, we calculate the number of standard deviations that would result in an outlier classification frequency of 0.135%.

Results

Three-sigma for non-transformed data. Based on analysis of the non-transformed data, 38 out of the total 2,512 samples (1.513%) were identified as outliers using the three-sigma method (Table 1). The frequency of outlier identification was 11.2 times higher than expected for normally distributed data. Exclusion of the 24 May and June 2005 flagged TP data results in a 1.568% frequency for classifying data as outliers (Table 2), which is 11.6 times higher than expected for normally distributed data. A plot of TP concentrations versus percentiles shows that the data are approximately log-normally distributed within the central portion of their distributions, but are more skewed than log-normal at percentiles above 93% (Figure 1).

Three-sigma for log-transformed data. Based on analysis of the log-transformed data, 21 out of the total 2,512 samples (0.836%) were identified as outliers using the three-sigma method (Table 1). This frequency of outlier identification is 6.2 times higher than expected for normally distributed values. Exclusion of the 24 May and June 2005 flagged TP data results in a 0.643% frequency for classifying data as outliers (Table 2), which is 4.8 times higher than expected for normally distributed data.

Total number of standard deviations necessary to reduce frequency of identifying outliers to 0.135%. For the case with non-transformed data, the number of standard deviations necessary to decrease the frequency for classifying data as outlier down to or below 0.135% was 9.6 for a frequency of 0.119%. For the case where May and June 2005 flagged data were not included in the analysis and the data were non-transformed, the number of standard deviations necessary to decrease the frequency for classifying outliers down to or below 0.135% is 8.44 for a frequency of 0.121%. For the case with log-transformed data, the number of standard deviations necessary to decrease the frequency for classifying data as outliers down to or below 0.135% is 4.6, which is a 0.119% frequency of classifying data as outliers. When excluding the May and June 2005 flagged data for the log-transformed data, the number of standard deviations necessary was 4.1 for a frequency of 0.121%.

Discussion

Limitation of applying three-sigma method to non-transformed data. The use of non-transformed data for identifying outliers is not appropriate as it violates a fundamental assumption (data normality) for the test and substantially over-predicts the frequency of identifying outliers. The TP data collected in the Refuge are mostly log-normally distributed, except at the higher concentrations observed at or above the 93rd percentile of the data distribution. The log-normal distribution is slightly violated when concentrations are above 22 parts per billion, but log-transforming the data gives it a better approximation to a normal distribution, which is required to apply the three-sigma method. Even with the data transformation, applying the three-sigma method to the log transformed data identified more outliers than expected with or without the flagged May and June 2005 data. As such, application of the three-sigma method to the Refuge data is not acceptable because the frequency of classifying outlier is too high.

Federal proposed method for identifying outliers. In consideration of the State's outlier analysis proposal, we believe that the best approach for identifying outliers is to use a 4.1- sigma method applied to log-transformed TP data for the dataset that does not include the flagged May and June 2005 data. This method would apply to individual sites and not to the aggregated data from all sites, and we believe that it will be rigorous enough to provide protection from over-predicting outliers in the TP dataset. Regardless of the test selected, the next step following outlier identification must be an expert review of the data and the outliers in questions by the full TOC. These data must be assessed to determine if there are other parameters (i.e., meteorological, hydrological, ecological, etc.) that support identifying the sample as an outlier or explain why the data should not be considered an outlier.

Table 1. Summary statistics and outlier identification thresholds (cutoffs) for TP data collected on the Refuge. All data, including the May and June 2005 flagged data, are included in this dataset.

	LOX10	LOX11	LOX12	LOX13	LOX14	LOX15	LOX16	LOX3	LOX4	LOX5	LOX6	LOX7	LOX8	LOX9	Total
mean	9.1	9.1	8.2	8.6	7.8	7.4	8.9	10.5	11.1	9.9	7.8	9.3	9.6	8.4	
stdev	4.8	6.0	4.6	4.2	3.1	3.4	6.2	6.7	6.8	7.3	5.2	7.9	5.3	4.5	
3stdev	14.3	18.0	13.8	12.6	9.3	10.3	18.7	20.1	20.5	22.0	15.7	23.8	15.9	13.4	
cutoff	23.4	27.0	22.0	21.2	17.2	17.7	27.6	30.6	31.6	32.0	23.5	33.1	25.5	21.8	
#outlier	2.0	2.0	2.0	3.0	4.0	5.0	3.0	3.0	3.0	1.0	2.0	1.0	3.0	4.0	38
#normal	146	194	211	183	202	204	198	115	150	135	190	195	199	152	2474
total	148	196	213	186	206	209	201	118	153	136	192	196	202	156	2512
%outlier	1.351%	1.020%	0.939%	1.613%	1.942%	2.392%	1.493%	2.542%	1.961%	0.735%	1.042%	0.510%	1.485%	2.564%	1.513%
Log trasnformed															
mean	2.1	2.1	2.0	2.1	2.0	1.9	2.1	2.2	2.3	2.2	1.9	2.1	2.2	2.0	
stdev	0.4	0.5	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.5	0.5	0.4	0.5	0.5	
3stdev	1.3	1.4	1.3	1.3	1.2	1.2	1.2	1.5	1.3	1.5	1.5	1.3	1.4	1.4	
cutoff	30.2	31.3	26.8	29.7	23.8	23.6	27.7	40.3	37.6	37.5	29.4	30.3	34.4	29.7	
#outlier	1	2	2	1	0	1	3	1	3	1	2	1	2	1	21
#normal	147	194	211	185	206	208	198	117	150	135	190	195	200	155	2491
total	148	196	213	186	206	209	201	118	153	136	192	196	202	156	2512
%outlier	0.676%	1.020%	0.939%	0.538%	0.000%	0.478%	1.493%	0.847%	1.961%	0.735%	1.042%	0.510%	0.990%	0.641%	0.836%

Table 2. Summary statistics and outlier identification thresholds (cutoffs) for TP data collected on the Refuge. This dataset excludes the May and June 2005 flagged data.

	LOX10	LOX11	LOX12	LOX13	LOX14	LOX15	LOX16	LOX3	LOX4	LOX5	LOX6	LOX7	LOX8	LOX9	Total
mean	9.0	8.6	8.2	8.6	7.8	7.4	8.7	10.4	10.8	9.8	7.5	8.7	9.4	8.3	
stdev	4.6	3.7	4.6	4.2	3.0	3.4	5.9	6.6	6.3	7.2	4.3	3.5	4.6	4.2	
3stdev	13.7	11.0	13.8	12.6	9.0	10.3	17.7	19.8	19.0	21.7	12.9	10.4	13.8	12.7	
cutoff	22.6	19.6	22.1	21.1	16.8	17.8	26.4	30.2	29.9	31.5	20.4	19.1	23.2	20.9	
#outlier	2.0	4.0	2.0	3.0	4.0	5.0	2.0	3.0	3.0	1.0	2.0	2.0	3.0	3.0	39
#normal	145	190	209	181	200	202	197	114	148	134	188	192	197	152	2449
total	147	194	211	184	204	207	199	117	151	135	190	194	200	155	2488
%outlier	1.361%	2.062%	0.948%	1.630%	1.961%	2.415%	1.005%	2.564%	1.987%	0.741%	1.053%	1.031%	1.500%	1.935%	1.568%
Log trasnformed															
mean	2.1	2.1	2.0	2.1	2.0	1.9	2.1	2.2	2.3	2.2	1.9	2.1	2.1	2.0	
stdev	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.5	0.4	0.5	0.5	0.4	0.4	0.4	
3stdev	1.3	1.2	1.3	1.3	1.2	1.2	1.2	1.5	1.3	1.4	1.4	1.2	1.3	1.3	
cutoff	29.1	27.3	27.0	29.5	23.2	23.7	26.3	39.4	35.4	36.4	27.2	25.9	32.4	28.6	
#outlier	1	0	2	1	0	1	2	2	2	1	1	1	1	1	16
#normal	146	194	209	183	204	206	197	115	149	134	189	193	199	154	2472
total	147	194	211	184	204	207	199	117	151	135	190	194	200	155	2488
%outlier	0.680%	0.000%	0.948%	0.543%	0.000%	0.483%	1.005%	1.709%	1.325%	0.741%	0.526%	0.515%	0.500%	0.645%	0.643%

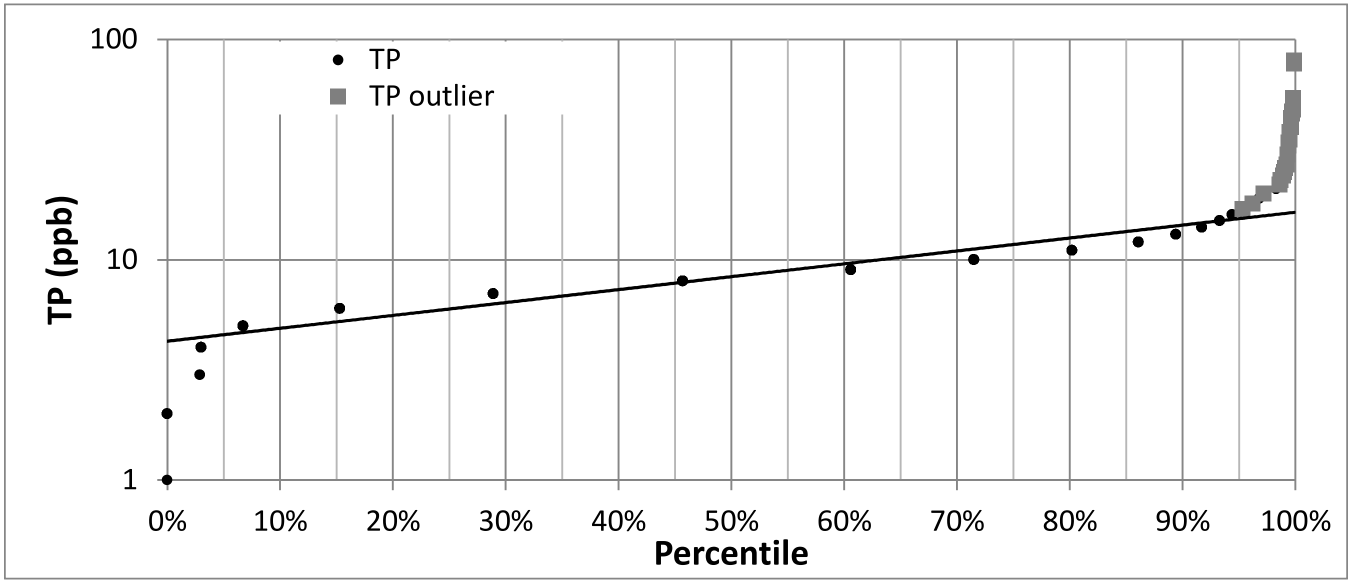

Figure 1. Total phosphorus (TP) concentrations and percentile distribution for data collected at the ARM Loxahatchee National Wildlife Refuge. Black circles represent TP concentrations and grey squares represent TP concentrations identified as outliers under the state proposed three-sigma rule. The black line is a simple linear trend line for comparison against the TP percentile distribution.