Appendix

B

Some Methods To Be Avoided

In the following sections some statistical methods that should be avoided are described. These methods are not available in DUMPStat.

Analysis of Variance - ANOVA

Application of ANOVA procedures to ground-water detection monitoring programs, both parametric and nonparametric is inadvisable for the following reasons.

1. Univariate ANOVA procedures do not adjust for multiple comparisons due to multiple constituents which can be devastating to the site-wide false positive rate. As such, a site with 10 indicator constituents will have a 40% chance of failing at least one on every monitoring event (USEPA 1992 section 5.2.1).

2. ANOVA is more sensitive to spatial variability than contamination. Spatial variability affects mean concentrations but typically not the variance, hence small yet consistent differences will achieve statistical significance. In contrast, contamination affects both variability and mean concentration, therefore a much larger effect is required to achieve statistical significance. In fact, application of ANOVA methods to pre-disposal ground-water monitoring data can result in statistically significant differences between upgradient and downgradient wells, despite the fact that there is no waste in between. The reasons for this are:

    a) The overall F-statistic tests the null hypothesis of no differences among any of the wells regardless of gradient (*i.e.*, it will be significant if two downgradient wells are different), and

    b) The distribution of the mean of 4 measurements (*i.e.*, four measurements collected from the same well within a six month period) is normal with mean μ and variance $\sigma^2/4$ whereas the distribution of each of the individual measurements is normal with mean μ and variance σ^2. This means that the standard deviation of the mean of four measurements is one-half the size of the standard deviation of the individual measurements themselves. As a result, small but consistent geochemical differences that are invariably observed naturally across a waste disposal facility will be attributed to contamination. To make matters worse, since there are far more downgradient than upgradient wells at these facilities, spatial variation has a far greater chance of occurrence downgradient than upgradient further increasing the likelihood of falsely concluding that contamination is present. While spatial variation is also a problem for prediction limits and tolerance limits for single future measurements, it is not nearly as severe a problem as for ANOVA since the distribution of the individual measurement is considered and not the more restrictive distribution of the sample mean.

3. Nonparametric ANOVA is often presented as if it protects the user from all of the weaknesses of its parametric counterpart. This is *not* the case. Both methods assume identical distributions for the analyte in *all* monitoring wells. The only difference is that the parametric ANOVA assumes that the distribution is normal and the nonparametric ANOVA is indifferent to what the distribution is. Both parametric and nonparametric ANOVA assume homogeneity of variance, a condition that almost never occurs in practice. This is not a weakness of methods for single future samples (*i.e.*, prediction and tolerance limits) since the variance estimates rely solely on the background data. Why would anyone want to use downgradient data from an existing site (which could be affected by the site) to characterize natural variability? Yet this is exactly what the ANOVA does. Furthermore, ANOVA is not a good statistical technique for detecting a narrow plume that might affect only one of 10 or 20 monitoring wells (USEPA 1992 section 5.2.1).

4. ANOVA requires the pooling of downgradient data. Specifically, USEPA has suggested that four samples per semi-annual monitoring event be collected (*i.e.*, eight samples per year). As such, on average, it will never most rapidly detect a release, since only a subset of the required four semi-annual samples will be affected by a site impact. This heterogeneity will decrease the mean concentration and dramatically increase the variance for the affected well thereby limiting the ability of the statistical test to detect contamination when it occurs. This is not true for tolerance limits, prediction limits and control charts, which can and *should* be applied to individual measurements. For these reasons, when applied to ground-water detection monitoring, ANOVA will maximize both false positive and false negative rates, and double the cost of monitoring (*i.e.*, ANOVA requires four samples per semi-annual event or eight per year versus a maximum of four quarterly samples per year for prediction or tolerance limits that test each new individual measurement and more typically only two samples per year).

To illustrate, consider the data in Table 2 which were obtained from a facility in which no disposal of waste has yet occurred (see Gibbons, 1994 *NSWMA Waste Tech Conference Proceedings*, Charleston SC, 1/14/94).

Table 2
Raw data for all detection monitoring wells and constituents (mg/l)
This facility has no garbage in it.

Well	Event	TOC	TKN	COD	ALK
MW01	1	5.2000	.8000	44.0000	58.0000
MW01	2	6.8500	.9000	13.0000	49.0000
MW01	3	4.1500	.5000	13.0000	40.0000
MW01	4	15.1500	.5000	40.0000	42.0000
MW02	1	1.6000	1.6000	11.0000	59.0000
MW02	2	6.2500	.3000	10.0000	82.0000
MW02	3	1.4500	.7000	10.0000	54.0000
MW02	4	1.0000	.2000	13.0000	51.0000
MW03	1	1.0000	1.8000	28.0000	39.0000
MW03	2	1.9500	.4000	10.0000	70.0000
MW03	3	1.5000	.3000	11.0000	42.0000
MW03	4	4.8000	.5000	26.0000	42.0000
MW04	1	4.1500	1.5000	41.0000	54.0000
MW04	2	1.0000	.3000	10.0000	40.0000
MW04	3	1.9500	.3000	24.0000	32.0000
MW04	4	1.2500	.4000	45.0000	28.0000
MW05	1	2.1500	.6000	39.0000	51.0000
MW05	2	1.0000	.4000	26.0000	55.0000
MW05	3	19.6000	.3000	31.0000	60.0000
MW05	4	1.0000	.2000	48.0000	52.0000
MW06	1	1.4000	.8000	22.0000	118.0000
MW06	2	1.0000	.2000	23.0000	66.0000
MW06	3	1.5000	.5000	25.0000	59.0000
MW06	4	20.5500	.4000	28.0000	63.0000
P14	1	2.0500	.2000	10.0000	79.0000
P14	2	1.0500	.3000	10.0000	96.0000
P14	3	5.1000	.5000	10.0000	89.0000

Results of applying both parametric and nonparametric ANOVA to these predisposal data yielded an effect that approached significance for Chemical Oxygen Demand (COD) (p < .072 parametric and p < .066 nonparametric) and a significant difference for Alkalinity (ALK) (p < .002 parametric and p < .009 nonparametric). In terms of individual comparisons, significantly increased COD levels were found for well MW05 (p < .026) and significantly increased ALK was found for wells MW06 (p < .026) and P14 (p < .003) relative to upgradient wells. Of course, these results represent false positives due to spatial variability, since there is no garbage. What is perhaps most remarkable, however, is the absence of any significant results for TOC, where some of the values are as much as 20 times higher than the others. The reason, of course, is that these extreme values tremendously increase the within-well variance estimate, rendering the ANOVA powerless to detect any differences regardless of magnitude. This is yet another testimonial to why it is environmentally negligent to average measurements from downgradient monitoring wells, a problem that is inherent to ANOVA-type analyses when applied to dynamic ground-water quality measurements. The elevated TOC data are clearly inconsistent with chance expectations and should be investigated. In this case, however, they are likely due to insects getting into the wells since this greenfield facility is in the middle of the Mohave desert.

Cochran's Approximation to the Behrens Fisher t-test

Although no longer required, for years the RCRA regulation was based on application of the Cochran's approximation to the Behrens Fisher (CABF) *t*-test. The test was incorrectly implemented by requiring that four quarterly upgradient samples from a single well and single samples from a minimum of three downgradient wells each be divided into four aliquots and treated as if there were $4n$ independent measurements. The net result was that every hazardous waste disposal facility regulated under RCRA was declared "leaking." As an illustration consider the data in Table 3.

Table 3

Illustration of pH data used in computing the CABF t-test.

Date	Replicate				Average
	1	2	3	4	
Background					
11/81	7.77	7.76	7.78	7.78	7.77
02/82	7.74	7.80	7.82	7.85	7.80
05/82	7.40	7.40	7.40	7.40	7.40
08/82	7.50	7.50	7.50	7.50	7.50
$\overline{X}_B$	7.62				7.62
SD_B	0.18				0.20
N_B	16				4
Monitoring					
09/83	7.39	7.40	7.38	7.42	7.40
$\overline{X}_B$	7.40				7.40
SD_B	0.02				
N_B	4				1

Note that the aliquots are almost perfectly correlated and add virtually no independent information yet they are assumed to be completely independent by the statistic. The CABF *t*-test is computed as:

$$t = \frac{\overline{X}_B - \overline{X}_M}{\sqrt{\frac{S^2{}_B}{N_B} + \frac{S^2{}_M}{N_M}}} = \frac{7.62 - 7.40}{\sqrt{\frac{.032}{16} + \frac{.0004}{4}}} = \frac{.22}{.05} = 4.82$$

The associated probability of this test statistic is 1 in 10,000 indicating that the chance that the new monitoring measurement came from the same population as the background measurements is 1 in 10,000. Note that in fact, the mean concentration of the four aliquots for the new monitoring measurement is identical to one of the four mean values for background (*i.e.*, 7.4), suggesting that intuitively the probability is closer to 1 in 4 rather than 1 in 10,000. Averaging the aliquots, which should have never been split in the first place, yields the statistic:

$$t = \frac{\overline{X}_B - \overline{X}_M}{S_B\sqrt{\frac{1}{N_B} + 1}} = \frac{7.62 - 7.40}{.20\sqrt{\frac{1}{4} + 1}} = \frac{.22}{.22} = 1$$

which has an associated probability of 1 in 2. Had the sample size been increased to $N_B = 20$ the probability would have decreased to 1 in 3. In 1988 U.S. EPA recognized this flaw and changed this regulation (see USEPA 1988).

Control of False Positive Rate by Constituent

Site-wide false positive and false negative rates are more important than choice of statistic, nonetheless, certain statistics make it impossible to control the site-wide false positive rate because the rate is controlled separately for each constituent (*e.g.*, parametric and nonparametric ANOVA - see USEPA 1992 section 5.2.1). The only important false positive rate is the one which includes all monitoring wells and all constituents, since any single exceedance can trigger an assessment. This criterion impacts greatly on the selection of statistical method. These error rates are dependent on the number of wells, number of constituents, number of background measurements, type of comparison (*i.e.*, intra-well versus inter-well), distributional form of the constituents, detection frequency of the constituents and the individual comparison false positive rate of the statistic being used. Invariably, this leads to a problem in interval estimation the solution of which is typically a prediction limit that incorporates the effects of verification resampling as well as multiple comparisons introduced by both multiple monitoring wells and multiple monitoring constituents.

Restriction of Background Samples

Certain states have interpreted the Subtitle D regulation as indicating that background be confined to the first four samples collected in a day or a semi-annual monitoring event or a year. The first approach (*i.e.*, four samples in a day) violates the assumption of independence and confounds day to day temporal and seasonal variability with potential contamination. As an analogy, consider setting limits on yearly ambient temperatures in Chicago by taking four temperature readings on July 4th. Say the temperature varied between 75 and 85 degrees on that day yielding a prediction interval from 70 to 90 degrees. As I write this, the temperature in Chicago is -20 degrees. Something is clearly amiss. In the second example of restricting background to the first four events taken in 6 months, the measurements may be independent if ground water flows fast enough, but seasonal variability is confounded with contamination. The net result is that comparisons of background water quality in the summer may not be representative of point of compliance water quality in the winter (*e.g.*, disposal of road salts increasing conductivity in the winter). In the third example in which background is restricted to the first four quarterly measurements, independence is typically not an issue and background versus point of compliance monitoring well comparisons are not confounded with season. However, as previously pointed out, restriction of background to only four samples dramatically increases the size of the statistical prediction limit thereby increasing the false negative rate of the test (*i.e.*, the prediction limit is over five standard deviation units above the background mean concentration). The reason for this is that the uncertainty in the true mean concentration covers the majority of the normal distribution. As such we could obtain virtually any mean and standard deviation by chance alone. If by chance the values are low, false positive results will occur. If by chance the values are high, false negative results will occur. By increasing the background sample size, uncertainty in the sample based mean and standard deviation decrease as does the size of the prediction limit, therefore both false positive and false negative rates are minimized. Furthermore, use of statistical outlier detection procedures applied to the background data will remove the possibility of spurious background results falsely inflating the size of the prediction limit.