PPC 9481.1991(01)

ANALYSIS OF RETESTING PROCEDURES PAPER

OFFICE OF SOLID WASTE AND EMERGENCY RESPONSE

October 16, 1991

Glen D. Johnson
Commonwealth of Pennsylvania
Dept. of Environmental Resources
P.O. Box 2063
Harrisburg, Pennsylvania 17105-2063

Dear Mr. Johnson,

Four questions were raised in your letter of August 21. Hopefully our responses will be satisfactory. If you still have any concerns please feel free to contact us again.

First, concerning the "Analysis of Retesting Procedures" paper, we have re-examined our simulation code used to generate the power results given in that paper and have examined the code you provided as well. It seems that the discrepancy between our results can be traced to the method by which you generate future values for comparison to the simulated prediction limits.

The theory behind prediction intervals assumes that not only are the background measurements drawn at random from a (Normal) distribution, but that the future values to be compared against the prediction limits are also randomly drawn from the same population. That is, two sources of variability are built into the equation used to calibrate the width of a prediction interval: variability in the background measurements (accounted for by the quantity t.99,n-1 and by 1/n under the root symbol) and variability in the future values (accounted for by adding I to 1/n under the root symbol).

In your SAS code, the future values for a given effect size are always fixed at the expected mean level of the downgradient well. No variation is built into these numbers; consequently our power results differ. We have made an additional run of our simulations (based again on 10,000 iterations) to give you approximate power levels in the table below for each of the effect sizes you used. These results make sense from the standpoint that if the alternative mean is close to the background mean, adding variation to the future values should increase how often these numbers fall above the prediction limit and hence increase the power over what you derived. The

reverse situation should be true when the background and alternative means are far apart, as seen in the table, for then the alternative mean will generally be above the prediction limit, and variability in the future values will tend to lower the power somewhat.

EFFECT	YOUR POWER	OUR POWER
0.0	0.0000	0.0100
0.5	0.0008	0.0257
1.0	0.0051	0.0613
1.5	0.0301	0.1189
2.0	0.1116	0.2073
2.5	0.2698	0.3319
3.0	0.4756	0.4802
3.5	0.6946	0.6256
4.0	0.8473	0.7570
4.5	0.9386	0.8568
5.0	0.9801	0.9223

With regard to the article by Robert Gibbons in Ground Water (Vol. 29, No. 4, 1991), our basic reaction to the method Gibbons proposes for monitoring large networks of downgradient wells is not wholly unfavorable. There do seem to be advantages to employing some type of retesting strategy in order to verify results from wells that show possible contamination. EPA has in fact already approved a proposal from the State of California that adopts a retesting strategy. However, the specific strategy put forward by Gibbons, that is, an initial tolerance interval followed by a prediction interval on any resampled wells, may or may not be the best retesting strategy. The approved California proposal, for instance, consists only of prediction intervals with double resampling of downgradient wells that initially exceed the prediction limit. Both resamples must be immediately collected from the monitoring well and both must pass the retest to verify that the initial failure was a false positive.

In addition, Gibbons' article raises a few questions concerning the statistical logic used. First, Gibbons' basic example hypothesizes a network of 20 downgradient wells, each being monitored quarterly for five constituents. As Gibbons notes, this results in 100 sample measurements per quarter that must be tested statistically. It is not true, however, that all 100 values would be compared to single tolerance (or prediction) limit calculated from the background data. Rather, each constituent would have to be tested separately, leading to five separate tolerance (or prediction) limits each used to test 20 measurements. Clearly, it would not be the case that all five constituents would have similar distributions of concentration

values. This point is not crucial to Gibbons case, but it is worth emphasizing that his hypothetical framework should actually lead to more conservative prediction limits than he describes.

Another point is that Gibbons' discussion of Type I error rates when comparing the performance of tolerance and prediction limits without resampling, particularly in regard to Figure 1 on p. 568, is somewhat misleading. Gibbons argues for instance that "the 95% prediction limit for the next 1000 measurements achieves its nominal error rate of 5%. However, the false positive rate for the 95% confidence 95% coverage tolerance limit is over 70%. Even 99% coverage produces a false positive rate of approximately 17%." What Gibbons means by false positive rate here is not the traditional Type I error probability, rather, it represents the frequency with which any of the 100 sample measurements falls above the prediction or tolerance limit.

For prediction limits these two definitions of error are the same, since the prediction limit is designed to contain all of the 100 samples. For tolerance limits, however, Gibbons' definition of the false positive rate is not correct, because a tolerance limit is designed to "miss" a certain fraction of the tested measurements. Under 95% coverage, a tolerance limit is expected to miss approximately 5 out of every 100 new samples. The tolerance limit only falls in the Type I error sense, if the actual coverage of the limit is less than expected amount (e.g, 10 out of 100 samples fall above the limit instead of the expected 5 or less). It is true enough that some measurements in a large enough sample will fall above the tolerance limit; however, this does not indicate a failure of the tolerance limit to do its job. If 100 new measurements were collected from a single downgradient well, and at least 95 of those values fell below the tolerance limit, there would be no need to designate the well as possibly contaminated.

Granting the above comments, Gibbons does recognize a basic problem in applying a tolerance limit approach to a set of measurements taken one per downgradient well. That is, measurements which fall above the tolerance limit may indeed indicate contamination at particular wells, because distinct wells may have different distributions of the constituent being tested. If data from many downgradient wells are pooled together, the tolerance limit approach assumes that each well has the same distribution of sample values and that values fall above the tolerance limit only because a large enough sample from any (normal) distribution will have a few extreme measurements. This assumption may not be true if just one or two downgradient wells have been contaminated, so that some of the extreme values are the result of contamination rather than just random variation in a large set of measurements. In other words, by allowing a certain fraction of the values to be above the tolerance limit (typically 1% or 5% of the concentrations),

actual contamination at a very few wells could be missed.

One solution to this problem is as Gibbons suggests to retest each well for which the sample measurement falls above the tolerance limit. A more practical alternative not discussed in the article relates to the likely nature of contaminated wells for many constituents. Experience with monitoring data suggests that an actual spill or leak from a monitored facility results in concentration levels elevated typically by one or more factors of magnitude above background levels. Samples from wells contaminated in this way should be much greater in concentration than even extreme values from uncontaminated wells. Consequently, it may be easy to identify contaminated wells by comparing the relative magnitudes of those samples which fall above the tolerance limit, even in the absence of any retesting strategy.

Your inclination concerning ground-water sample independence with respect to quarterly measurements is consistent with our experience in evaluating ground-water monitoring data. Keep in mind, however, that the 40 CFR Part 264, Subpart F regulations require at least semiannual sampling, which may improve the likelihood of sample independence in slow moving ground water. Further, well purging procedures that are implemented prior to sample collection also improve sample independence.

EPA is in the process of developing software for assisting Regional and State personnel in evaluating ground-water monitoring data. The system (GRITS-ground-water research information system) is an enhancement to an EPA Region VII data base that uses Lotus files for data input and will perform all of the Subpart F statistical procedures. We plan to provide training on the system and the included statistical procedures throughout the late summer and fall of 1992 (Philadelphia or a nearby metropolitan area will be a host training site).

I hope that these comments have been useful. Please contact me at (202) 260-3240 if I can be of further assistance.

Sincerely,

Original Document signed

James R. Brown

CC: Denise Keehner
Vernon Myers