Statistical Methods for Ground-Water Monitoring at the USPCI/Laidlaw Grassy Mountain Facility

Prepared

by

Robert D. Gibbons Ph.D.

June 1996

Dr. R.D. Gibbons
Robert D. Gibbons LTD
2021 N. Mohawk
Chicago, IL 60614
312-413-7755

Executive Summary

This report has two specific aims. First to describe a general statistical strategy for ground-water detection monitoring that is applicable at the USPCI/Laidlaw Grassy Mountain Facility and second, to apply this methodology to existing data at the facility. The methodology is first described in considerable detail, appropriately referenced to both the scientific literature and USEPA regulation and guidance and then applied to existing data at the facility. For completeness we describe appropriate statistical methodologies for both inter-well (*i.e.*, upgradient versus downgradient) and intra-well comparisons.

There were several exceedances of upgradient limits for manganese, two for sulfide and two for TSS. The absence of clear historical trends in these wells for these constituents and the absence of VOCs suggest that these differences are due to spatial variability and not an impact from the site. Indeed, there is considerable spatial variability in manganese levels among the four upgradient wells (see Table 1 Appendix A). Intra-well comparisons revealed a single verified exceedance of manganese in well W39 which is also above upgradient limits. Three values have exceeded control limits and are awaiting verification (manganese in W2 and sulfide in W40A and W9). In light of these results, intra-well comparisons are recommended for routine monitoring at this facility. Statistical power analysis based on site specific conditions indicate that the current site-wide false positive rate is much too high (approximately an 80% chance of a verified exceedance of at least one out of 1220 statistical comparisons). To reduce this false positive rate to a reasonable level, a minimum of 8 background samples in each well are required and the number of statistical comparisons should be reduced. The best way to accomplish the latter goal is to reduce the number of monitoring constituents used in the statistical evaluation by selecting a subset that are high in the facility's leachate relative to their concentration in upgradient wells. New leachate data are being collected and a reduced monitoring list of leachate indicator constituents will be proposed.

Specifically, we propose the following:

1. Intra-well comparisons using combined Shewart-CUSUM control charts will be performed for all wells and constituents.

2. For new wells, background will be obtained using an accelerated sampling plan of quarterly sampling for a period of two years. In the interim, new monitoring measurements for those wells with less than eight background samples will be compared to upgradient prediction limits.

3. Every two years all data that are within control limits will be pooled with background and control limits will be recomputed.

4. Intra-well comparisons will also be computed for the upgradient wells to insure that increasing trends are not due to regional or climactic fluctuations.

5. Nonparametric prediction limits will be used for those wells and constituents that have detection frequency less than 25%.

6. New leachate data will be obtained, and leachate concentrations will be compared to upgradient prediction limits. At a later date, we will propose to remove any constituent that is not significantly higher in leachate relative to upgradient ground water.

Table of Contents

Overview

In the context of ground-water monitoring at waste disposal facilities, legislation has required statistical methods as the basis for investigating potential environmental impact due to waste disposal facility operation. Owner/Operators must perform a statistical analysis on a quarterly or semi-annual basis. A statistical test is performed on each of many constituents (*i.e.*, 10 to 50) for each of many wells (5 to 100 or more). The result is potentially hundreds, and in some cases, a thousand or more statistical comparisons performed on each monitoring event. Even if the false positive rate for a single test is small (*e.g.*, 1%), the possibility of failing at least one test on any monitoring event is virtually guaranteed. This assumes you have done the correct statistic in the first place.

In the following sections, a statistical plan is developed that includes: an effective verification resampling plan, and selection of appropriate statistical methods (*e.g.*, parametric and nonparametric prediction limits or control charts for intra-well comparison) that detect contamination when it is present and do not falsely conclude that the site is contaminated. Statistical significance of contamination detection cannot be properly determined without verification resampling. It is noted from the information presented herein that the final statistical detection monitoring plan cannot be fully specified until background samples for the required list of indicator constituents are available. In general, it is unwise to perform statistical computations on any less than eight background samples. This may be four quarterly samples in each of two upgradient wells, or eight samples taken in each well where intra-well comparisons are to be performed. To take any fewer samples will lead to high false negative rates due to the large size of the prediction limit (*i.e.*, with four samples and three degrees of freedom, the uncertainty in the true mean and standard deviation (μ and σ) given the sample based estimates ($\bar{x}$ and s) is enormous, resulting in extremely high prediction limits). Conversely, with only a few background measurements, our knowledge of the true sampling variability, distributional form and detection frequency may be completely inaccurate leading to a high false positive rate.

Yet another major concern is whether the upgradient wells accurately characterize the natural spatial variability that is observed in the downgradient wells. The alternative is to perform intra-well comparisons which are gen-

erally preferable, however, we must first demonstrate that the well has not been impacted by the site. To this end, we will first test the appropriateness of upgradient versus downgradient comparisons for each well and constituent, and in those cases where intra-well comparisons are applicable, demonstrate (1) the absence of any significant trend in that well and constituent and (2) demonstrate the absence of any constituents of concern (*e.g.*, volatile organic priority pollutant list compounds or other constituents that characterize the leachate from the facility and would not be expected in the natural ground water).

It is noted that when justified, intra-well comparisons are always more powerful than their inter-well counterparts because they completely eliminate the spatial component of variability. Due to the absence of spatial variability, the uncertainty in measured concentrations is decreased making intra-well comparisons more sensitive to real releases (*i.e.*, false negatives) and false positive results due to spatial variability are completely eliminated.

The following provides an outline of the general statistical procedure for ground-water monitoring under the Subtitle D regulation, which is also described in the flowchart at the end of this report.

A. Detection Monitoring

1. Upgradient Versus Downgradient Comparisons

   (a) Detection frequency > 50%

      i. If normal, compute normal prediction limit (40CFR 264 Subpart F), selecting false positive rate based on number of wells, constituents and verification resamples (40CFR 264 Subpart F), adjusting estimates of sample mean and variance for nondetects.

      ii. If lognormal, compute a lognormal prediction limit (40CFR 264 Subpart F).

      iii. If neither normal nor lognormal, compute nonparametric prediction limit (40CFR 264 Subpart F) unless background is insufficient to achieve a 5% site-wide false positive rate. In this case, use a normal distribution (40CFR 264 Subpart F).

(b) If the background detection frequency is greater than zero but less than 50%, compute a nonparametric prediction limit and determine if the background sample size will provide adequate protection from false positives. If insufficient data exist to provide a site-wide false positive rate of 5%, more background data must be collected (40CFR 264 Subpart F).

(c) If the background detection frequency equals zero, use the laboratory specific PQL (recommended) or limits required by applicable regulatory agency (40CFR 264 Subpart F). This only applies for those wells and constituents that have at least 13 background samples. Thirteen samples provides a 99% confidence nonparametric prediction limit with one resample (see Table 1). If less than 13 samples are available more background data must be collected.

(d) As an alternative to (c), use a Poisson prediction limit which can be computed from only 4 background measurements regardless of the detection frequency (USEPA, 1992 section 2.2.4).

(e) If downgradient wells fail, determine cause.

i. If the downgradient wells fail because of natural or off-site causes, select constituents for intra-well comparisons (40CFR 264 Subpart F).

ii. If site impacts are found, a site plan for assessment monitoring and detection monitoring (at unaffected wells) may be necessary (40CFR 264 Subpart F).

2. Intra-well Comparisons

(a) For those facilities that either

i. Have no definable gradient,

ii. Have no existing contamination from an on-site-off-site landfill or other source,

iii. Have too few upgradient wells to meaningfully characterize spatial variability (*e.g.*, a site with one upgradient well or a facility in which upgradient water quality is not representative of downgradient water quality),

iv. Satisfy specific hydrogeological criteria (*e.g.*. slow moving ground-water zones, no access to upgradient ground water, inappropriate ground-water migration pathways) as defined by a ground-water professional,

compute intra-well comparisons using combined Shewart-CUSUM control charts (40CFR 264 Subpart F).

(b) For those wells and constituents that fail upgradient versus downgradient comparisons, compute combined Shewart-CUSUM control charts. If no VOCs or hazardous metals are detected and no trend is detected in other indicator constituents. use intra-well comparisons for detection monitoring of those wells and constituents.

(c) If data are all non-detects after 13 quarterly sampling events, use PQL as statistical decision limit (40CFR 264 Subpart F). Thirteen samples provides a 99% confidence nonparametric prediction limit with one resample (40CFR 264 Subpart F) and USEPA 1992 section 5.2.3). Note that 99% confidence is equivalent to a 1% false positive rate, and pertains to a single comparison (*i.e.*, well and constituent) and not the site-wide error rate (*i.e.*, all wells and constituents) that is set to 5%.

(d) If detection frequency is greater than zero (*i.e.*, the constituent is detected in at least one background sample) but less than 25% set control limit to the largest of at least 13 background samples.

(e) As an alternative to (c) and (d) compute a Poisson prediction limit following collection of at least 4 background samples (USEPA 1992 section 2.2.4). Since the mean and variance of the Poisson distribution are the same, the Poisson prediction limit is defined even there is no variability (*e.g.*, even if then constituent is never detected in background). In this case, the reporting limits are used in place of the measurements and the Poisson prediction limit can be computed directly.

3. Verification Resampling

(a) Verification resampling is an integral part of the statistical methodology (USEPA 1992 section 5).

(b) Without verification resampling much larger prediction limits would be required to obtain a site-wide false positive rate of 5%. The resulting false negative rate would be dramatically increased.

(c) Verification resampling allows sequential application of a much smaller prediction limit, therefore minimizing both false positive and false negative rates.

(d) A statistically significant exceedance is not declared and should not be reported until the results of the verification resample are known. The probability of an initial exceedance is much higher than 5% for the site as a whole.

(e) Note that requiring passage of two verification resamples (*e.g.*, in the state of California regulation) will lead to higher false negative rates because larger prediction limits are required to achieve a site-wide false positive rate of 5% than for a single verification resample; hence, the preferred method is one verification resample. Also note that for nonparametric limits, requiring passage of two verification resamples may result in need for a larger number of background samples than are typically available (see Gibbons, 1994).

4. False Positives and False Negative Rates

(a) Conduct simulation study based on current monitoring network, constituents, detection frequencies, and distributional form of each monitoring constituent (USEPA 1992 Appendix B).

(b) Project frequency of verification resamples and false assessments for site as a whole for each monitoring event based on the results of the simulation study.

(c) As a general guideline, we require a site-wide false positive rate of 5% and a false negative rate of approximately 5% for differences on the order of 3 to 4 standard deviation units (see USEPA 1992 Appendix B). Note that following USEPA we simulate the most conservative case of a release that effects a single constituent in a single downgradient well. In practice. multiple constituents in multiple wells will be impacted, therefore, the actual false negative rates will be considerably smaller than estimates obtained via simulation.

5. Use of MDLs and PQLs in Ground-Water Monitoring

   (a) MDLs indicate that the analyte is present in the sample with confidence.

   (b) PQLs indicate that the true quantitative value of the analyte is close to the measured value.

   (c) For analytes with estimated concentration exceeding the MDL but not the PQL, it can only be concluded that the true concentration is greater than zero - there is no way of knowing the actual concentration.

   (d) If the laboratory-specific MDL for a given compound is 3 μg/l, and the PQL for the same compound is 6 μg/l, then a detection of that compound at 4 μg/l could actually represent a true concentration of anywhere between 0 and 6 μg/l. The true concentration may well be *less than* the MDL (see Currie 1968, Hubaux and Vos, 1970 and Gibbons 1994).

   (e) Comparison of such a value to a maximum contaminant level (MCL), or any other concentration limit, is not meaningful unless the concentration is larger than the PQL.

   (f) Verification resampling applies to this case as well.

B. Assessment or Corrective Action Monitoring

1. Comparison to Background

   (a) Define background for any Appendix II compounds detected (*i.e.*, a minimum of four background samples (40CFR 264 Subpart F).

   (b) Compute appropriate prediction limit based on distributional tests and detection frequency as previously described, based on upgradient data or historical data from each well (40CFR 264 Subpart F).

   (c) Compare any Appendix II constituent concentrations found to the background prediction limit. If all values are below the prediction limit for two consecutive sampling events return to detection monitoring (40CFR 264 Subpart F).

(d) In Corrective Action (required if background is exceeded) use same statistic until background is achieved for three years. (40CFR 264 Subpart F). Use Sen's test to evaluate trends (declining) to demonstrate effectiveness of corrective action.

2. Comparison to a Standard

(a) If a maximum contaminant level (MCL) or alternate concentration limit (ACL) is used, and the ACL or MCL is greater than the background prediction limit, then new concentrations in the assessment or corrective action wells should be compared to the standard (*i.e.*, ACL or MCL) using the upper 95% normal confidence limit computed from the last four independent samples (USEPA 1992).

(b) In the case of anthropogenic compounds such as VOCs, if the standard is less than the PQL, then the standard becomes the PQL, since no smaller value can be quantified.

(c) Use Sen's test to evaluate trends (both increasing and decreasing) to demonstrate the effectiveness of corrective action.

C. Implementation

1. The computer program used to implement the detection monitoring plan will encompass all aspects of the previously presented statistical decision tree.

2. The program will be automatic with respect to selection of statistical methods based on the decision tree and all wells and analytes will be input as a complete file and analyzed on the basis of a single instruction. Cumbersome programs such as GRITS/STAT which require extensive user input for analysis of each well and constituent individually will be avoided.

3. Once the program is configured no further statistical decisions, choices or selections will be made so that it can be run by someone with or without adequate statistical background to make these decisions.

4. The program will have a graphical user interface that allows the user to communicate the data format and to add new data to an existing database rather than requiring a complete new database each quarter.

5. The computer program DUMPStat (Downgradient Upgradient Monitoring Program Statistics) distributed by Discerning Systems, Vancouver CA is the only existing program that provides these features.

D. Technical Details

The purpose of this section is to provide a description of the specific statistical methods used in DUMPStat, which is the computer program that will be used in performing the routine statistical analysis of detection monitoring data at the facility. Please note, however, that specific recommendations for any given facility require an interdisciplinary site-specific study that encompasses knowledge of the facility, it's hydrogeology. geochemistry, and study of the false positive and false negative error rates that will result. In general, the appropriate statistical methods are available in DUMPStat, however the program must be properly configured for each site to insure that the methods are properly implemented. Performing a correct statistical analysis, such as nonparametric prediction limits, in the wrong situation (*e.g.*, when there are too few background measurements) can lead to disaster. It is for this reason that DUMPStat's simulation capabilities are so important. In the following, the general DUMPStat algorithm is described.

1. Upgradient Versus Downgradient Comparisons

   For those wells and constituents that show similar variability in upgradient and downgradient monitoring zones inter-well comparisons can be performed by computing limits based on historical upgradient data to which individual new downgradient monitoring measurements can be compared. In the following, the decision rules by which various prediction limits can be computed is outlined. The decision points are based on detection frequency and distributional form of the upgradient data.

(a) Case 1: Compounds Quantified in All Background Samples

i. Test normality of distribution using the multiple group version of the Shapiro-Wilk test (Wilk and Shapiro, 1968) applied to n background measurements. The multiple group version of the original Shapiro-Wilk test (Shapiro and Wilk, 1965) takes into consideration that upgradient measurements are nested within different upgradient monitoring wells, hence the original Shapiro-Wilk test does not apply (USEPA, 1992 section 1.1.4).

ii. If normality is not rejected, compute the 95% prediction limit as:

$$\bar{x} + t_{[n-1,\alpha]} s \sqrt{1 + \frac{1}{n}}$$

where

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

α is the false positive rate for each individual test,

$t_{[n-1,\alpha]}$ is the one-sided $(1-\alpha)100\%$ point of Student's t distribution on $n-1$ degrees of freedom.

and n is the number of background measurements.

iii. Select α as the minimum of .01 or one of the following:

A. Pass the first or one of one verification resample

$$\alpha = \left(1 - .95^{1/k}\right)^{1/2}$$

B. Pass the first or one of two verification resamples

$$\alpha = \left(1 - .95^{1/k}\right)^{1/3}$$

C. Pass the first or two of two verification resamples

$$\alpha = \sqrt{1 - .95^{1/k}}\sqrt{1/2}$$

where k is the number of comparisons (*i.e.*, monitoring wells times constituents - see USEPA 1992 section 5.2.2).

iv. If normality is rejected, take natural logarithms of the n background measurements and recompute the multiple group Shapiro-Wilk test.

v. If the transformation results in a nonsignificant G statistic (*i.e.*, the values $log_e(x)$ are normally distributed - see USEPA 1992 section 1.1), compute the lognormal prediction limit as:

$$exp\left(\bar{y} + t_{[n-1,\alpha]} s_y \sqrt{1 + \frac{1}{n}}\right)$$

where

$$\bar{y} = \sum_{i=1}^{n} \frac{log_e(x_i)}{n}$$

and

$$s_y = \sqrt{\sum_{i=1}^{n} \frac{(log_e(x_i) - \bar{y})^2}{n-1}}$$

vi. If log transformation does not bring about normality (*i.e.*, the probability of G is less than 0.01), compute nonparametric prediction limits as in section 3 (USEPA 1992 section 5.2.3). (Option - compute Poisson prediction limits as in section 3.4 - see USEPA 1992 section 2.2.4).

(b) Case 2: Compounds Quantified in at Least 50% of All Background Samples

i. Apply the multiple group Shapiro-Wilk test to the n_1 quantified measurements only.

ii. If the data are normally distributed compute the mean of the n background samples as:

$$\bar{x} = \left(1 - \frac{n_0}{n}\right) \bar{x}'$$

where $\bar{x}'$ is the average of the n_1 detected values, and n_0 is the number of samples in which the compound is not detected or is below the method detection limit. The standard deviation is:

$$s = \sqrt{\left(1 - \frac{n_0}{n}\right) s^{2'} + \frac{n_0}{n}\left(1 - \frac{n_0 - 1}{n-1}\right) \bar{x}^{2'}}$$

where s' is the standard deviation of the n_1 detected measurements. The normal prediction limit can then be computed as previously described. This method is due to Aitchison (1955) - (see USEPA 1992 section 2.2.2).

iii. If the multiple group Shapiro-Wilk test reveals that the data are lognormally distributed, replace $\bar{x}'$ with $\bar{y}'$ and s' with s'_y in the equations for $\bar{x}$ and s.

iv. The lognormal prediction limit may then be computed as previously described.

v. Note that this adjustment only applies to positive random variables. The natural logarithm of concentrations less than 1 are negative and therefore the adjustment does not apply. For this reason we add 1 to each value (*i.e.*, $log_e(x_i + 1) \geq 0$), compute the prediction limit on a log scale and then subtract one from the antilog of the prediction limit.

vi. If the data are neither normally or lognormally distributed, compute a nonparametric prediction limit. (Option - compute normal prediction limit).

(c) Case 3: Compounds Quantified in less than 50% of All Background Samples

i. In this application, the nonparametric prediction limit is the largest concentration found in n upgradient measurements (USEPA 1992 section 4.2.1).

ii. Gibbons (1990, 1991) has shown that the confidence associated with this decision rule, following one or more verification resamples, is a function of the multivariate extension of the hypergeometric distribution (USEPA 1992 section 5.2.3).

iii. Complete tabulations of confidence levels for $n = 4, \ldots, 100$, $k = 1, \ldots, 100$ future comparisons (*e.g.*, monitoring wells), and a variety of verification resampling plans are presented in Gibbons (1994). For example with 5 monitoring wells and 10 constituents (*i.e.*, 50 comparisons). we would require 40 background measurements to provide 95% confidence (USEPA 1992 section 5.2.3). Table 1 displays confidence levels for a single verification resample.

iv. As an option to the nonparametric prediction limits, DUMPStat can compute Poisson prediction limits. Poisson prediction limits are useful for those cases in which there are too few background measurements to achieve an adequate site-wide false positive rate using the nonparametric approach. Gibbons (1987) derived the Poisson prediction limit as

$$\text{Poisson PL} = y/n + \frac{t^2}{2n} + t/n\sqrt{y(1+n) + t^2/4}\,.$$

where y is the sum of the detected measurements or reporting limit for those samples in which the constituent was not detected and t is the $(1-\alpha)100$ upper percentage point of Student's t-distribution (USEPA 1992 section 2.2.4). More recent work in this area suggests that a more conservative approach is to substitute the normal multiplier z for t using a value of *alpha* as previously described. The normal multiplier is now used in DUMPStat.

TABLE 1

PROBABILITY THAT THE FIRST SAMPLE OR THE VERIFICATION RESAMPLE WILL BE BELOW THE MAXIMUM OF n BACKGROUND MEASUREMENTS AT EACH OF k MONITORING WELLS FOR A SINGLE CONSTITUENT

Previous	Number of Monitoring Wells (k)														
n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	.933	.881	.838	.802	.771	.744	.720	.698	.679	.661	.645	.630	.617	.604	.592
5	.952	.913	.879	.849	.823	.800	.779	.760	.742	.726	.711	.697	.684	.672	.661
6	.964	.933	.906	.882	.860	.840	.822	.805	.789	.774	.761	.748	.736	.725	.714
7	.972	.947	.925	.905	.886	.869	.853	.838	.825	.812	.799	.788	.777	.766	.757
8	.978	.958	.939	.922	.906	.891	.878	.864	.852	.841	.830	.819	.809	.800	.791
9	.982	.965	.949	.935	.921	.908	.896	.885	.874	.864	.854	.844	.835	.827	.818
10	.985	.971	.957	.945	.933	.922	.911	.901	.891	.882	.873	.865	.857	.849	.841
11	.987	.975	.964	.953	.942	.933	.923	.914	.906	.897	.889	.882	.874	.867	.860
12	.989	.979	.969	.959	.950	.941	.933	.925	.917	.910	.902	.896	.889	.882	.876
13	.990	.981	.973	.964	.956	.948	.941	.934	.927	.920	.914	.907	.901	.895	.889
14	.992	.984	.976	.969	.961	.954	.948	.941	.935	.929	.923	.917	.912	.906	.901
15	.993	.986	.979	.972	.966	.959	.953	.947	.942	.936	.931	.926	.920	.915	.910
16	.993	.987	.981	.975	.969	.964	.958	.953	.948	.943	.938	.933	.928	.923	.919
17	.994	.988	.983	.978	.972	.967	.962	.957	.953	.948	.943	.939	.935	.930	.926
18	.995	.990	.985	.980	.975	.970	.966	.961	.957	.953	.949	.944	.940	.937	.933
19	.995	.991	.986	.982	.977	.973	.969	.965	.961	.957	.953	.949	.946	.942	.938
20	.996	.991	.987	.983	.979	.975	.972	.968	.964	.960	.957	.953	.950	.947	.943
25	.997	.994	.992	.989	.986	.984	.981	.978	.976	.973	.971	.968	.966	.964	.961
30	.998	.996	.994	.992	.990	.988	.986	984	.983	.981	.979	.977	.975	.974	.972
35	.998	.997	.996	.994	.993	.991	.990	.988	.987	.986	.984	.983	.981	.980	.979
40	.999	.998	.997	.995	994	.993	.992	.991	.990	.989	988	.987	.985	.984	.983
45	.999	.998	.997	.996	.995	.995	.994	.993	.992	.991	.990	.989	.988	.987	.987
50	.999	.998	.998	.997	.996	.996	.995	.994	.993	.993	.992	.991	.990	.990	989
60	.999	.999	.998	.998	.997	.997	.996	.996	.995	.995	.994	.994	.993	.993	.992
70	1.00	.999	.999	.998	.998	.998	.997	.997	.997	.996	.996	.995	.995	.995	.994
80	1.00	.999	.999	.999	.998	.998	.998	.998	.997	.997	.997	.996	.996	.996	.996
90	1.00	1.00	.999	.999	.999	.999	.998	.998	.998	.998	.997	.997	.997	.997	.996
100	1.00	1.00	.999	.999	.999	.999	.999	.998	.998	.998	.998	998	997	.997	.997

Previous	Number of Monitoring Wells (k)														
n	20	25	30	35	40	45	50	55	60	65	70	75	80	90	100
4	.542	.504	.474	.449	428	.410	.394	.380	.367	.356	.345	336	.327	.312	299
5	.612	.574	.543	.517	.495	.476	.459	443	.430	417	.406	396	.386	.369	.355
6	.668	.631	.600	.574	.552	.532	.514	.499	.484	.472	.460	.449	.439	.420	.405
7	.713	.678	.648	.623	.600	.580	.563	.547	.532	.519	.507	.496	.485	.466	.450
8	.750	.717	.688	.664	.642	.622	.605	.589	.574	561	.549	537	.527	.507	.490
9	.781	.750	.723	.699	.678	.659	.642	626	.612	.598	.586	574	.564	.544	.527
10	.807	.777	.752	.729	.709	691	.674	659	644	.631	.619	.608	.597	.578	.560
11	.828	.801	.777	.755	.736	.718	.702	.687	.674	661	.649	.638	.627	.608	.590
12	847	.821	799	.778	.760	.743	.727	713	.700	687	.675	.664	.654	.635	.618
13	862	.839	.817	.798	.781	.764	.750	.736	.723	.711	.699	.689	.678	.660	.643
14	.876	.854	.834	.816	.799	.784	.769	756	.744	.732	.721	.710	.701	.682	.666
15	.888	.867	.848	.831	.815	.801	.787	774	.762	.751	.740	.730	.721	.703	.686
16	.898	.879	.861	.845	.830	.816	.803	.791	.779	.768	.758	.748	.739	722	706
17	.907	.889	.872	.857	843	.830	.817	.806	.794	.784	.774	765	.756	739	.723
18	.914	.898	.882	.868	.855	.842	.830	819	.808	.798	.789	.780	.771	.754	.739
19	.921	.906	.891	.878	865	.853	.842	.831	.821	.811	.802	.793	.785	.769	.754
20	.928	.913	.899	.886	.874	.863	.852	.842	.832	.823	.814	.806	.798	.782	.768
25	.950	.939	.929	.919	.910	.901	.892	.884	.876	.869	.862	.855	.848	.835	.823
30	.963	.955	.947	.940	.932	.925	.919	912	.906	.900	.894	888	.882	872	.861
35	.972	.966	.959	.954	.948	.942	.937	931	.926	.921	.916	.911	.907	898	889
40	.978	.973	.968	.963	.958	.954	.949	.945	.941	.936	.932	.928	.924	.917	909
45	.982	.978	.974	.970	.966	.962	.959	955	951	948	.944	.941	.938	.931	925
50	.985	.982	.979	975	.972	.969	.966	.963	.959	.956	.954	.951	.948	942	.937
60	.990	.987	.985	.982	.980	.978	.975	973	.971	.968	.966	.964	.962	958	954
70	.992	.990	.989	.987	.985	.983	.981	990	.978	.976	.974	.973	971	.968	.965
80	.994	.993	.991	.990	.988	.987	.986	984	.983	.981	.980	.979	977	.975	.972
90	.995	.994	.993	.992	.991	.990	988	987	.986	.985	.984	.983	.982	.980	.978
100	.996	.995	.994	.993	.992	.991	.991	990	.989	988	.987	.986	985	.983	.982

2. Intra-Well Comparisons

One particularly good method for computing intra-well comparisons is the combined Shewart-CUSUM control chart (USEPA 1992 section 6.1).

The method is sensitive to both gradual and rapid releases and is also useful as a method of detecting "trends" in data. Note that this method should be used on wells unaffected by the landfill. There are several approaches to implementing the method and in the following one useful way is described as well as discussion of some statistical properties.

(a) Assumptions

The combined Shewart-CUSUM control chart procedure assumes that the data are *independent* and *normally* distributed with a *fixed* mean μ and constant variance σ^2. The most important assumption is independence, and as a result wells should be sampled *no more* frequently than quarterly. In some cases, where ground-water moves relatively quickly, it may be possible to accelerate background sampling to eight samples in a single year; however, this should only be done to establish background and not for routine monitoring. The assumption of normality is somewhat less of a concern, and if problematic, natural log or square root transformation of the observed data should be adequate for most practical applications. For this method, nondetects can be replaced by the method detection limit without serious consequence. This procedure should *only* be applied to those constituents that are detected at least in 25% of all samples, otherwise, σ^2 is not adequately defined.

(b) *Nondetects*

i. For those well and constituent combinations in which the detection frequency is less than 25%, we will provide graphical display of these data until a sufficient number of measurements are available to provide 99% confidence (*i.e.*, 1% false positive rate) for an individual well and constituent using a nonparametric prediction limit, which in this context is the maximum detected value out of the n historical measurements. As previously discussed this amounts to 13 background samples for 1 resample, 8 background samples for pass 1 of 2 resamples and 18 background samples for pass 2 of 2 resamples. It should be obvious that if nonparametric prediction limits are to be used for intra-well comparisons of rarely detected constituents, two verification resamples will often be required and failure will

only be indicated if *both* measurements exceed the limit (*i.e.*, the maximum of the first 8 samples).

ii. For those cases in which the detection frequency is greater than 25%, DUMPStat substitutes the median reporting limit for the nondetects. In this way, changes in reporting limits do not appear to be significant trends.

iii. If nothing is detected in 8, 13 or 18 independent samples (depending on resampling strategy), DUMPStat uses the reporting limit as the control limit.

iv. As in the previously described inter-well comparisons, DUMPStat provides optional use of Poisson prediction limits as an alternative to nonparametric prediction limits for rarely detected constituents (*i.e.*, less than 25% detects). Poisson prediction limits can be computed after 8 background measurements regardless of detection frequency.

(c) Procedure

i. DUMPStat requires that at least 8 historical independent samples are available to provide reliable estimates of the mean μ and standard deviation σ, of the constituent's concentration in each well.

ii. DUMPStat selects the three Shewart-CUSUM parameters h (the value against which the cumulative sum will be compared), k (a parameter related to the displacement that should be quickly detected), and SCL (the upper Shewart limit which is the number of standard deviation units for an immediate release). Lucas (1982) and Starks (1988) suggest that $k = 1$, $h = 5$, and $SCL = 4.5$ are most appropriate for ground-water monitoring applications. This sentiment is echoed by USEPA in their interim final guidance document *Statistical analysis of ground-water monitoring data at RCRA facilities* (April, 1989). Also see USEPA 1992 section 6.1. For ease of application, however, we have selected $h = SCL = 4.5$, which is slightly more conservative than the value of $h = 5$ suggested by USEPA.

iii. Denote the new measurement at time-point t_i as x_i.

iv. Compute the standardized value z_i

$$z_i = \frac{x_i - \bar{x}}{s}$$

where $\bar{x}$ and s are the mean and standard deviation of the at least 8 historical measurements for that well and constituent (collected in a period of no less than one year).

v. At each time period, t_i, compute the cumulative sum S_i, as

$$S_i = \max[0, (z_i - k) + S_{i-1}]$$

where max[A, B] is the maximum of A and B, starting with $S_0 = 0$.

vi. Plot the values of S_i (y-axis) versus t_i (x-axis) on a time chart. Declare an "out-of-control" situation on sampling period t_i if for the first time, $S_i \geq h$ or $z_i \geq SCL$. Any such designation, however, must be verified on the next round of sampling, before further investigation is indicated.

vii. The reader should note that unlike prediction limits which provide a fixed confidence level (*e.g.*, 95%) for a given number of future comparisons, control charts do not provide explicit confidence levels, and do not adjust for the number of future comparisons. The selection of $h = SCL = 4.5$ and $k = 1$ is based on USEPA's own review of the literature and simulations (see Lucas, 1982; Starks, 1988; and USEPA, 1989). USEPA indicates that these values "allow a displacement of two standard deviations to be detected quickly." Since 1.96 standard deviation units corresponds to 95% confidence on a normal distribution, we can have approximately 95% confidence for this method as well.

viii. In terms of plotting the results, it is more intuitive to plot values in their original metric (*e.g.*, μg/l) rather than in standard deviation units. In this case $h = SCL = \bar{x} + 4.5s$ and the S_i are converted to the concentration metric by the transformation $S_i * s + \bar{x}$, noting that when normalized (*i.e.*, in standard deviation units) $\bar{x} = 0$ and $s = 1$ so that $h = SCL = 4.5$ and $S_i * 1 + 0 = S_i$.

ix. When $n \geq 12$ Starks (1988) and USEPA (1992) suggest that $k = .75$, and $h = SCL = 4.0$ provide more conservative control

limits and this approach is now used in DUMPStat.

(d) Outliers

i. From time to time, inconsistently large or small values (outliers) can be observed due to sampling, laboratory, transportation, transcription errors, or even by chance alone. The verification resampling procedure that we have proposed will tremendously reduce the probability of concluding that an impact has occurred if such an anomalous value is obtained for any of these reasons. However, nothing has eliminated the chance that such errors might be included in the historical measurements for a particular well and constituent. If such erroneous values (either too high or too low) are included in the historical database, the result would be an artificial increase in the magnitude of the control limit, and a corresponding increase in the false negative rate of the statistical test (*i.e.*, conclude that there is no site impact when in fact there is).

ii. To remove the possibility of this type of error, the historical data are screened for each well and constituent for the existence of outliers (USEPA 1992 section 6.2) using the well known method described by Dixon (*Biometrics*, 1953, 9, 74-89). These outlying data points are indicated on the control charts (using a different symbol), but are excluded from the measurements that are used to compute the background mean and standard deviation. In the future, new measurements that turn out to be outliers, in that they exceed the control limit, will be dealt with by verification resampling in downgradient wells only.

iii. This same outlier detection algorithm is applied to each upgradient well and constituent to screen outliers for inter-well comparisons as well.

(e) Existing Trends

If contamination is pre-existing, trends will often be observed in the background database from which the mean and variance are computed. This will lead to upward biased estimates and grossly inflated control limits. To remove this possibility, we first screen the background data for each well and constituent for trend using Sen's

(1986) nonparametric estimate of trend. Confidence limits for this trend estimate are given by Gilbert (1987). A significant trend is one in which the 99% lower confidence bound is greater than zero. In this way, even pre-existing trends in the background dataset will be detected.

(f) A Note on Verification Sampling

i. It should be noted that when a new monitoring value is an outlier, perhaps due to a transcription error, sampling error, or analytical error, the Shewart and CUSUM portions of the control chart are affected quite differently. The Shewart portion of the control chart compares each individual new measurement to the control limit, therefore, the next monitoring event measurement constitutes an independent verification of the original result. In contrast, however, the CUSUM procedure incorporates *all* historical values in the computation, therefore, the effect of the outlier will be present for both the initial and verification sample; hence the statistical test will be invalid.

ii. For example, assume $\bar{x} = 50$, and $s = 10$. On quarter 1 the new monitoring value is 50, so $z = (50 - 50)/10 = 0$ and $S_i = \max[0, (z - 1) + 0] = 0$. On quarter 2, a sampling error occurs and the reported value is 200, yielding $z = (200 - 50)/10 = 15$ and $S_i = \max[0, (15 - 1) + 0] = 14$, which is considerably larger than 4.5; hence an initial exceedance is recorded. On the next round of sampling, the previous result is not confirmed, because the result is back to 50. Inspection of the CUSUM, however, yields $z = (50 - 50)/10 = 0$ and $S_i = \max[0, (0 - 1) + 14] = 13$, which would be taken as a confirmation of the exceedance, when in fact, no such confirmation was observed. For this reason, the verification must *replace* the suspected result in order to have an unbiased confirmation.

(g) Updating the Control Chart

i. As monitoring continues and the process is shown to be in control, the background mean and variance should be updated periodically to incorporate these new data. Every year or two, all new data that are *in control* should be pooled with the

initial samples and $\bar{x}$ and s recomputed. These new values of $\bar{x}$ and s will then be used in constructing future control charts. This updating process should continue for the life of the facility and/or monitoring program (USEPA 1992 section 6.2).

ii. DUMPStat allows the user to update background by changing the time window menu option. This option sets a window of time for which background summary statistics are computed. Changing the maximum date will incorporate new data into the background limit estimate. Note that this time window applies to computing background for both inter-well and intra-well comparisons.

(h) An Alternative Based on Prediction Limits

i. An alternative approach to intra-well comparisons involves computation of well-specific prediction limits. Prediction limits are somewhat more sensitive to immediate releases but less sensitive to gradual releases than the combined Shewart-CUSUM control charts. Prediction limits are also less robust to deviations from distributional assumptions.

ii. As an alternative to combined Shewart-CUSUM control charts DUMPStat can compute normal prediction limits as described in the previous section on inter-well comparisons.

iii. For detection frequencies greater than 25%, nondetects are replaced with the median reporting limit. For detection frequencies less than 25%, either nonparametric or Poisson prediction limits are computed depending on what option the user has selected (*i.e.*, rare-event statistic window).

3. Comparison to a Standard

(a) For assessment or corrective action, it is often required that samples from a potentially impacted well be compared to a ground-water quality protection standard such as an MCL or ACL. DUMPStat's assessment monitoring module provides tabular and graphical display of this comparison based on tests of increasing and decreasing trend and comparison of the standard to the upper 95% normal confidence limit applied to the last four independent samples.

(b) The 95% confidence limit for the mean of the last four measurements is computed as

$$\bar{x} + t_{[3,.05]}\frac{s}{2} .$$

(c) Nondetects are replaced by one-half of the reporting limit since with only four measurements, more sophisticated statistical adjustments are not appropriate.

E. Some Methods to be Avoided

In the following sections some statistical methods that should be avoided are described.

1. Analysis of Variance - ANOVA

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

(a) 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).

(b) ANOVA is more sensitive to spatial variability than contamination. Spatial variability effects mean concentrations but typically not the variance, hence small yet consistent differences will achieve statistical significance. In contrast, contamination effects 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.

(c) Nonparametric ANOVA is often presented by USEPA 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 effect only one of 10 or 20 monitoring wells (USEPA 1992 section 5.2.1).

(d) 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. USEPA may like ANOVA because it will appear to be more powerful than prediction and tolerance limits for single future values. The increased power, however, is only realized when all four measurements from a single well are equally affected by the site impact which on average will only occur 25% of the time (*i.e.*, if four semi-annual sampling events are evenly spaced, all four will be impacted by a new release only one in four times). 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).

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 WasteTech Conference Proceedings*, Charleston SC, 1/14/94).

TABLE 2

Raw Data for All Detection Monitoring
Wells and Constituents (mg/l)
This Landfill 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.

2. Cochran's Approximation to the Behrens Fisher t-test

   Although no longer required, for years the USEPA 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
2/82	7.74	7.80	7.82	7.85	7.80
5/82	7.40	7.40	7.40	7.40	7.40
8/82	7.50	7.50	7.50	7.50	7.50
$\bar{X}_B$		7.62			7.62
SD_B		0.18			0.20
N_B		16			4
Monitoring					
9/83	7.39	7.40	7.38	7.42	7.40
$\bar{X}_B$		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

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

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, 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

$$t = \frac{\bar{X}_B - \bar{X}_M}{S_B\sqrt{\frac{1}{N_B}+1}} = \frac{7.62 - 7.40}{.20\sqrt{\frac{1}{4}+1}} = \frac{.22}{.22} = 1.0$$

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. It took U.S. EPA six years to recognize this flaw and to change this regulation (see USEPA 1988).

3. 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.

4. 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 in the site-specific illustration, 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.

F. Results of Application at the USPCI/Laidlaw Grassy Mountain Facility

In the following, results of site-specific analysis of the existing monitoring program are described.

1. Monitoring Well Network

   A list of upgradient and downgradient monitoring wells are provided in the following Table.

Current Upgradient and Downgradient Monitoring Wells

Upgradient	Downgradient
PZ06	W10
PZ07	W11
PZ08	W12
W1	W18
	W19A
	W2
	W21
	W22
	W23
	W24
	W25
	W27
	W28
	W29A
	W30A
	W32A
	W33
	W34
	W35
	W36
	W37A
	W38A
	W39
	W40A
	W41
	W42
	W43
	W44
	W45
	W46
	W47
	W48
	W49
	W5
	W50
	W51
	W52
	W53
	W54
	W55
	W56
	W57
	W58
	W59
	W60
	W67
	W68
	W69
	WT0
	W71
	W72
	W73
	W74
	W75
	W8
	W9

A list of the constituents used in the analysis is provided in the following Table.

Constituents used in the Analysis

Constituent
Arsenic (total)
Barium (total)
Beryllium (total)
Cadmium (total)
Chromium (total)
Copper (total)
Dissolved solids (total)
Lead (total)
Manganese (total)
Mercury (total)
Molybdenum (total)
Nickel (total)
pH
Purgable organic halides
Selenium (total)
Silver (total)
Sulfide
Suspended solids (total)
Total organic carbon
Zinc (total)
VOCs

For the purpose of this initial analysis, background was set to all data prior to 1995.

2. Upgradient versus Downgradient Comparisons

Results of upgradient versus downgradient comparisons are presented in Appendix A. All historical data for each downgradient well and constituent is displayed graphically along with the upgradient prediction limit (*i.e.*, horizontal line). All historical upgradient data were used in computing the prediction limits, hence the shaded background time line

covers the entire x-axis. Raw upgradient data with outliers indicated are displayed in Table 1 for all constituents. Current downgradient monitoring results with statistical exceedances noted are displayed in Table 2. Comparison of detection frequencies in upgradient and downgradient wells is presented in Table 3. Tests of distributional form and corresponding type of prediction limit selected are displayed in Table 4. Computed prediction limit values and intermediate statistics for normal and lognormal prediction limits and confidence levels for nonparametric prediction limits are displayed in Table 5. Historical data for those downgradient monitoring wells that exceeded an upgradient prediction limit (whether they were verified or not) are displayed in Table 6.

Inspection of Table 1 reveals considerable spatial variability as reflected in differences between the four background wells (*e.g.*, see arsenic in Table 1 of Appendix A). This spatial variability limits the usefulness of upgradient versus downgradient comparisons because spatial variability will be confused with a potential site impact.

Inspection of Table 2 of Appendix A (and graphs at the end of this report) reveals verified exceedances of upgradient limits for manganese in W10, W11, W2, W27, W30A, W38A, W39, W44, W45, W46, W58, W59, W69 and W70, sulfide in wells W24 and W60 and total suspended solids in wells W30A and W51. Inspection of historical data for these wells and constituents (see Table 6 and graphs at end of the report and/or in Appendix A) reveal that in general, these concentrations have historically exceeded upgradient background with either no evidence of increasing trend or gradual trends over time.

3. Intra-well Comparisons

Given (1) the presence of spatial variability, (2) the absence of any detected volatile organic compounds (which are present in large concentrations in the facility's leachate) and (3) the absence of any significant trend in historical concentrations, intra-well comparisons are the method of choice. Combined Shewart-CUSUM control charts are displayed graphically for all wells and constituents in Appendix B. Summary statistics and intermediate computations are displayed in Table 1 of Appendix B. All wells and constituents were automatically tested for trend using Sen's nonparametric test prior to analysis. A single significant verified

significant increase was detected for manganese in well W39 (see Graphs at end of report and/or Appendix B). Three values have exceeded control limits and are awaiting verification (manganese in W2 and sulfide in W40A and W9). In addition, the following significant trends were found (TDS in W12, W18, W25, W26, W9; Manganese W11 and W46; TOC in W33). In general, these trends are quite gradual and may reflect changes in sampling and analytical protocols over the last five to 10 years of the data record. These results are also consistent with chance expectations given that we have performed 1120 statistical tests.

4. Statistical Power

   Statistical power curves for the facility-wide false positive and false negative rates are presented at the end of each Appendix. For upgradient versus downgradient comparisons the false positive rate is 61% and the test becomes sensitive to 3 to 4 standard deviation unit increases over background. For intra-well comparisons the false positive rate is 83% and the test becomes sensitive to 2 to 3 standard deviation unit increases over background. This means that there is an 83% chance of at least one verified exceedance (*i.e.*, for one well and constituent) out of the 1220 statistical comparisons performed. These high false positive rates are due to the huge number of comparisons (*i.e.*, 61 wells and 20 constituents) and the fact that some of these wells have quite limited background databases. These estimates were, however, based only on those wells and constituents that had a minimum of eight background samples.

5. VOCs

   Inspection of Table 1 in Appendix C reveals that there have been a scattered detections of bis(2-ethylhexyl)phthalate in both upgradient and downgradient wells. 2,4,6-trichlorophenol has also been occasionally detected, however, there are no clear trends or consistent detections of VOCs in any well.

6. Proposed Statistical Methods

   Specifically, we propose the following:

(a) Intra-well comparisons using combined Shewart-CUSUM control charts will be performed for all wells and constituents.

(b) For new wells, background will be obtained using an accelerated sampling plan of quarterly sampling for a period of two years. In the interim, new monitoring measurements for those wells with less than eight background samples will be compared to upgradient prediction limits.

(c) Every two years all data that are within control limits will be pooled with background and control limits will be recomputed.

(d) Intra-well comparisons will also be computed for the upgradient wells to insure that increasing trends are not due to regional or climactic fluctuations.

(e) Nonparametric prediction limits will be used for those wells and constituents that have detection frequency less than 25%.

(f) New leachate data will be obtained, and leachate concentrations will be compared to upgradient prediction limits. At a later date, we will propose to remove any constituent that is not significantly higher in leachate relative to upgradient ground water.

7. Summary

There were several exceedances of upgradient limits for manganese, two for sulfide and two for TSS. The absence of clear historical trends in these wells for these constituents and the absence of VOCs suggest that these differences are due to spatial variability and not an impact from the site. Indeed, there is considerable spatial variability in manganese levels among the four upgradient wells (see Table 1 Appendix A). Intra-well comparisons revealed a single verified exceedance of manganese in well W39 which is also above upgradient limits. Three values have exceeded control limits and are awaiting verification (manganese in W2 and sulfide in W40A and W9). In light of these results, intra-well comparisons are recommended for routine monitoring at this facility. Statistical power analysis based on site specific conditions indicate that the current site-wide false positive rate is much too high (approximately an 80% chance of a verified exceedance of at least one out of 1220 statistical comparisons). To reduce this false positive rate to a reasonable level,

a minimum of 8 background samples in each well are required and the number of statistical comparisons should be reduced. The best way to accomplish the latter goal is to reduce the number of monitoring constituents used in the statistical evaluation by selecting a subset that are high in the facility's leachate relative to their concentration in upgradient wells. New leachate data are being collected and a reduced monitoring list of leachate indicator constituents will be proposed.

Some Relevant Literature

References

[1] Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association*, **50**, 901-908.

[2] Currie, L.A. (1968). Limits for qualitative detection and quantitative determination: Application to radiochemistry. *Analytical Chemistry*, **40**, 586-593.

[3] Davis, C. B. & McNichols, R. J. (1987). One-sided intervals for at least p of m observations from a normal population on each of r future occasions. *Technometrics*, **29**, 359-370.

[4] Davis, C. B. (1993). Environmental regulatory statistics. in *Handbook of Statistics, Vol. 12: Environmental Statistics* G.P. Patil & C.R. Rao, editors, Elsevier.

[5] Dixon, W. J. (1953). Processing data for outliers. *Biometrics*, **9**, 74-89.

[6] Gibbons, R. D. (1987). Statistical prediction intervals for the evaluation of ground-water quality. *Ground Water*, **25**, 455-465.

[7] Gibbons, R. D. (1987). Statistical models for the analysis of volatile organic compounds in waste disposal facilities. *Ground Water*, **25**, 572-580.

[8] Gibbons, R. D., Jarke, F. H., & Stoub, K. P. (1989). Method detection Limits. Proceedings of *Fifth Annual USEPA Waste Testing and Quality Assurance Symposium*, Vol. **2**, 292-319.

[9] Gibbons, R. D. (1990). A general statistical procedure for Ground-Water Detection Monitoring at waste disposal facilities. *Ground Water*, **28**, 235-243.

[10] Gibbons, R. D. (1990). Estimating the precision of ground-water elevation data. *Ground Water*, **28**, 357-360.

[11] Gibbons, R. D., Grams N. E., Jarke F. H., & Stoub K. P. (1990). Practical quantitation limits. *Proceedings of Sixth Annual USEPA Waste Testing and Quality Assurance Symposium Vol 1*, 126-142.

[12] Gibbons, R. D. & Baker J. (1991). The properties of various statistical prediction limits. *Journal of Environmental Science and Health, A26-4*, 535-553.

[13] Gibbons, R. D. (1991). Statistical tolerance limits for ground-water monitoring. *Ground Water*, **29**.

[14] Gibbons, R. D. (1991). Some additional nonparametric prediction limits for ground-water monitoring at waste disposal facilities. *Ground Water*, **29**, 729-736.

[15] Gibbons, R. D., Jarke, F. H., & Stoub, K. P. (1991). Detection Limits: for linear calibration curves with increasing variance and multiple future detection decisions. *Waste Testing and Quality Assurance*, **3**, ASTM SPT 1075, 377-390.

[16] Gibbons, R. D., Grams, N. E., Jarke, F. H., & Stoub, K. P. (1992). Practical quantitation limits: *Chemometrics and Intelligent Laboratory Systems*, **12**, 225-235.

[17] Gibbons, R. D. (1992). An overview of statistical methods for ground-water detection monitoring at waste disposal facilities. IN *Ground-Water Contamination at Hazardous Waste Sites: Chemical Analysis*, S. Lesage & R.E. Jackson (eds.), New York: Marcel Dekker, Inc.

[18] Gibbons, R. D., Dolan, D., Keough H., O'Leary, K. & O'Hara R. (1992). A comparison of chemical constituents in leachate from industrial hazardous waste & municipal solid waste landfills. Proceedings of the *Fifteenth Annual Madison Waste Conference*, University of Wisconsin, Madison.

[19] Gibbons, R. D. *Statistical Methods for Ground-Water Monitoring*, John Wiley & Sons, 1994.

[20] Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring*, Van Nostrand Reinhold, New York.

[21] Hubaux, A. and Vos, G. (1970). Decision and detection limits for linear calibration curves. *Analytical Chemistry,* **42**, 849-855.

[22] Lucas, J. M. (1982). Combined Shewart-CUSUM quality control schemes. *Journal of Quality Technology,* **14**, 51-59.

[23] Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 63,* 1379-1389.

[24] Shapiro, S.S., and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika,* **52**, 591-611.

[25] Starks T. H. (1988). Evaluation of control chart methodologies for RCRA waste sites. USEPA technical report CR814342-01-3.

[26] USEPA, 40CFR Part 264: Statistical methods for evaluating groundwater monitoring from hazardous waste facilities; final rule. *Federal Register,* 53, 196 (1988) 39720-39731.

[27] USEPA, Interim Final Guidance Document *Statistical analysis of groundwater monitoring data at RCRA facilities* (April, 1989).

[28] USEPA, Addendum to Interim Final Guidance Document *Statistical analysis of ground-water monitoring data at RCRA facilities* (July, 1992).

[29] Wilk, M.B., and Shapiro, S.S. (1968). The joint assessment of normality of several independent samples. *Technometrics,* **10, no 4.** 825-839.

APPENDIX A

Site-Specific Results

Upgradient Versus Downgradient

Comparisons

APPENDIX B

Site-Specific Results

Intra-Well Comparisons

APPENDIX C

Site-Specific Results

Historical VOC Detections