'On-the-Fly' Goodness of Fit and Outlier Testing for Left-Censored Data

Kirk Cameron, Ph.D.*

Abstract

A longstanding conundrum in environmental data with multiply left-censored measurements (i.e., non-detects) is how to best fit parametric data models. The magnitude and pattern of censoring jointly impact the information available for testing specific models or for identifying outliers. Yet such testing is important for many applications, e.g., estimating limits on contaminants in soil or groundwater. This paper presents a novel strategy for computing percentage points under nearly arbitrary left-censoring for two common tests: the probability plot correlation coefficient goodness-of-fit test and Rosner's block outlier screening test. The new strategy: (1) eliminates the need to 'fudge' percentage points computed from complete samples, (2) allows a unique set of percentage points to be computed for each dataset, depending on the magnitude and pattern of censoring, and (3) partially alleviates the complication of left-censoring in the 'chicken-and-egg' problem of needing a distributional model to identify outliers, but also needing to remove outliers prior to fitting a data model.

Key Words: goodness-of-fit testing, outlier testing, left-censored data, environmental

1. Introduction

As is well known, traditional diagnostic tests of normality and outlier identification were developed for complete, uncensored samples (e.g., Shapiro-Wilk, Anderson-Darling, Dixon, Rosner, etc.), including nearly all of the published algorithms and available percentage points. In practice, the same tests are applied to left-censored data (i.e., non-detects) — especially in environmental data analysis — by 'fudging': censored measurements are either ignored or perhaps imputed to some fraction of the detection/reporting limit. Then the sample is assumed (pretended) to be complete.

Does such 'fudging' make an important difference in goodness-of-fit and/or outlier testing? If so, is there a practical way to accommodate data censoring and still construct accurate diagnostic tests?

1.1 Non-detects (left-censored data) are a challenge with environmental data

Non-detects are one of the 'bug-a-boos' of environmental data analysis. Typically, the degree of bias in one-sample estimates, confidence intervals, prediction limits, etc. is either unknown or crudely bounded. Especially problematic are samples containing multiple levels of censoring (e.g., varying reporting limits) interleaved or interwoven in magnitude with quantified (uncensored) measurements. At best, partial rankings of such samples are possible. Many pairwise orderings are indeterminate (e.g., 4 ppb vs. < 10 ppb) and may have to be treated as ties in nonparametric procedures. Further, the possible levels of censoring and patterns of 'mixing' of groups of uncensored and censored measurements are nearly infinite. It would be impossible to enumerate all cases in published tables or algorithms.

1.2 Further challenge to parametric model fitting and outlier testing

Not only do percentage points for normality tests assume complete samples, but formal outlier testing is always parametric: one must assume a known or working distributional

*MacStat Consulting, Colorado Springs, CO

model. A 'Catch-22' occurs if *both* left-censored data and outlier(s) are present in the same sample. One needs to identify a model to classify possible outlier(s), but also needs to eliminate outlier(s) to properly fit the model in the first place. Again, since all the percentage points for outlier tests also assume complete samples (e.g., Barnett and Lewis, 1994), this not uncommon dilemma is solved in practice by assuming a complete sample even when left-censored data are present.

1.3 A practical solution

Since pre-published percentage points to account for possible censored data configurations are not feasible, a practical solution is to generate them 'on-the-fly.' That is, embed Monte Carlo algorithms into diagnostic tests to compute percentage points that account for sample-specific censoring configurations. We have constructed such algorithms for two common diagnostic/model fitting techniques:

- Filliben's normality test (Filliben, 1975), and
- Rosner's outlier test (Rosner, 1975)

Depending on the sample size and processor, 30-50,000 Monte Carlo replications of either algorithm can be run in typically 10-60 seconds per test on modern PCs. This is slower than commercial software implementations of Filliben's or Rosners tests, but still practical for regular data analysis and, we would argue, more accurate. Each algorithm is described in turn below, but they both depend on first constructing a partial ranking of the censored sample and then estimating (probability) plotting positions for the subsample of uncensored (detected) measurements. This basic step accounts for the observed level and pattern of censoring.

2. Extending Filliben's Normality and Rosner's Outlier Tests

2.1 Filliben's Test

In complete, uncensored samples, Filliben's test computes the correlation between the order statistics ($X_{(i)}$) and the approximate expected median normal quantiles of the sample ($Q_{(i)}$), as on a Q-Q plot. The higher the correlation, the closer the fit of the data to a normal model. To extend Filliben's procedure to left-censored samples, we first partially rank the sample (of size N) using either a modification of the well-known Kaplan-Meier algorithm (Kaplan and Meier, 1958; USEPA, 2009) or a similar construction popular in environmental circles, regression on order statistics (ROS) (Gilliom and Helsel, 1986). From the plotting positions (p_i) estimated on the ordered subsample U of uncensored values (size $N - m$, where $m = \#censored$), we then compute associated normal quantiles ($Q_{(i)} = \Phi^{-1}(p_i)$) and correlate those quantiles with the values in U. This gives the extended Filliben test statistic:

$$r_{cen} = \frac{\sum_{i \epsilon U} (X_{(i)} - \bar{X}_U)(Q_{(i)} - \bar{Q}_U)}{\sqrt{\sum_{i \epsilon U} (X_{(i)} - \bar{X}_U)^2 \sum_{i \epsilon U} (Q_{(i)} - \bar{Q}_U)^2}}$$

where U is the set of subscripts associated with the uncensored subsample, and $\bar{X}_U$ and $\bar{Q}_U$ are the subsample means of $X_{(i)}$ and $Q_{(i)}$ respectively.

To compute 'on-the-fly' percentage points for this test, we generate a large series (N_{sim}) of ordered, standard normal samples of size N. In each Monte Carlo sample, we

select the subset corresponding to the same estimated plotting positions as the original uncensored subsample ($i \epsilon U$), thus mimicking the observed level and pattern of left-censoring. By then correlating each subset with expected quantiles $Q_{(i),i \epsilon U}$, we generate an 'on-the-fly' null distribution for the extended Filliben statistic. This allows easy computation of percentage points and p-values under the hypothesis of normality.

2.2 Rosner's Test

For complete, uncensored samples, the most common form of Rosner's algorithm does the following:

1. Identify a block of $k \geq 2$ possible outliers in a sample of size N.
2. Iteratively compute studentized residuals of the outlier subsample, beginning with the most extreme possible outlier. That is, compute $R_1 = max_{j \epsilon 1,\ldots,N}((y_j - \bar{y})/s_y)$.
3. Exclude the most extreme remaining value and repeat steps (2-3) until k studentized residuals have been computed; these residuals form a vector of Rosner test statistics (R_i) of length k.
4. Test each block size ($k \geq 2$) in order from k down to 2 as a possible group of outliers; conduct each test by comparing the vector of Rosner statistics against a k-vector of Rosner critical points (v_i). If any $R_i > v_i$, declare the $j = max_i(R_i > v_i)$ most extreme values to be a block of j outliers.

To extend Rosner's test to left-censored samples, we modify the algorithm as follows:

1. Identify a block of $k \geq 1$ possible outliers.
2. Exclude the possible outliers, then partially rank the reduced sample using regression on order statistics (ROS).
3. Impute values for the subsample of censored measurements (C) using the estimated ROS plotting positions from step (2). This first entails fitting an acceptable distribution to the (left-censored) reduced sample.
4. Add back the excluded block of possible outliers to the imputed, reduced sample, then iteratively compute a vector of k Rosner statistics on the combined (size N) data.

To compute 'on-the-fly' percentage points for the extended test, we generate a large Monte Carlo series (N_{sim}) of ordered standard normal samples of size N, as with Filliben's extension. In each generated sample — using the same ROS-based plotting positions estimated from the original sample, respectively, for the censored (C) and uncensored (U) subsamples — exclude those values corresponding to the ranks of subsample C. Then use ROS to impute values for each excluded slot, treating this subsample *as if* it were left-censored, again to mimic the level and pattern of censoring in the original sample. Finally, compute a k-length vector of (censored) Rosner statistics (R_i^{cen}) as in the algorithm above.

The result of this algorithm is an $N_{sim} \times k$ array of simulated Rosner statistics. Rosner (1977, pp. 307-08) shows how to use the marginal distributions of $R_1^{cen}, \ldots, R_k^{cen}$ to compute joint percentage points under the null hypothesis of no outliers by finding β and $v_1, \ldots, v_k$ such that

$$Pr[R_i > v_i(\beta)] = \beta, \text{ for } i = 1, \ldots, k$$

Table 1: Percentage Points for Complete, Uncensored Samples ($N_{sim} = 50,000$)

	Filliben (1975)			On-the-Fly		
N	$r_{0.01}$	$r_{0.05}$	$r_{0.1}$	$r_{0.01}$	$r_{0.05}$	$r_{0.1}$
4	0.822	0.864	0.898	0.820	0.865	0.894
8	0.859	0.905	0.924	0.856	0.902	0.922
20	0.925	0.950	0.960	0.923	0.947	0.958
40	0.958	0.972	0.977	0.956	0.970	0.975
100	0.981	0.987	0.989	0.980	0.986	0.989

and

$$Pr\left\{\bigcup_{i=1}^{k}[R_i > v_i(\beta)]\right\} = \alpha$$

where α is the significance level. The k-vector $v_i, i = 1, \ldots, k$ is the desired set of percentage points.

3. Performance and Benefits

'On-the-fly' percentage points for Filliben's and Rosner's diagnostic tests offer several benefits and few drawbacks. Probably the biggest disadvantage is computational speed. Neither test is instantaneous, and the computation time will vary by speed of Monte Carlo generation, the number of default replications, sample size, and degree of censoring. In tests on a 2012 iMac (3.4 GHz Intel Core i7), the modified Filliben's test averages 10-30s per data set, while the modified Rosner's test typically runs 20-60s. The default code runs 50,000 replications in order to limit the Monte Carlo error associated with each test.

The advantages include:

- **Reproducibility**. Each modified algorithm matches published percentage points for complete, uncensored samples within Monte Carlo error ($s.e. \sim 0.0022$ for 50,000 replications).

Table 1 compares examples of varying sample sizes for Filliben's test. The number of replications can be adjusted as desired to balance the performance/accuracy tradeoff.

- **Accuracy**. Depending on the degree and pattern of left-censoring, percentage points and/or p-values for these tests can differ substantially. Therefore, relying on published tables designed for complete samples can give inaccurate results.

Table 2 gives comparative examples for Filliben's test. Samples of metals concentrations in groundwater are compared using each of three methods: (1) the proposed Monte Carlo algorithm, (2) ignoring non-detects completely and treating the uncensored measurements as a complete sample, and (3) imputing half the reporting limit for each non-detect and then treating the imputed sample as complete (and uncensored). The probability plot correlation coefficient test statistic is given in column Q-Q Corr, and the generated or published percentage points for common significance levels are given in the three right-hand columns.

For lithium with zero censoring, the sample fails the normality test, but does so identically for all three approaches. For radium with nearly 24% censoring, not only do the

Table 2: Example Comparisons for Filliben's Test

Metal	Method	N/ND%	Q-Q Corr	$r_{0.01}$	$r_{0.05}$	$r_{0.1}$
Radium	on-the-fly	38/23.7	0.9830	0.939	0.961	0.970
	ignore	29	0.9618	0.945	0.962	0.969
	impute	38	0.9271	0.956	0.970	0.975
Molybdenum	on-the-fly	46/69.6	0.9921	0.881	0.922	0.938
	ignore	14	0.9661	0.901	0.934	0.947
	impute	46	0.7480	0.962	0.974	0.979
Lithium	on-the-fly	48/0	0.9272	0.962	0.974	0.979
	ignore	48	0.9272	0.963	0.975	0.980
	impute	48	0.9272	0.963	0.975	0.980

percentage points differ by method, but also the test statistic and, importantly, the *outcome* of the test. The 'on-the-fly' method computes the probability plot correlation of the uncensored values after first adjusting the plotting positions for the level and pattern of censoring. No such adjustment is made if the non-detects are either ignored or treated with simple imputation. In fact, in this example, the radium sample passes the normality test using the 'on-the-fly' algorithm, but fails if the censored values are imputed and fails at the 0.05 level if the non-detects are ignored.

Molybdenum, with almost 70% censoring, also exhibits instructive differences. Simple imputation fails badly, while ignoring the non-detects altogether passes the test, largely due to the much smaller percentage points associated with disregarding most of the data. The highest correlation stems from the 'on-the-fly' approach, and in turn, the high level of censoring leads to the lowest set of percentage points. The key point in these comparisons is that naive application of these diagnostic tests may lead to incorrect decisions, as well as inaccurate percentage points.

- **Flexibility**. The modified tests can be used not only with left-censored data, but also with a wide variety of (unpublished) censoring configurations and/or sample sizes. The algorithms are coded in R to enable widespread accessibility and possible extensions.

Many such extensions are feasible. Rosner's test, for example, was originally developed for testing blocks of two (2) or more outliers, with published percentage points to match. The modified test is coded to also test single outliers. Both Rosner and Filliben assume normality under H_0; this has been extended by considering a range of Box-Cox type transformations with left-censoring. A prototype of the R code for the modified procedures also tests the Weibull and gamma distributions. Indeed, the same logic could be used to modify other diagnostic tests to handle left- and/or right-censoring.

Figures 1 through 3 illustrate upgradient background data for barium, chromium, and fluoride in groundwater, respectively. These examples are distinguished by different numbers of possible outliers and levels of left-censoring. The results from applying the 'on-the-fly' Rosner's algorithm to these cases is summarized in Table 3.

In all three examples, the seemingly obvious outliers are confirmed no matter how the censored measurements are treated. However, the 0.01 percentage points differ by method; the 'on-the-fly' critical points generally fall between ignoring the non-detects altogether and imputing each non-detect to half its reporting limit. For Rosner's procedure, the lower

Table 3: Example Comparisons for Rosner's Test

Metal	Method	N/ND%	Block Size	R_i	$v_{0.01}$
Barium	on-the-fly	61/0	1	7.18	3.56
	ignore	61		7.18	3.56
	impute	61		7.18	3.56
Chromium	on-the-fly	182/51.6	2	(12.6, 12.5)	(3.95, 3.27)
	ignore	88		(8.75, 8.99)	(3.84, 3.25)
	impute	182		(12.6, 12.5)	(4.08, 3.45)
Fluoride	on-the-fly	66/72.7	1	6.23	3.26
	ignore	18		3.92	2.94
	impute	66		7.54	3.59

the critical point, the more sensitive the test, so accounting for left-censoring in the proposed manner appears to offer more powerful outlier tests than simply imputing the sample and treating it as complete. On the other hand, eliminating the non-detects altogether ignores substantial information in the censored samples and also sharply lowers the comparative set of Rosner statistics. Again, this would appear to lower the power of the test relative to the 'on-the-fly' approach.

4. Conclusion

Environmental data analysis still makes frequent use of diagnostic testing in order to fit parametric data models or identify outliers. If the samples contain non-detect (i.e., left-censored) measurements, the diagnostics are typically run by 'fudging' or perhaps ignoring the censoring content, in part because published guidance is lacking on proper adjustment for left-censored data.

Yet, naive use of goodness-of-fit and/or outlier testing with left-censored data may easily lead to biased or inaccurate diagnostic tests. Our method for computing 'on-the-fly' Monte Carlo percentage points is now practical for at least two common diagnostic tests: Filliben's probability plot correlation coefficient test of normality and Rosner's outlier test. While the power of the modified tests must still be formally investigated, informal application of each test on hundreds of real data sets indicates very good agreement between the test result and what you might expect from visual examination of the data.

To encourage greater use of these techniques, R routines for both tests are available by request from the author (kcmacstat@gmail.com).

REFERENCES

Filliben, J. J. (1975), "The Probability Plot Correlation Coefficient Test for Normality," *Technometrics*, 17, 111–117.

Gilliom, R. J., and Helsel, D. R. (1986), "Estimation of Distributional Paramters For Censored Trace Level Water Quality Data: I. Estimation Techniques," *Water Resources Research*, 22, 135–146.

Kaplan, E. L., and Meier, P. (1958), "Non-parametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Rosner, B. (1975), "On the Detection of Many Outliers," *Technometrics*, 17, 221–227.

Rosner, B. (1977), "Percentage Points for the RST Many Outlier Procedure," *Technometrics*, 19, 307–312.

USEPA (2009), *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Unified Guidance*, Office of Solid Waste, Washington DC.

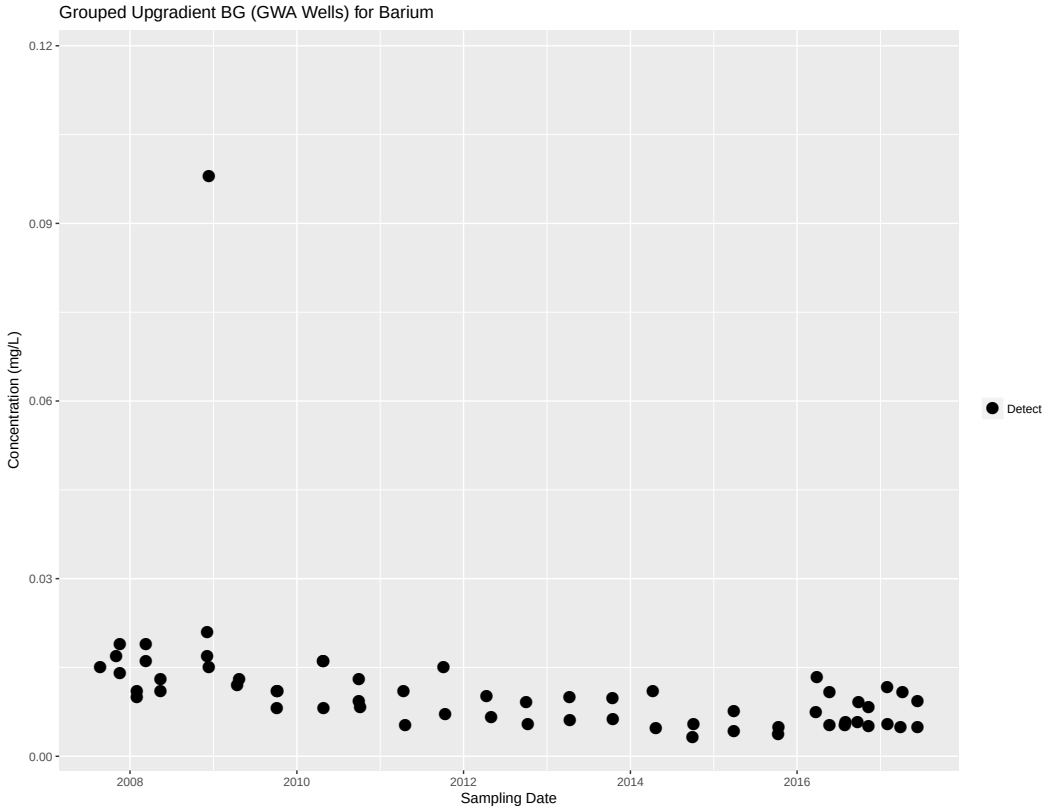

Figure 1: Grouped Upgradient Barium in Groundwater

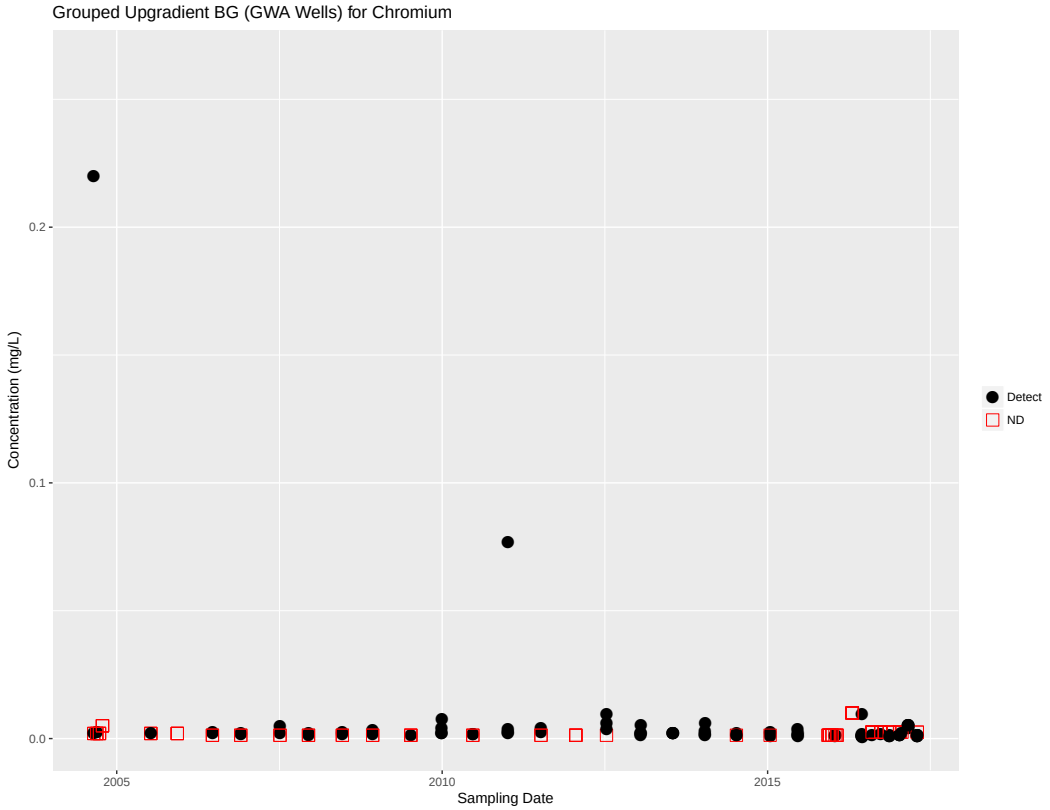

Figure 2: Grouped Upgradient Chromium in Groundwater

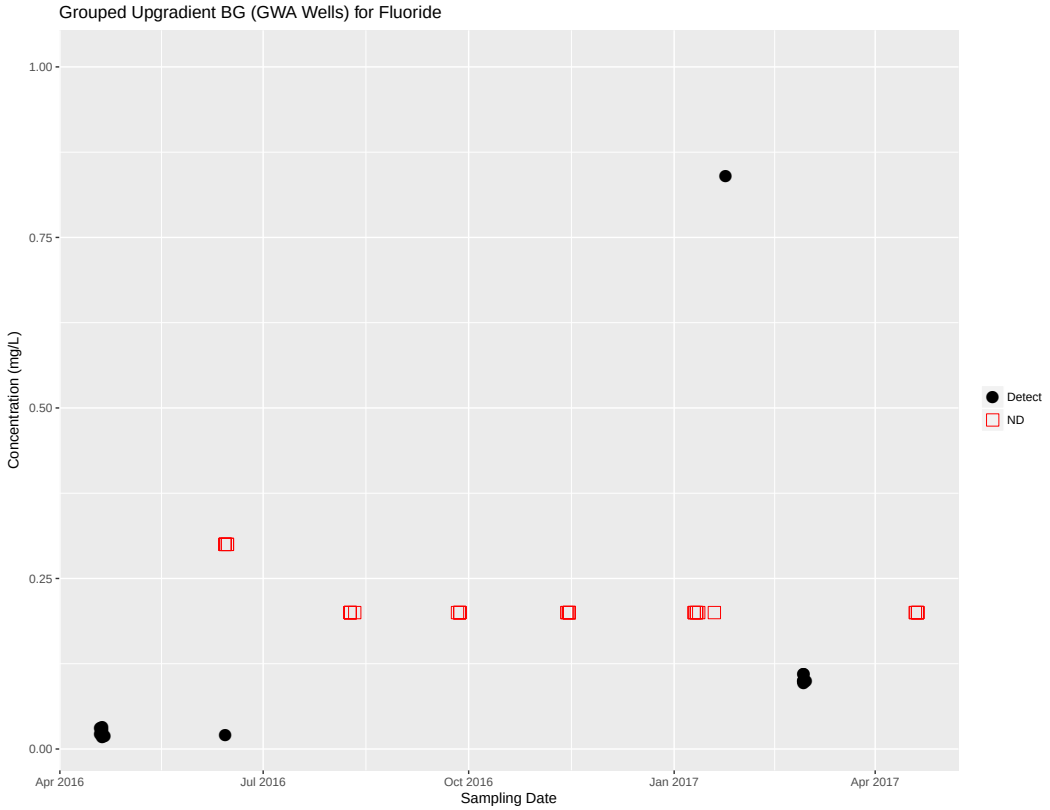

Figure 3: Grouped Upgradient Fluoride in Groundwater