Method for absolute status based on cdf scoring functions
Elgin Perry
4/15/2010

The absolute status algorithm proposed here is a tool for assessing the status of the distribution of a water quality parameter in an assessment region relative to a limiting acceptable distribution. For cases where high values of a parameter are unacceptable, the limiting acceptable distribution defines an upper threshold distribution of acceptable values. Note that there is no hardline number for whether a single observation is acceptable and thus no strict sample by sample pass-fail criterion . The idea is that large numbers are acceptable if they occur in low frequencies. Large numbers in high frequencies are not acceptable. Based on this logic, there may be considerable overlap in the range of an acceptable distribution and an impaired distribution, but the two will differ in the frequency and magnitude of undesirable observations.

This absolute status algorithm is based on the relative status algorithm that was used by CBP 1997-200?. It is similar to the relative status algorithm in that it employs a similar method of scoring individual observations against a reference condition. The scoring function is the cumulative distribution function (CDF) of the reference condition. It differs from the relative status algorithm in that it uses a different definition of the reference condition. The relative status method used the CDF of the first 6 years of monitoring data as the reference condition. The absolute status method uses a CDF describing a threshold of acceptable conditions for each parameter as the reference condition.

In what follows, we present the logic of the method and details for implementation. This is followed by an example for chlorophyll. Following the example is a is a discussion of the advantages and flexibility of this approach. The discussion is concluded with a review of outstanding issues that might lead to improvement in the method.

The scoring algorithm:

The method can be summarized as a sequence of steps, but as will be developed in discussion, there is much room for flexibility.

1. For each parameter to be assessed, identify a statistical distribution that is a reasonable model for the distribution of that parameter when it is observed at the extreme of it's acceptable conditions. In the chlorophyll example, the log-normal distribution is used.

2. Estimate the parameters of the threshold distribution based on some standard that defines the limits of what is acceptable. This standard might be one of the following: reference data, the water quality model, values taken from literature review, best professional judgement, or any source is available to help define the boundary between acceptable and unacceptable.

It is desirable to make these estimates spatially and seasonally explicit. In the chlorophyll example, reference data are used.

3. Using the results of 1. and 2. construct a cumulative distribution function(s) (CDF(s)) for the threshold distribution(s). The log-normal distribution for chlorophyll is estimated by finding the mean and variance of log-transformed data.

4. Use the CDF(s) defined in 3. to score observations in the assessment data into the interval (0,1).   This scoring uses what is known as the probability integral transform.

5. Compute a summary statistic for the (0,1) scores of the assessment data.

6. Assess the data summary statistic against a predetermined set of thresholds for compliance vs. noncompliance.

7. It may be desirable to statistically test the null hypotheses that the distribution for the assessed data exceeds the threshold distribution.

Chlorophyll Example

Here we illustrate this absolute status method using Chesapeake Bay chlorophyll. This chlorophyll example uses prior work (Williams et al., 200?, Buchanan et al. 2005) to estimate chlorophyll threshold distributions. Williams identified thresholds for reference conditions for chlorophyll by season and salinity zone (Table 1.0) based on Buchanan's work.   More useful to this exercise, Buchanan partitioned chlorophyll observations into categories based on habitat (Table 2.0) that were subsequently grouped as Good, Bad, and Mixed.   We use Buchanan's Better, Best category to defined the threshold distribution (Figure 1.0).   To illustrate assessment, we compute scores for and summarize the worst chlorophyll ('Bad ') (Figure 2, blue step function). For simplicity, we consider assessent only for tidal fresh-summer.

Step 1. The log-normal distribution is frequently used to model the distribution of Chlorophyll and that is the distribution that will be used for this example. In discussion other possibilities are set forth.

Step 2. To estimate the parameters of the log-normal distribution,   the "Better-Best" tidal fresh - summer data (Buchanan et al. 2005) are log-transformed and the mean and variance are computed in the logarithm metric.

Step 3. To obtain a scoring function for chlorophyll, we estimate the CDF of the distribution as a smooth log-normal curve (smooth black function, Figure 1) using the log-mean and log-variance from the reference data. For reference, we also show Empirical Cumulative

Distribution Function (ECDF) (step function, Figure 1.0) for this sample. Note that the 70th percentile which was used by Williams as a threshold occurs at a point where the ECDF deviates from the estimated CDF.

Step 4. To illustrate chlorophyll scoring, we use 9 observations drawn at rougly equally spaced ranks from the 'bad' population (Figure 3., Table 3.) These observations are represented on the chlorophyll scale by blue circles on the x-axis. These observations are projected first vertically to the reference CDF function and the horizontally to the (0,1) score scale on the y-axis. For this example where the log-normal is used to model the threshold population, there is no closed form mathematical function that will score the data. Thus we must rely on probability integral transform functions that are available in high level statistics languages such as SAS or R. However, it is proposed below that nearly identical results can be obtained using the log-logistic CDF. To score data using the log-logistic, one needs only a common exponential function.

Step 5. To illustrate summary statistics, the mean and median of the scores for the impaired population are computed as 0.8336 and 0.9879 . Which statistic is optimal for assessment is an outstanding question, but either of these will suffice for illustration.

Step 6. The cutpoints for pass/fail should have a scientific basis. For this example we use a cutpoint of 0.7 as was proposed by Williams (Table 1.0). Based on this cutpoint, the 'Bad' data are judged to be not meeting the designated use based on either the mean or the median.

Step 7. There are numerous approaches available for statistical inference in this method. Some of these are discussed below. To illustrate statistical inference, we use the fact that under null hypothesis that the assessed data follow the assumed threshold distribution, the probability integral transformed data (i.e. the (0,1) scores) follow a uniform distribution. A median of n observations from a uniform distribution follows a Beta(n/2, n/2) distribution. If the median of the scores from the assess population exceed the 95% percentile of this Beta distribution, then the exceedance is statistically significant. The sample size of bad data is 187 and the appropriate distribution is Beta (93.5,93.5). A 90% confidence interval for 0.5 based on this Beta distribution is (0.44, 0.56). The median of the Bad data of 0.9879 exceeds the 95% upper bound of 0.56 by quite a lot. Thus the exceedance is statistically significant.

Discussion of the Chlorophyll Example

There are many options to consider when implementing this absolute status algorithm. Those chosen here for Chlorophyll were chosen for expediency and not because they are thought to be the best and most defensible choices. None-the-less it remains interesting to discuss these choices relative to others to illustrate the flexibility of this method.

Choice of Distribution.

The log-normal distribution was chosen for chlorophyll because it is often found to be a good model for chlorophyll and many other water quality parameters. As noted above, the log-normal and normal have the disadvantage that the CDF for these distribution requires evaluation of the integral

$$F(x;\mu,\sigma) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

which does not have a closed form solution. Most high level languages use an interpolating polynomial to evaluate this function, but spread sheet users may not have access to this. An alternative is to us the log-logistic for which the CDF is

$$F(x;\alpha,\beta) = \frac{1}{(1+e^{-(x-\alpha)/\beta})^2}$$

where the mean = α and the variance is $\beta\pi^2/3$. This requires only the more common exponential function.

Choice of definition for reference

estimation methods will depend on optimal estimators for the distrubution chose in step 1.

Once steps 1-3 are complete, step 4 is a given.

Step 5 - as noted above the choice of summary statistic is open for research. Median has advantage of 1-to-1 relationship between scores and raw data.

Step 6 -

Under the null hypothesis that the chlorophyll distribution being assessed is following the threshold distribution, the mean and median score for the assessed population has expected value 0.5. For chlorophyll, an impaired population will be shifted to the right (blue curve, Figure 2.).

If the chl distribution is just at the threshold distribution, the both the median and the mean of these CDF scores will be 0.5. A shift in central tendency to the high side of 0.5 indicates moving toward unacceptable chlorophyll.

Science based definition of meaningful departure from reference

0.7 for this case is a shift in median chlorophyll from 8.4 to 10.9.
Assuming that log variance doesn't change, it corresponds to a shift in distribution shown in Figure 4.0. Under the reference distribution chlorophyll exceeds 19 about 5% of the time under the impaired distribution chlorophyll exceeds 19 about 13% of the time.

caution against using 'statistical criterion'

step 7 - Mean v Median v Midrange.
95% upper bound of 0.7 corresponds to a sample size of 16.

distribution shifted to the right (blue line) is exceeding acceptable conditions.

The threshold distribution may vary seasonally and across subregions of a reporting region (Table 1.0). The CFD scoring function puts all observations on a standard (0,1) scale relative to the appropriate threshold distribution which will facilitate combining scores over seasons or subregions.

Discussion

Note that scoring observations into the 0-1 scores faciliates aggregation of summary statistics over space and time when the threshold distribution may vary over space and time.

For many water quality parameters, the log-normal is a reasonable choice. The distribution at the extreme of acceptable conditions will be called the 'threshold' distribution.

This estimation at a minimum should include some information on the central tendency and acceptable spread of the threshold distribution. For example the threshold distribution for chlorophyll might center on 10 mg/l and exceed 18 only ten percent of the time.

The probability integral transform results in data in the interval (0,1) which follows a uniform distribution under the null hypothesis that the data are from the threshold distribution (Roussas, 1973).

Under the null hypothesis that the assessment data follow the threshold distribution, the median and the mean of the (0,1) scores have the value 0.5. Thus if the median, mean, or mid-range of the assessment data exceed 0.5 for cases where large values are bad, it would indicate unacceptable conditions.

We feel this approach will offer two advantages over the current status approach based on the percent of violations metric that is implemented for the report card. One advantage is that this method will retain information about the degree of violation (i.e. distance above the threshold). Another advantage is that this approach will enable statistical inference so that 33% based on 1 sample in 3 will be interpreted differently from 33% based on 10 samples in 30.

Roussas, George G. (1973)
A First Course in Mathematical Statistics. Addison-Wesley, Reading, Mass.

Outstanding issues:

Nonparametric density estimation

What happens when distribution is mis-specified.

Statistical summary, median vs mean vs mid-range (max+min)/2
The 3 year median of this 0-1 data is computed as and indicator of status in the current three year period. The median of n observations taken from a uniform distribution follows a Beta distribution with parameters (m,m) where m = (n+1)/2 and n is the number of observations(Roussas, 1973)

Space - Time issues

Potential for 'Context Sensitive Criteria'.

Table 1.0 Chlorophyll thresholds (Williams,) by season and salinity zone as absolute chlorophyll (mg/l) and distribution percentiles.

		salinity zone			
season	measure	TF sal<0.5	OH 0.5 <sal< 5.0	MH 5.0<sal<18	PH 18<sal
Mar-Jun15	chl	13.98	20.93	6.17	2.80
	%ile of Ref	95	90	65	50
Jun15-Sep	chl	12.00	9.47	7.7	4.52
	%ile of Ref	70	70	55	50

Table 2.0 Habitat Categories as defined by Buchanan and used by Williams.

description	number	light	DIN	PO_4	group
worst	1	Low	excess	excess	BAD
poor	2	Low	excess	excess	
mixed poor light	3	Low	limiting	excess	MIXED
	4	Low	excess	limiting	
	5	Low	limiting	limiting	
mixed better light	6	High	excess	excess	
	7	High	excess	limiting	
	8	High	limiting	excess	
better	9	High	limiting	limiting	GOOD
best	10	High	limiting	limiting	

Table 3.0 Chlorophyll observations taken from the "bad" chlorophyll group and scored by the CDF estimated from the "good" chlorophyll group.

observation rank	observed chlorophyll	CDF score
11	3.440	0.0369
31	8.720	0.5314
51	16.450	0.9118
71	21.680	0.9717
91	25.420	0.9870
111	31.585	0.9961
131	37.380	0.9986
151	47.100	0.9997
171	61.300	0.9999
mean	28.12	0.8259
median	25.42	0.9869

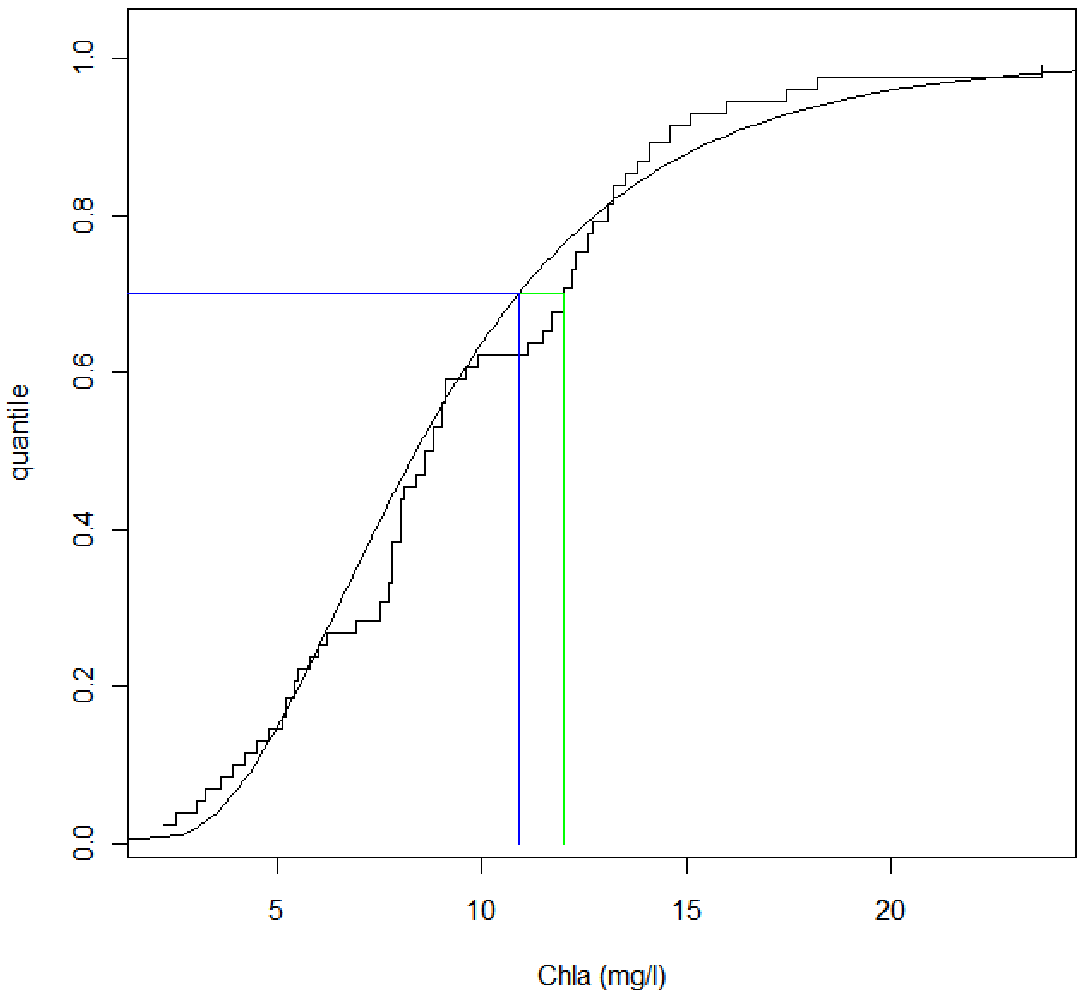

Figure 1.0 The empical distribution function (step function) of chlorophyll a from 'good' regions with an overlay of the log-normal cumulative distribution function (smooth curve). Reference lines show the 70th percentile estimated by the EDF (green) and the CDF (blue).

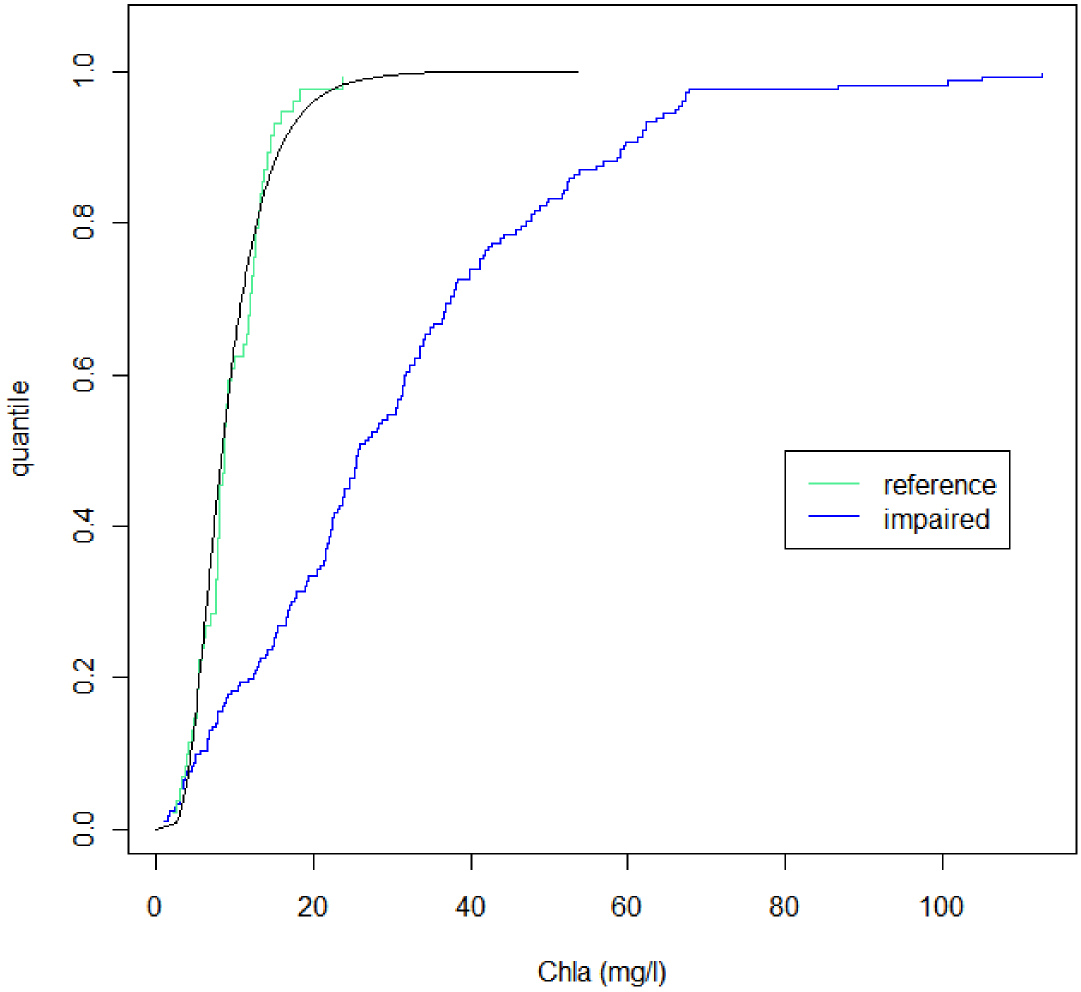

Figure 2.0 The empical distribution function of chlorophyll a from 'good' regions (green step function) with an overlay of the log-normal scoring function (smooth black curve) and the edf of chlorophyll a from 'bad' regions (blue step function).

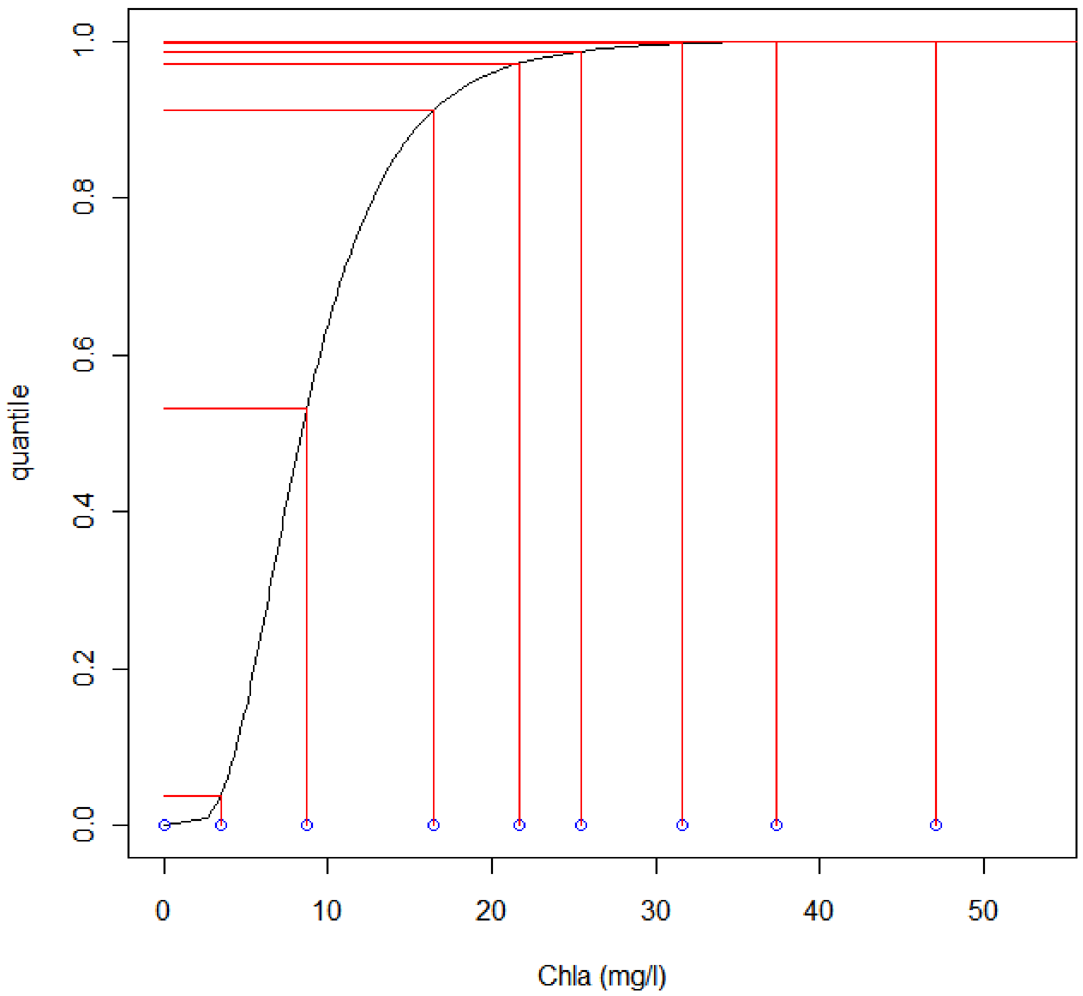

Figure 3.0 The estimated CDF or the log-normal scoring function (smooth black curve) based on 'good' regions with a subset of chlorophyll observations from 'bad' conditions (blue circles). The red lines show scoring of 9 observations from the 'bad' distribution as they map onto the (0.0,1.0) ordinate using the black scoring function.

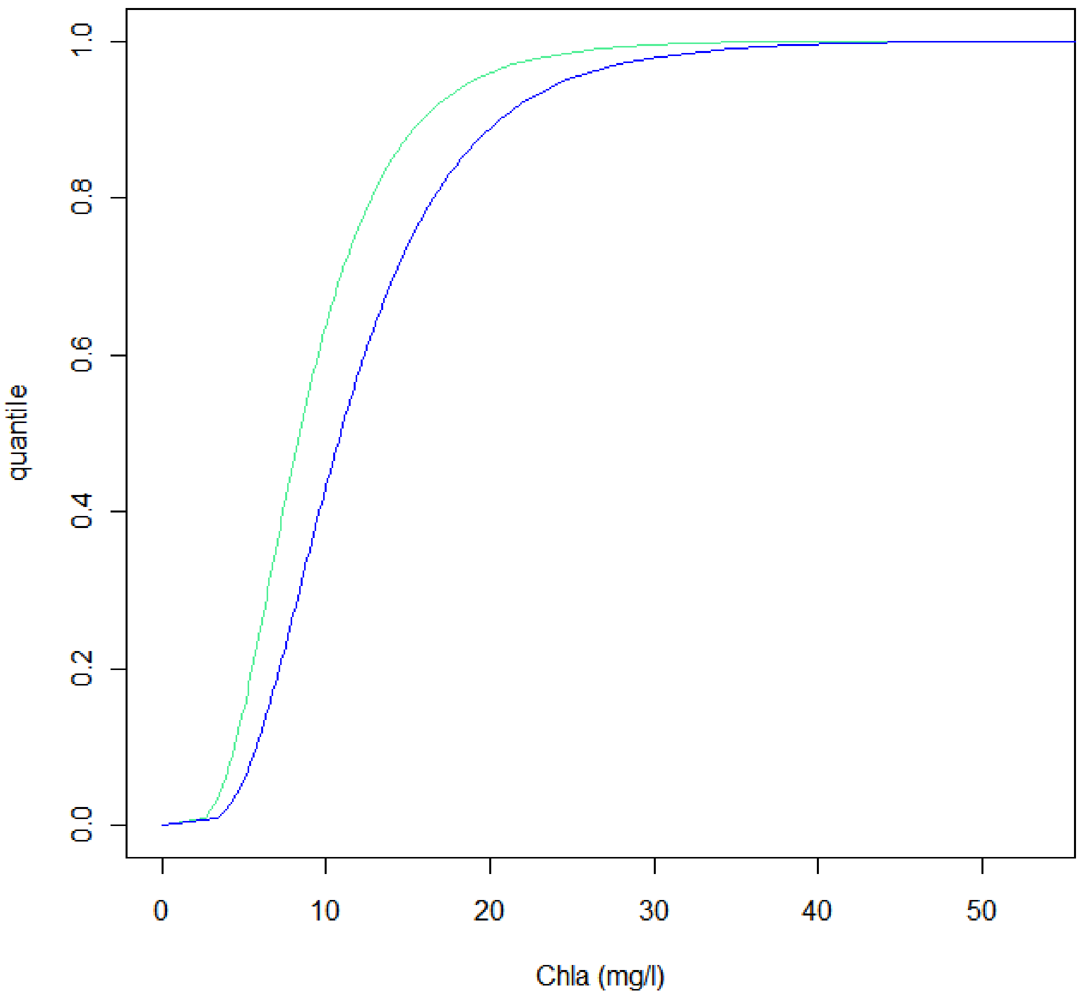

Figure 4.0 Shift in distribution is criterion threshold for scores is 0.7.