Risk Analysis
DOI: 10.1111/risa.12820

Incremental Sampling Methodology: Applications for Background Screening Assessments

Penelope S. Pooler,[1] Philip E. Goodrum ———,[2,*] Deana Crumbling,[3] Leah D. Stuchal,[4] and Stephen M. Roberts[4]

This article presents the findings from a numerical simulation study that was conducted to evaluate the performance of alternative statistical analysis methods for background screening assessments when data sets are generated with incremental sampling methods (ISMs). A wide range of background and site conditions are represented in order to test different ISM sampling designs. Both hypothesis tests and upper tolerance limit (UTL) screening methods were implemented following U.S. Environmental Protection Agency (USEPA) guidance for specifying error rates. The simulations show that hypothesis testing using two-sample *t*-tests can meet standard performance criteria under a wide range of conditions, even with relatively small sample sizes. Key factors that affect the performance include unequal population variances and small absolute differences in population means. UTL methods are generally not recommended due to conceptual limitations in the technique when applied to ISM data sets from single decision units and due to insufficient power given standard statistical sample sizes from ISM.

KEY WORDS: Background screening assessment; composite sampling; hypothesis testing; incremental sampling methodology (ISM); risk assessment

1. INTRODUCTION

Many chemicals that are listed as target analytes due to their potential for contributing to unacceptable risks in human and ecological risk assessments, or service losses in natural resource damage assessments, are also present naturally or are ubiquitous at low concentrations due to anthropomorphic activities. For these chemicals, concentrations found in environmental media must be interpreted in the context of background levels. Common examples include inorganics found naturally in soil, sediment, and groundwater, and organics formed naturally and/or widely dispersed from a variety of human activities such as polycyclic aromatic hydrocarbons. A determination of whether or not concentrations found in a specific area are elevated above background often plays a critical role in developing risk management strategies.

In practice, it is rare to have historical data to characterize background concentrations of chemicals for a site before contamination might have occurred, so background conditions are determined by collecting samples from one or more nearby reference areas that have similar physical, chemical, geological, and biological characteristics as the site.[1] With appropriate sampling designs applied to both site and reference areas, and well-defined tolerances for decision

[1]Syracuse University, Whitman School of Management, Syracuse, NY, USA.
[2]Integral Consulting, Inc., Fayetteville, NY, USA.
[3]U.S. Environmental Protection Agency, Cleanup Technology Innovation Program, Office of Superfund Remediation & Technology Innovation, Washington, DC, USA.
[4]University of Florida, Center for Environmental & Human Toxicology, Gainesville, FL USA.
*Address correspondence to Philip E. Goodrum, Integral Consulting, Inc., 7030 E Genesee Street, Suite 105, Fayetteville, NY 13066, USA; tel: +315-396-6655; pgoodrum@integral-corp.com.

errors, a determination whether site concentrations represent background conditions is readily accomplished through statistical analysis. The analysis may indicate that the concentrations at the site are elevated, at which point a chemical would be determined to be a contaminant and further evaluated in the risk assessment. Alternatively, the analysis may show that the differences in the sampling distributions are sufficiently small to conclude that the site and reference areas share the same population distribution. In this case, the chemical concentrations at the site would be considered to be within the range of background conditions and the chemical would "screen out" from further investigation.[1–3]

The availability of regulatory guidance and user-friendly software has helped to promote consistency in the application of statistical concepts to inform sampling designs as well as to conduct background screening assessments with the data generated.[1,3–9] Practitioners are generally familiar with strengths and limitations of various approaches, including defining upper-bound statistics for "background" (e.g., 95/95 upper tolerance limits [UTLs]) and two-sample hypothesis tests. But this experience is largely with sampling of environmental media with discrete grab samples of small volumes, which hereafter will be referred to as "discrete sampling."[10] Recently, there has been considerable interest in using incremental sampling methodology (ISM), a form of statistically structured composite sampling that utilizes elements of Gy sampling theory to collect material from across a decision unit (DU), the area over which a single decision will be made.[10,11] The regulatory and regulated communities are increasingly opting to replace discrete samples with ISM because ISM is more scientifically defensible, practical, relatively easy to implement (and reproduce) in the field and laboratory, cost effective for most investigation objectives, and readily communicated to stakeholders. Perhaps the strongest incentive is the potential for ISM to greatly reduce decision errors that can lead to excessive contaminant exposures or unnecessary costly remedial action as compared to discrete sampling methods.

ISM sampling strategies typically divide a DU into gridded subareas (cells) and a uniform amount of soil is sampled from each cell. These small soil samples (termed "increments"), typically numbering 30–100 over a DU, are combined into a single composite sample. The concentrations of chemicals of interest are measured after the composite has undergone specific laboratory processing and subsampling procedures that minimize matrix heterogeneity and subsampling error. The field ISM sample concentration represents an estimate of the mean chemical concentration within the DU derived through physical averaging of systematic, randomly selected soil from across the DU.[10] This process can be repeated with increments taken from different locations within the cells to generate replicate ISM samples, each of which is an independent estimate of the DU mean.

For any environmental sampling protocol, a statistical sample provides an estimate of the underlying population distribution. In the case of ISM, each physical sample is also the average of a statistical sample, so the population of ISM samples within a DU is a distribution of means. This is of particular interest for statistical tests with ISM data sets because the process of compositing a relatively large set of increments, of 30 or more, motivates practitioners to evoke the central limit theorem, which states that the distribution of sample means tends toward normal as the sample size (number of increments for ISM) increases, regardless of the shape of the underlying distribution. In practice, the degree to which the distribution shape approaches normal is determined by the contaminant distribution at the scale of an increment and the choice of sampling design parameters, including the area or volume of the DU, the volume of soil collected for each increment, and the number of increments in each ISM sample. The relevant sample statistics for ISM are (1) the mean of the field replicate samples, sometimes thought of as "the mean of the means" or "the grand mean" and (2) the standard deviation (*SD*) of the replicates. The ratio of the *SD* divided by the mean is the relative standard deviation, also known as the coefficient of variation (CV). In contrast, when a discrete sampling approach is used, each discrete sample represents the concentration for a vanishingly small volume of the DU, and the results are averaged arithmetically to obtain an estimate of the mean concentration over the entire DU. Variability in concentrations among discrete samples represents heterogeneity in spot-to-spot concentrations across the DU rather than variability in estimates of the mean as with ISM samples. Thus, while both ISM and discrete sampling can generate estimates of the mean, the variability terms (*SD* and CV) have different meanings and different implications for statistical analyses.

There are a number of practical issues or questions that must be addressed when using ISM data in a background screening assessment. A common question is whether or not data from ISM and

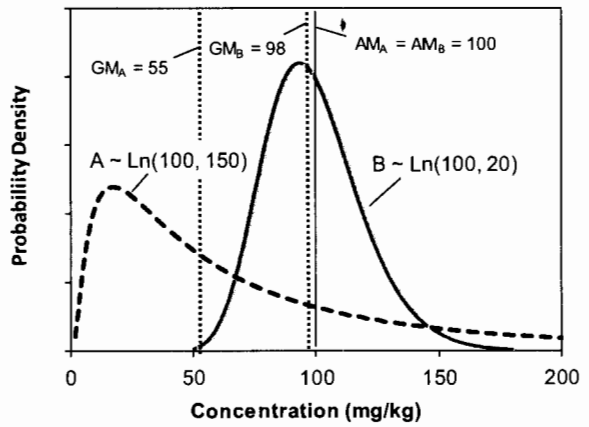

Fig. 1. Examples of population distributions for discrete sampling (graph A) and ISM sampling (graph B) applied to the same decision unit. The distribution for (A) is lognormal with an arithmetic mean (AM) and standard deviation of (100, 150) and corresponding geometric mean (GM, equivalent to median) of 55 mg/kg. The distribution for (B) is lognormal (100, 20) with GM = 98 mg/kg.

discrete data sets can be compared (e.g., background data derived using discrete sampling and site data obtained through ISM or vice versa). The problem with this type of comparison is illustrated as follows. Consider a DU where both ISM sample results and discrete sample results are generated. One sampling event involves collecting 30 individual masses of soil from locations evenly spaced across the DU. Each soil mass is evenly split in a manner that controls matrix heterogeneity. One-half of each split is analyzed for chemical concentration (yielding 30 concentration values) and the other half is used to create a single composite sample, which is analyzed to give a single ISM concentration result. Repeat this sampling event many times and plot the discrete and ISM sample results. In the example shown in Fig. 1, the population distribution for discrete sampling (graph A) is lognormal with a mean (μ) of 100 mg/kg, *SD* (σ) of 150 mg/kg, and geometric mean (GM) of 55 mg/kg. The population distribution for ISM sampling (graph B) shares the same mean of 100 mg/kg, but the *SD* is lower (20 mg/kg), which makes the GM (equivalent to the median for lognormal distributions) closer to the arithmetic mean. The *SD* for the ISM data is expected to be lower because it represents variability in estimates of the mean concentration for the whole DU mass (several tons), while the *SD* for the discrete data represents variability in the concentrations of the individual tiny soil masses actually analyzed (maybe 10 g each). Correspondingly, the CV (σ/μ) for graph A is 1.5, whereas the CV for graph B is 0.2. Note that Fig. 1 is an idealized example where the population distributions are known, which in reality is never the case. However, if we had generated two data sets of random samples from this distribution—one with discrete sampling and the other with an ISM sampling design—we would clearly have had a very high chance of concluding (incorrectly) that the samples were obtained from two areas with different degrees of contamination, even though the samples were collected from the very same DU. Because the statistical tests depend in part on the *SD* for the populations being compared, and *SD* has different meaning for discrete and ISM data, it is an invalid comparison. In view of this, it should be evident that statistical analyses should not be applied to compare data sets that are a mix of discrete and ISM samples.

A second question is whether or not the formal hypothesis testing approaches commonly used to compare background and site data sets from discrete sampling designs can be used with ISM data. Most hypothesis tests are designed so that when the assumptions of the test are met, the test will perform as specified.[6] In practical terms, this means that if a large number of repeated trials were conducted, the frequency of falsely deciding a site is not contaminated when in fact it is would be no greater than the specified error rate.[6] The ability of various statistical tests to provide type I and type II error rates that are not higher than expected has been evaluated using discrete data sets relevant to environmental contamination. Information on performance of these tests is the basis for extensive guidance and software tools available from EPA to assist with the selection of the appropriate type(s) of tests given the environmental conditions and properties of the sample data.[1,6–9,12,13] A similar evaluation has not been conducted for the types of data generated by ISM. Of particular interest is the ability of hypothesis testing conducted with ISM samples to have sufficient statistical power given the limited degrees of freedom, as represented by the number of observations (i.e., ISM replicates) that will typically be available for individual site and reference area DUs. Related to this is the question of whether a difference in the sampling designs underlying two ISM data sets may introduce bias in the results of a statistical test. The sampling designs for each ISM field event may include different specifications for the total volume of soil that comprises a replicate. The simplest example of this is when a different number of

increments is included in the composite. For example, one ISM event may be based on three replicates of 30 increments apiece (for a total of 90 increments across the DU) and a second ISM event may be based on three replicates of 20 increments apiece (for a total of 60 increments across the DU). The ISM sampling design with the larger number of increments will not only provide greater spatial coverage across the DU, but also, it can be expected to yield a lower standard deviation (on average) for the set of replicates. This is because subareas of high and low concentrations that may exist throughout a DU (i.e., the "chemical footprint") will tend to be represented with greater precision as the sample size increases. The ratio of standard deviations of population distributions with different sample sizes (number of increments) will be roughly proportional to the ratio of the inverse of the square root of the number of increments. Continuing with the example of $n = 30$ and $n = 20$ increments:

$$\frac{\sigma_{30}}{\sigma_{20}} \approx \frac{\frac{1}{\sqrt{30}}}{\frac{1}{\sqrt{20}}} = 0.816$$

so the (population) standard deviation for the ISM replicates based on 30 increments will be approximately 20% lower than the standard deviation based on 20 increments.

In addition to the sample size, the total volume that comprises a replicate may also depend on the volume of material that comprises each increment (referred to as "sample support") (ITRC 2012). For example, one ISM sampling event with three replicates and 30 increments may call for 1 g of soil per increment, whereas the second ISM sampling event calls for three replicates and 30 increments of 2 g of soil, thereby doubling the total volume of soil. The spatial coverage (increments × replicates) may be the same, but the design with greater sample support will tend to exhibit lower variability across replicates (thereby changing the variance of the underlying population distribution). Specific recommendations and equations for calculating sample support volumes are discussed in the ITRC Technical Support Document and in citations included therein (ITRC 2012). For this article, we apply two simplifying assumptions: (1) the sample support is the same for each hypothetical ISM sampling event; and (2) all of the sources of variability are collectively represented by the population distribution that is defined for the DU—no attempt is made to apportion this variability to various factors associated with sample collection and processing.

The fact that the population distribution represented by a sample depends on the sampling design has important implications for the way in which we use sample results in site investigations. Most importantly, when comparing data from different DUs (such as is done with background screening assessments), the comparisons should be based on samples generated with similar ISM sampling designs to the extent practicable.

A simulation study was conducted to examine the performance of standard statistical methods for comparing site and background concentrations using ISM data with different sample sizes, distribution variances, and magnitude of differences in site and background means. Because actual conditions at a site are never fully known, but can only be approximated by a sample, we can use numerical simulations to systematically evaluate questions regarding statistical performance of a sampling design.[6,8,10] By defining the population distribution for both the reference and site areas, we can simulate thousands of hypothetical sampling events and determine the frequency of making "decision errors," such as concluding the site and reference areas are the same when in fact the concentrations at a site are elevated. The following questions are examined here with respect to background screening assessments:

(1) Are there thresholds for minimum sample size (increments and replicates) and/or maximum distribution variance when using hypothesis tests to compare the distributions of one site DU and a single reference area DU?
(2) Does the performance of hypothesis tests with ISM depend on whether or not the population variances for background and site populations are equal?
(3) Can nonparametric hypothesis testing methods be considered with small sample sizes typically generated by ISM (e.g., 3–7 replicates)?
(4) Does a UTL (e.g., 95/95 UTL) concept apply when data are based on ISM sampling (measures of the mean across a single DU) rather than discrete sampling (measures of variability at spatial scales much smaller than the DU)?

Based on the findings from this investigation, recommendations for future investigations and

Table I. Specifications of Lognormal Population Distributions for Individual ISM Increments Collected from Background (BG) and Potentially Impacted (PI) Decision Units

Lognormal Population Distribution Specifications	
CV (σ/μ) Ranges Both PI-DU and BG-DU	$\Delta = \ln(\mu_{PI}) - \ln(\mu_{BG})$
0.1–1.0	[–0.1 to 1.0] in increments of 0.1
1.0–2.0	
2.0–4.0	
4.0–6.0	
6.0–8.0	

σ = population standard deviation.
μ_{BG} = arithmetic mean of background area (set equal to 1.0).
μ_{PI} = arithmetic mean of potentially impacted area.
Δ = difference between the log-transformed population mean concentrations μ_{BG} and μ_{PI}; since $\mu_{BG} = 1.0$ and $\ln(\mu_{BG}) = 0$, $\Delta = \ln(\mu_{PI}) - 0 = \ln(\mu_{PI})$.
CV = coefficient of variation.
ln = natural logarithm (base e).
Note: Each of the five CV ranges for the PI-DU was paired with each of the five CV ranges for the BG-DU, resulting in 25 pairs. These 25 sets of variability conditions were examined for each of 12 Δ values.

performance assessments of background screening with ISM are provided at the end of the article.

2. METHODS

2.1. Simulation Scenarios

Numerical simulations were conducted to evaluate a range of plausible scenarios for background screening. Simulation scenarios were developed by first defining conditions of chemical contamination in surface soil at a potentially impacted decision unit (PI-DU) and a single background decision unit (BG-DU), and then applying hypothetical sampling designs to those conditions. A range of lognormal population distributions for ISM sampling was defined by specifying five different CVs for each DU, for a total of 25 possible combinations (Table I). The population distribution in this context refers to the distribution of concentrations in 1 g increments (prior to the compositing step).

The arithmetic mean for the potentially impacted DU (μ_{PI}) was determined by specifying 12 different levels of delta, Δ, which is the difference in the log-transformed population means, $\ln(\mu_{PI}) - \ln(\mu_{BG})$. For simplicity, the arithmetic mean for the background DU (μ_{BG}) was set equal to a constant, 1.0 (unitless), for each scenario. Therefore, since ln(1) is zero, Δ simplifies to $\ln(\mu_{PI})$. Observations were randomly selected from the lognormal distributions and log-transformed prior to performing statistical analysis in order to achieve approximate normal distributions for each data set.

Table II. Delta (Δ) Conversion from log (Base e) Scale to Raw (Untransformed) Data Scale

$\Delta = \ln(\mu_{PI}) - \ln(\mu_{BG})$	μ_{PI}/μ_{BG}	$\mu_{PI} - \mu_{BG}$	% Difference
–0.1	0.90	–0.10	–10%
0.0	1.00	0.00	0%
0.1	1.11	0.11	11%
0.2	1.22	0.22	22%
0.3	1.35	0.35	35%
0.4	1.49	0.49	49%
0.5	1.65	0.65	65%
0.6	1.82	0.82	82%
0.7	2.01	1.01	101%
0.8	2.23	1.23	123%
0.9	2.46	1.46	146%
1.0	2.71	1.71	171%

See Table I for definitions.
% Difference = $100\% \times (\mu_{PI} - \mu_{BG}) / \mu_{BG}$
$\mu_{BG} = 1$, so $(\mu_{PI} / \mu_{BG}) = (\mu_{PI} / 1) = \mu_{PI}$

Table II summarizes the 12 different values for Δ (i.e., $\ln(\mu_{PI})$) along with the percent difference between the population means on a raw (untransformed) scale. For this study, scenarios include differences in site and background means ranging from –0.1 to 1.0 on a log scale, equivalent to –10% to +171% on an untransformed scale. These 12 conditions for Δ represent common situations when concentrations at the potentially impacted site may be slightly elevated, but also include a condition where the site mean is less than the background mean ($\Delta = -0.1$ on a log scale) and the site and background means are equal ($\Delta = 0$).

For each of the 300 population distribution specifications summarized in Tables I and II, eight different ISM sampling designs were applied using different replicate and increment combinations. As summarized in Table III, the scenarios were selected to examine the effects of varying both the increments and replicates in order to retain the same total number of samples (represented by the product of increments × replicates), and varying only the number of replicates (and holding the increments constant), thereby changing the total number of samples for each scenario.

Hypothetical ISM sampling for 2,400 scenarios (300 populations × 8 statistical sample size combinations) was repeated 1,000 times, for a total

Table III. Eight ISM Sample Designs were Examined with Different Combinations of Replicate and Increment Sample Sizes

No. of Replicates (r)	No. of Increments (n)	Total Field Effort ($r \times n$)	Purpose
3	20	60	Examine effect of changing number of replicates
4	20	80	
5	20	100	
6	20	120	
7	20	140	
3	33	99	Examine effect of tradeoff between increments and replicates for the same field effort
7	14	98	
10	10	100	

Note: For each design, 1,000 sampling events were simulated at random for each of the population distribution specifications for both the PI-DU and BG-DU given in Table I.

Table IV. Definitions of Type I and Type II Error Rates (Based on Table 3.3 in USEPA[1])

Decision Based on Sample Data	Actual Site Conditions	
	H_0 is true	H_0 is not true
H_0 is not rejected	Correct decision $(1-\alpha)$	Type II error false negative (β)
H_0 is rejected	Type I error false positive (α)	Correct decision $(1-\beta)$

Confidence level = 100% × $(1-\alpha)$; likelihood of type I error increases as α increases (confidence increases).
Power = 100% × $(1-\beta)$; likelihood of type II error increases as β increases (power decreases).

of 2,400,000 trials. For each trial, a background screening assessment was conducted on the sample results obtained for the paired background DU and potentially impacted site DU. Random sampling, statistical analysis, and corresponding data summaries including the frequency of decision errors for each scenario were performed in R (Version 3.1.2). (See online Supplementary File for the source code.)[14,15]

2.2. Hypothesis Tests

As shown in Table IV, two error rates are relevant in hypothesis testing. The false positive, or type I error, occurs when we incorrectly reject the null in favor of the alternative. The significance level, α, is the probability of making a type I error. The lower the α, that a hypothesis test can achieve, the greater our confidence level in the test result. The false negative, or type II error, occurs when we incorrectly fail to reject the null in favor of the alternative. The probability of making a type II error is given by β. The lower the β, the greater the power of the hypothesis test. We state that a hypothesis test is underpowered if we (incorrectly) fail to reject the null hypothesis more often than would be expected by random chance. Conversely, the test is overpowered if we reject the null hypothesis too frequently.

There are two fundamental forms of the null hypothesis for background screening assessments. Using terminology from EPA guidance,[1,6,13] Background Test Form 1 refers to the case where we state that the mean (or median) of the PI-DU is less than or equal to that of the BG-DU unless data provide evidence to the contrary. Background Test Form 2 reverses this condition, but also adds an extra term, S, the user defined "Substantial Difference" between the population parameters. With Test Form 2, we state that the site mean concentration exceeds the background mean concentration by S or more.

The following three hypotheses tests were applied to each trial data set in order to examine the performance of ISM sampling for both background test forms of the null hypothesis as well as differences in S for Test Form 2:

Hypothesis Test	Background Test Form	One-Sided Hypotheses
1	1	$H_0: \Delta \leq 0$ versus $H_A: \Delta > 0$
2	2	$H_0: \Delta > 0.1$ versus $H_A: \Delta \leq 0.1$
3	2	$H_0: \Delta > 0.2$ versus $H_A: \Delta \leq 0.2$

where $\Delta = \log(\mu_{PI}) - \log(\mu_{BG})$. Delta is shown on the log scale in figures, and the raw data equivalent difference and % difference are given in Table II.

As with most statistical tests, performance metrics for two-sample hypothesis tests are contingent

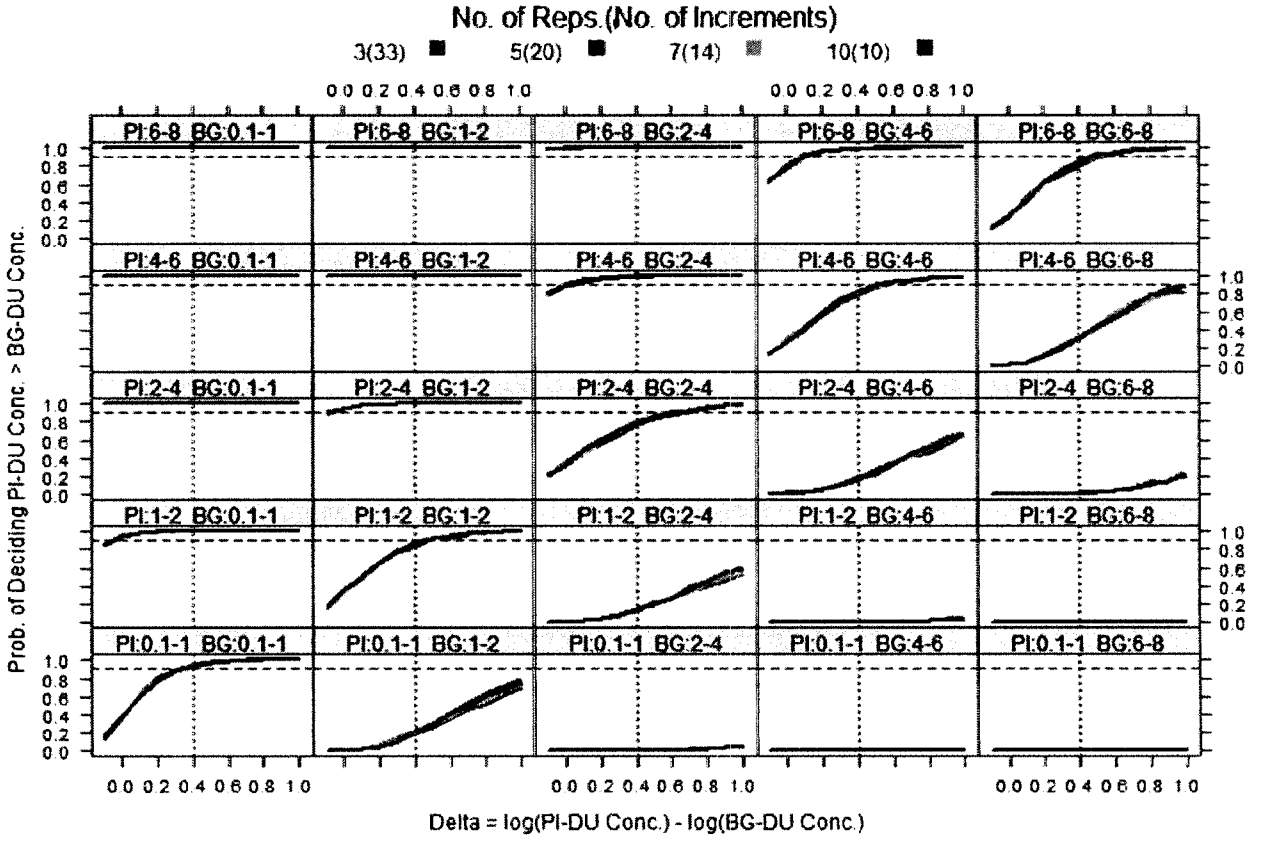

Fig. 2. Results for hypothesis 1, Background Test Form 1. Probability of deciding that the log-transformed mean concentration for PI-DU ($\log(\mu_{PI})$)is greater than $\log(\mu_{BG})$ based on results from a two-sample t-test assuming equal variance. The null hypothesis (hypothesis 1) is $\log(\mu_{PI}) < \log(\mu_{BG})$. The heading of each plot shows the population CV for both the PI-DU and BG-DU. Different colored lines represent distinct ISM sampling designs with roughly the same number of total increments sampled. The horizontal line at probability = 0.9 represents the minimal acceptable power (90%, so $\beta = 0.1$) for the specified confidence level of 80% (or $\alpha = 0.2$). The vertical line at $\Delta = 0.4$ indicates the delta for which minimal power is achieved in designs where CVs are equal. Color graphics are available in the online version of this article.

on the populations adhering to a few assumptions. For "parametric" tests (e.g., Student's t-test), we assume the population variances are equal and the shape of both population distributions is normal. For some parametric tests (e.g., Welch's or Satterthwaite's t), we can relax the assumption of equal variance while still requiring an assumption of normality. A nonparametric alternative to the Student's t-test, the Wilcoxon Mann–Whitney (WMW) test, assumes that the population variances are identical, though not necessarily normal.[6,8]

Typically, when an exploratory analysis suggests that assumptions are significantly violated, the Welch's test is recommended for unequal variance, and the WMW test is recommended for nonnormal distributions.[1,7,8,13] The performance of Student's t-test is considered to be robust to moderate differences in population variances (e.g., less than a factor of three). This rule of thumb can be examined in simulation studies by running both Student's t and Welch's t-test on each trial.

Nonparametric hypothesis tests were not applied in this study for two reasons. First, with ISM sampling, the number of replicates is typically too small (e.g., $n < 10$) to evaluate distribution assumptions either graphically or with statistical goodness-of-fit tests. Second, small data sets typically lack the power required to reject the null with any supportable probability, whether parametric or nonparametric tests are used. Nonparametric hypothesis tests compare rank statistics from the two groups rather than evaluating differences in the means or medians.

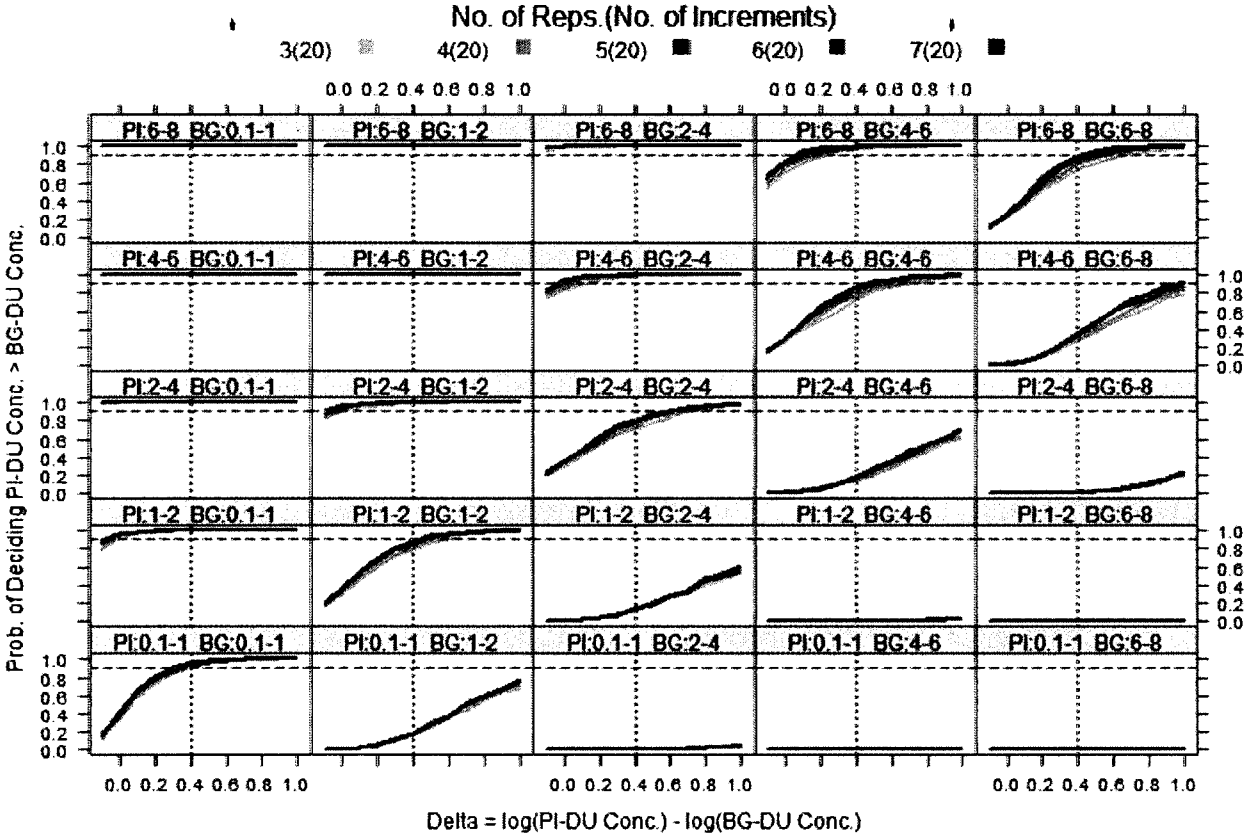

Fig. 3. Same as Fig. 2, except the sample sizes ($r \times n$) illustrate the sensitivity of the test performance to change in replicates ($r = 3$–7) for a fixed number of increments ($n = 20$). Color graphics are available in the online version of this article.

Because of this, these methods have substantially lower power than their parametric counterpart, especially when sample sizes are very small, and are essentially ineffective.[16,17] For example, with sample sizes of three ISM replicates for the PI-DU and BG-DU, even if all three observations for the PI-DU were greater than the maximum from the BG-DU, we would have insufficient power to reject a Background Test Form 1 null hypothesis that $\mu_{PI} \leq \mu_{BG}$ with a nonparametric test; therefore, the outcome of the background screening assessment would be based exclusively on which background test form was selected, rather than the measured concentrations.

For this simulation study, all data were log-transformed prior to running Student's t or Welch's t-test. Importantly, this approach means that these are tests of medians because the mean of a normal distribution of logs is the median of the corresponding arithmetic values.[6] The use of hypothesis tests applied to log-transformed data is acknowledged as a viable approach in USEPA guidance on hypothesis testing for background screening. In 2007, USEPA also used numerical simulations to examine the performance of Welch's t-test applied to log-transformed data in an assessment of alternative hypothesis testing approaches applied to samples collected with discrete sampling protocols.[6] In that study, USEPA concluded: "*The Welch's log t-test performed similarly to the [non-parametric] test for Test Form 1. For those specific scenarios where the [non-parametric] test with Test Form 1 was equal to or better than the untransformed t-tests with Test Form 2, the Welch's log t-test equaled or slightly outperformed the [non-parametric] test.*" USEPA's current (2015) guidance on the use of ProUCL software does indicate a preference for a nonparametric test applied to nontransformed data over a t-test applied to log-transformed data, but with the limited rationale that "a t-test on log-transformed data tests the equality of medians and not the equality of means."[9] However, in that same guidance, USEPA acknowledges

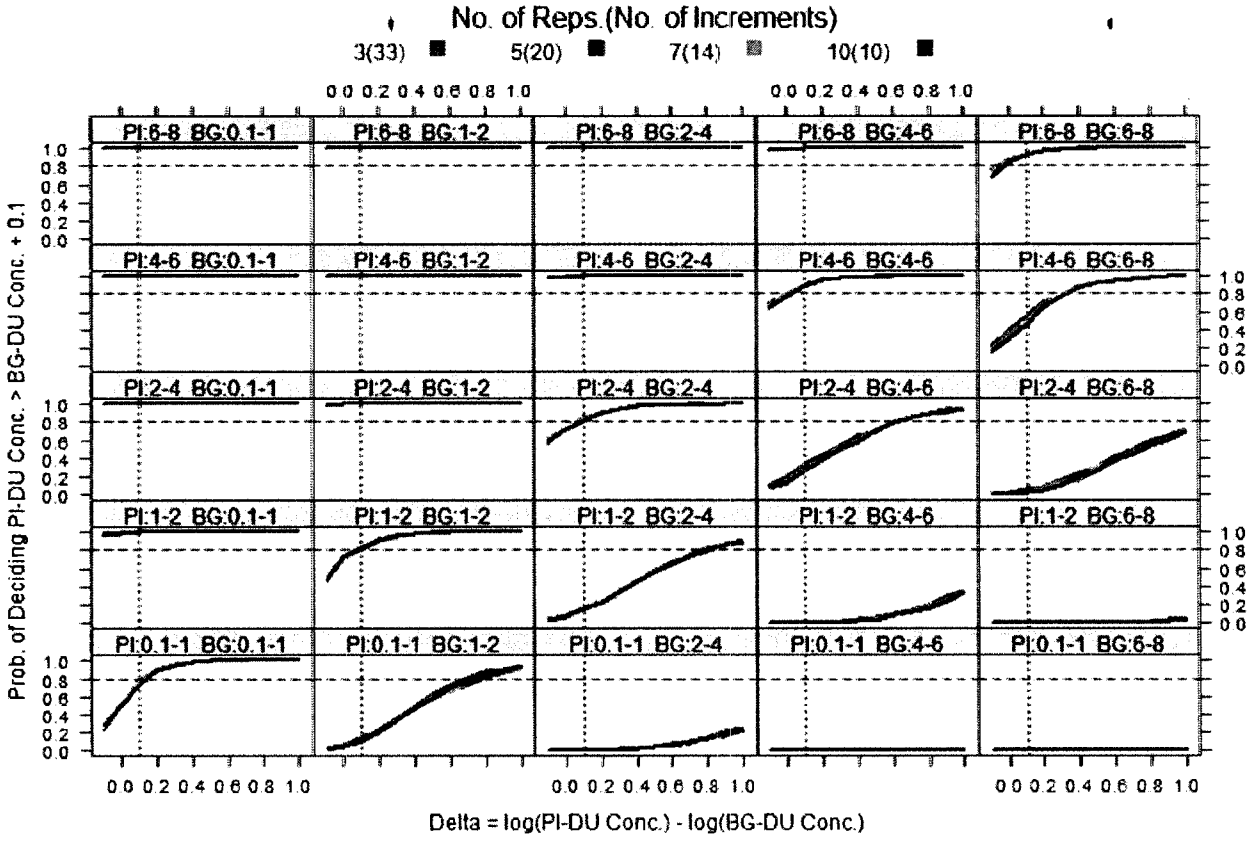

Fig. 4. Results for hypothesis test 2, Background Test Form 2. Probability of deciding that the log-transformed mean concentration for the PI-DU ($\log(\mu_{PI})$) is greater than $\log(\mu_{BG}) + 0.1$ based on results from a two-sample t-test assuming equal variance, which occurs if we fail to reject the null hypothesis (hypothesis 2, H_0: $\log(\mu_{PI}) > \log(\mu_{BG}) + 0.1$). Results are grouped by population CV ranges for the PI-DU and BG-DU. For example, the bottom-left corner graphic shows the results for population CVs ranging from 0.1 to 1.0 for both DUs. Different colored lines indicate the (replicate × increment) sampling design. The horizontal line at probability = 0.8 represents the minimal acceptable power (80%, so $\beta = 0.20$) for the specified confidence level (90%, so $\alpha = 0.10$). The vertical line at $\Delta = 0.1$ indicates the specified significant difference. Color graphics are available in the online version of this article.

that if the distributions are skewed (nonnormal), nonparametric methods also evaluate differences in the medians rather than the means.[9] Therefore, the main reason that USEPA appears to guide users to selecting a nonparametric method (for discrete data) is that nonparametric methods exhibit favorable performance under conditions of nonnormality as well as moderate censoring (i.e., one or more nondetects). As noted above, nonparametric methods are generally not an option for ISM data sets due to insufficient sample sizes (number of replicates).

Results from testing hypothesis 1 were summarized by estimating the probability of deciding that the log-transformed mean concentration of the PI-DU is greater than the log-transformed mean concentration of the BG-DU based on 1,000 sampling events. Results from testing hypotheses 2 and 3 were summarized by estimating the probability of deciding that the log-transformed mean concentration of the PI-DU is greater than the log-transformed mean concentration of the BG-DU + S based on the 1,000 sampling events. Empirical probabilities were estimated for a confidence level of 80% ($\alpha = 0.20$) for hypothesis 1 (EPA Background Test Form 1), and a confidence level of 90% ($\alpha = 0.10$) for hypotheses 2 and 3 (EPA Background Test Form 2), consistent with EPA guidance.[1]

2.3. Upper Tolerance Limit

The UTL is the one-sided upper confidence limit of an upper percentile of the BG-DU data set. This is

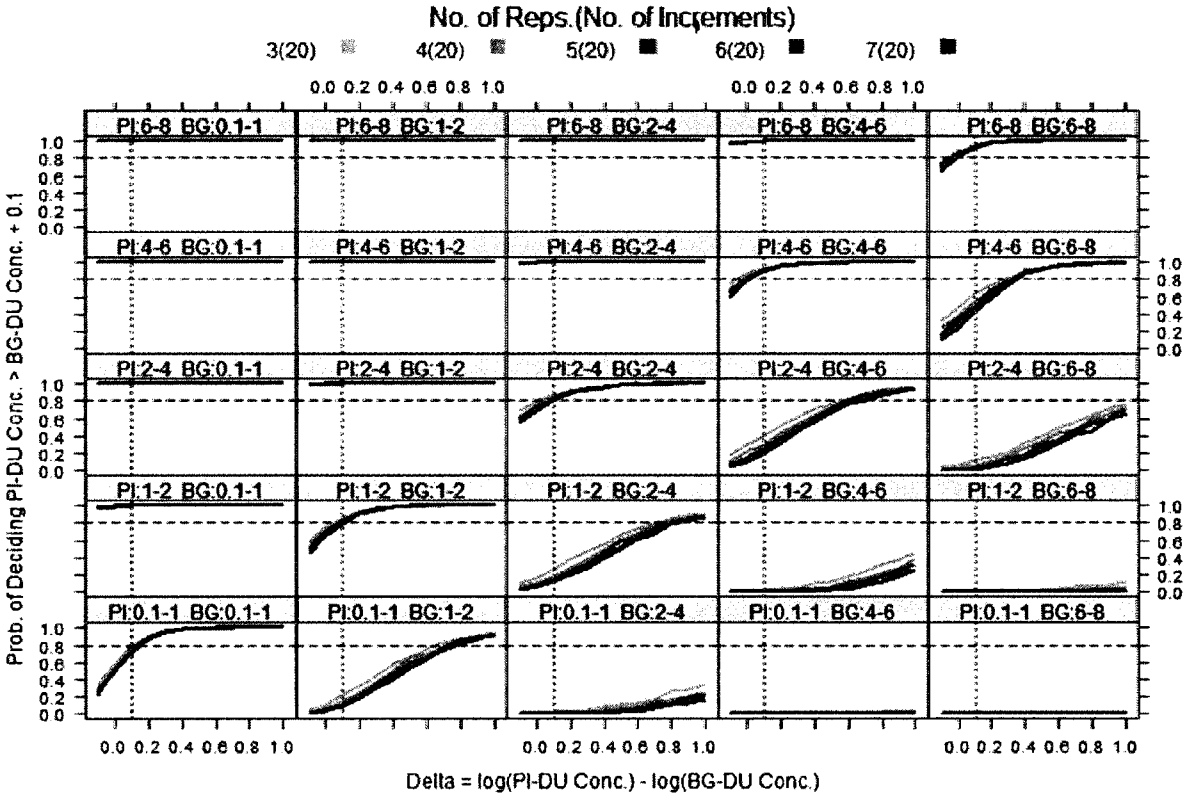

Fig. 5. Same as Fig. 4, except the sample sizes ($r \times n$) illustrate the sensitivity of the test performance to change in replicates ($r = 3$–7) for a fixed number of increments ($n = 20$). Color graphics are available in the online version of this article.

the second technique commonly used in background screening assessments that are traditionally applied to discrete samples. If one or more observations from the PI-DU exceeds the UTL, there is evidence that the site has higher concentrations than would be expected by chance alone if the population distributions were the same at the PI-DU and BG-DU. There are two design criteria for this statistic—the quantile (or percentile) and the confidence level.[8,18] In this simulation study, consistent with convention for background screening assessments, the UTL is calculated for the 95th percentile and 95% upper confidence limit, or "95/95 UTL."[8] As described above for hypothesis testing, the number of replicates with ISM is generally insufficient to rely on nonparametric (rank order) methods. Therefore, only parametric methods are considered in this study.

For simplicity, the conventional lognormal 95/95 UTL was calculated for the BG-DU data set (ISM replicates) for each trial. The lognormal 95/95 UTL is calculated as:

$$UTL_{0.95}\ (x_{0.95}) = \bar{x}\ + s K_{0.95,0.95}$$

where $\bar{x}$ and s are the mean and standard deviation of the log-transformed data for BG-DU, and K is a tolerance factor for estimating the 95% upper bound on the 95th quantile given the sample size (i.e., number of ISM replicates). K factors for the simulated ISM samples presented here were calculated based on the noncentral t-distribution using the tolerance package in R.[19] The K values were determined to be nearly identical to those presented by Gilbert[18] and Owen.[20]

UTL comparisons were summarized by finding the proportion of the 1,000 sampling events for which the maximum replicate from the ISM sample exceeded the background UTL.

3. RESULTS AND DISCUSSION

Graphical summaries are presented for the hypothesis tests (Figs. 2–5) and the 95/95 UTL screening (Figs. 6 and 7) and key observations are noted below. In addition, the relationship between the population distribution (of individual

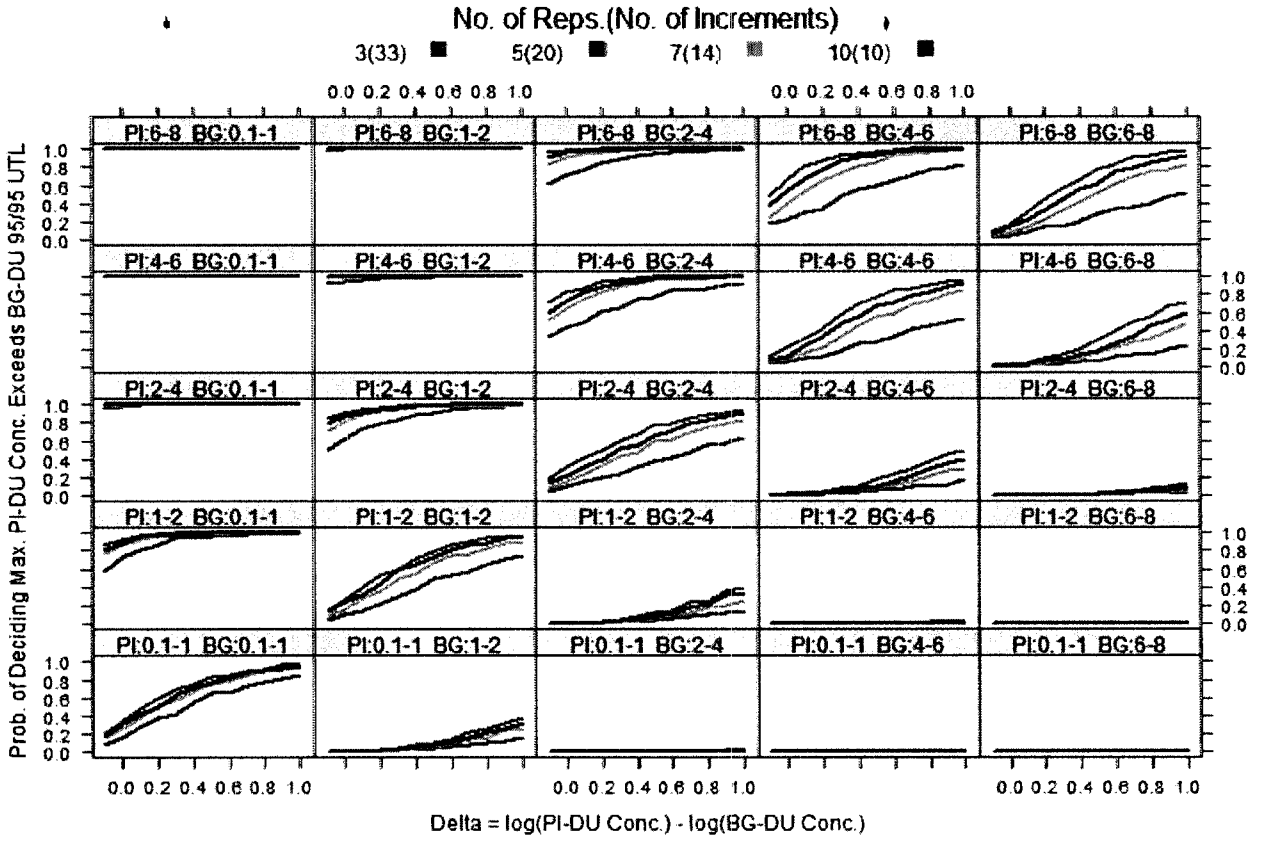

Fig. 6. Probability that the maximum observed ISM replicate concentration from the PI-DU exceeds the 95/95 UTL. Data from both DUs were log-transformed. The heading of each plot shows the population CV for both the PI-DU and BG-DU. Different colored lines represent distinct ISM sampling designs with roughly the same number of total increments sampled. Color graphics are available in the online version of this article.

increments) and the sampling distribution (of replicates) is given for the full range of sample designs (Fig. 8). This information, organized by groups of CVs, helps to explain the differences in the sensitivity of the outcome of a background screening assessment to the choice of statistical test and sample size selection.

3.1. Hypothesis Testing Results

Performance curves for all 25 specified CV combinations (Table I) indicate that a two-sample t-test of hypothesis 1 applied to ISM data performs well, meaning the test achieves the EPA-recommended power of 90% ($\beta = 0.1$) and confidence level of 80% ($\alpha = 0.2$) when both of the following conditions are met (Fig. 2):

(1) CVs for the population distribution (of individual increments) for the PI-DU and BG-DU are approximately equal (represented by graphics in the diagonal of Figs. 2 and 3); and
(2) The true difference in the means (delta) is 0.4 log units, or about 50% on an untransformed scale (Table II).

When delta is greater than 0.4 (and the CVs are approximately equal), the probability of correctly rejecting the null hypothesis and deciding the PI-DU is elevated approaches 100%. For lower delta values, the probability of correctly declaring the mean of the log-transformed concentrations from the PI-DU to be elevated drops considerably below 90%. Given that this is an implementation of hypothesis test 1, this result is an indication that the difference in the sample means (on a log scale) may frequently be too small to be declared statistically significant at $\alpha = 0.2$. Somewhat surprisingly, these findings are relatively insensitive to changes in the number

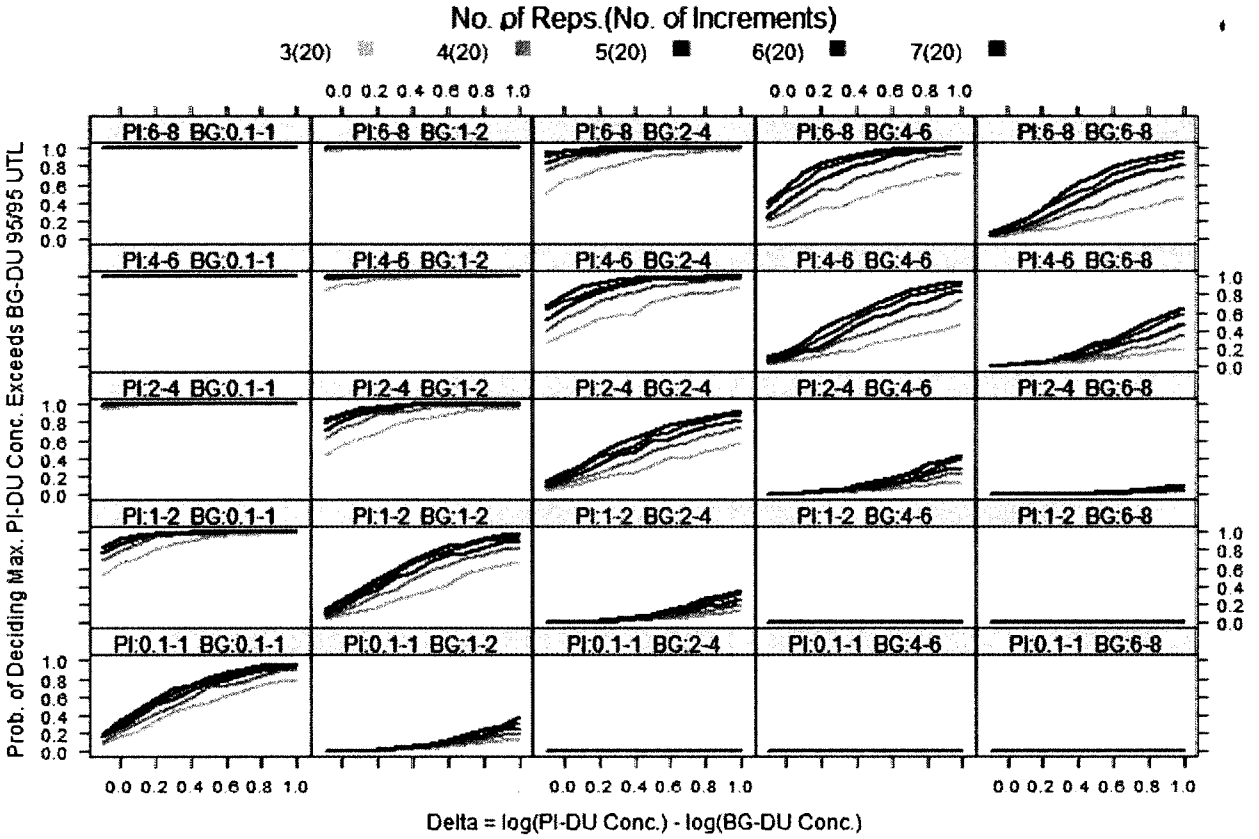

Fig. 7. Same as Fig. 6, except the sample sizes ($r \times n$) illustrate the sensitivity of the test performance to change in replicates ($r = 3$–7) for a fixed number of increments ($n = 20$). Color graphics are available in the online version of this article.

of increments and replicates, as indicated by the overlapping series in each graphic (Fig. 2) and the similar findings displayed in Fig. 3. In addition, the results are insensitive to the choice of tests based on assumptions of equal variance (displayed here) and unequal variance (not displayed). In other words, we cannot rely on the test that "corrects for" unequal variance to improve the decision error rate.

When the CVs of the population distributions are not the same, the probability of correctly rejecting the null and deciding the PI-DU has higher concentrations approaches one of two extremes, regardless of delta or the number of replicates. If the CV of the PI-DU is much greater than that of the BG-DU (e.g., CV = 1–2 at the PI-DU compared with 0.1–1 at the BU-DU), the test is greatly overpowered (the null is rejected with nearly 100% frequency, even when there is no difference between the two means). This is illustrated well by the graphics to the left of the diagonal in Figs. 2 and 3. This is a potentially costly outcome because it may lead to unnecessary remedial action. Conversely, if the CV of the BG-DU is much greater than that of the PI-DU (see graphics to the right of the diagonal), the test is greatly underpowered. From the perspective of health protection, a deficiency in the statistical power of a test is the more concerning outcome.

Figs. 4 and 5 also illustrate the results of hypothesis testing, but under the conditions of the Background Test Form 2 whereby the null hypothesis places the burden on the data to provide evidence that the concentrations in the PI-DU are not elevated above background. Each hypothesis test was implemented under the assumption that $S = 0.1$ (shown here) and $S = 0.2$ (not shown, but generated similar findings). The figures are annotated to show that the test is powered correctly when the colored probability lines cross the intersection made by the horizontal confidence line ($\alpha = 90\%$) and the vertical line indicating the specified value of $S = 0.1$. Similar to simulations with Background Test Form 1, hypothesis testing performs well in terms of intersecting

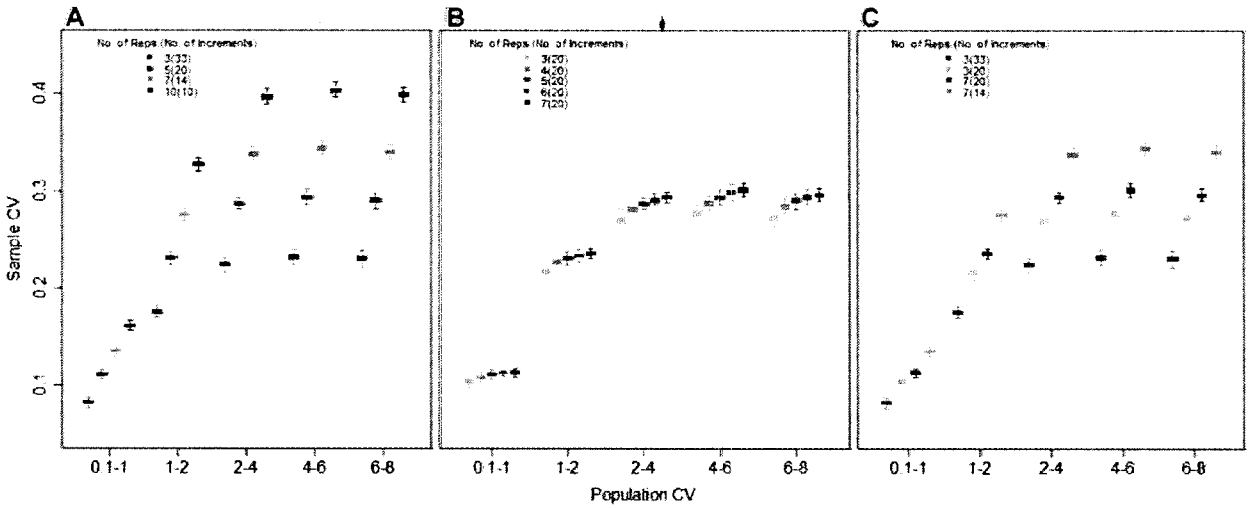

Fig. 8. Summary of sample CV (coefficient of variation) of simulated ISM results obtained from specified ranges of population CVs for lognormal distributions. Each figure illustrates the different performance assessments: (1) varying number of replicates (r) and increments (n), but approximately equal spatial coverage ($r \times n$); (2) constant n with varying r; and (3) constant r with varying n. Results for each design are shown as a boxplot of the distribution of the average sample CV from 1,000 simulated ISM sampling events. Color graphics are available in the online version of this article.

the desired power (80%) when the CVs are approximately the same and the delta is equal to 0.1. The use of Background Test Form 2 provides more power than Background Test Form 1 to detect smaller significant differences, S, between the PI-DU and BG-DU. However, a lower than desired probability of correctly deciding the PI-DU has higher concentrations may still occur under two conditions:

(1) The population CV for the BG-DU is at least double that of the PI-DU; or
(2) The population CVs are approximately equal, but delta is less than S. Since S can be set fairly low, this second observation is less concerning.

In practice, the population CVs and delta are not known, so the findings of the simulation studies can only serve as a guide, rather than hard-and-fast rules. Although the results indicate that error rates are not affected by the range of statistical sample sizes (number of replicates) selected here, the sample sizes are nonetheless a very important element of the sample statistics. Ultimately, the summary statistics of the replicate results, including the ratio of the sample CVs and the percent difference in the sample means, will be the only empirical information available to guide a background screening assessment. Additional considerations regarding the implications of the sampling design, and especially the number of increments, on the sample CVs are discussed below, after reviewing the findings of the UTL calculations.

3.2. Upper Tolerance Limit Screening Results

For most of the scenarios considered, the 95/95 UTL method of screening the maximum ISM replicate from the PI-DU tends to infrequently support a decision that the PI-DU is elevated, even across a wide range of deltas (Figs. 6 and 7). As with hypothesis testing, these comparisons yield lower error rates when the CV from the BG-DU is approximately equal to the CV from the PI-DU. When the BG-DU CV is substantially lower than the PI-DU CV, then these comparisons are overly sensitive and inaccurate, i.e., they will determine that there are exceedances that are not present. Similarly, when the BG-DU CV is substantially higher than the PI-DU CV, the parametric UTL comparisons will not detect exceedances that are present.

3.3. Effect of Sampling Design on CV

The number of increments is one of the design parameters that can affect the standard deviation of the population distribution of ISM replicates. Specifically, the population standard deviation is proportional to the inverse of the square root of the number of increments. For this reason, it is helpful to

examine the effect of the two sampling design parameters, number of replicates and increments ($r \times n$), on the sample CV.

ISM sampling designs with fewer increments tend to have higher sample CVs (Fig. 8A). As expected, the magnitude of the differences is roughly proportional to the ratios of the square root of n. For the same spatial coverage (i.e., approximately 100 total increments), the sample CV is consistently highest with 10×10 ($r \times n$) and lowest for 3×33. The change in sample CV is most noticeable for population CVs (for the distribution of increments) less than 2. Between CVs of 2–8, the effect of changing the number of increments becomes negligible, as demonstrated by the leveling off of the curve created by the boxplots for each design as CV increases in Fig. 8A.

Increasing the number of replicates also has a small but notable effect on the sample CV (Fig. 8B). This contradicts a standard assumption that increasing the number of replicates will generate tighter confidence intervals for parameter estimates, but will not otherwise alter the "fixed" population parameters (mean and variance). Unlike the influence of changing increments, which plateaus at population CV = 2, the small effect of the replicate sample size was consistent across the full range of population CVs.

Fig. 8C presents a hybrid of the factors illustrated in Figs. 8A and 8B. The sample CV is more sensitive to changes in number of increments than number of replicates.

These findings have interesting implications for background screening assessments in situations where the ISM data are derived from different sampling designs. As discussed in Section 1, differences in sampling designs translate into differences in the population distributions from which samples are collected. Consider the realistic scenario in which the population distributions are unknown for the PI-DU and BG-DU, and a 3×30 ISM design is applied to the PI-DU and a 3×20 ISM design is applied to the BG-DU. We know before even running a hypothesis test that the smaller number of increments will inflate the population standard deviation of the BG-DU relative to the PI-DU, if the population means are the same. We might consider the test to be slightly underpowered in this case–-likely to (incorrectly) fail to reject the null hypothesis more often than the type II error rate (β) would have indicated. The extent to which this design difference introduces a meaningful bias depends on how robust the test performance is to unequal variances.

One strategy for addressing differences in design parameters might be to introduce "correction factors" that would adjust the sample statistics—the mean, standard deviation, or both—so that the data sets are more comparable. The purpose of the correction factor would be to align the population distributions if, in fact, the distributions from the two areas are the same. This strategy sounds simple enough, but should be approached with caution. In practice, we do not actually know what the population distributions are to verify that the correction factor produces the desired error rates. Should the data set with smaller sample size be adjusted "up," the higher sample size adjusted "down," or should both be adjusted to some intermediate sample size? It is unclear how differences in the shape of the distribution may factor into the relationship between the population mean and standard deviation—for nonnormal distributions, these parameters are proportional. It is possible that the effect that changing the number of increments has on the corresponding sample statistics may differ depending on the shape of the distribution prior to the adjustment. The findings from this study (Fig. 8C) suggest that the solution may indeed be more complicated than simply adjusting the standard deviation by the ratio of the square root of the n because both the population mean and standard deviation may be impacted. Ultimately, a correction factor scheme needs to be accompanied by evidence that it generally improves the performance of the statistical tests. This research topic can be more fully explored in subsequent simulation studies.

4. CONCLUSIONS AND RECOMMENDATIONS

This simulation study provides a foundation for understanding whether and how ISM sampling can be used for background screening assessments. In this simulation study, for purposes of illustration, we defined background conditions based on a hypothetical single DU, although alternative sampling designs could involve multiple background DUs. Despite the small sample sizes available for statistical analysis (i.e., the number of replicate results), ISM is amenable to hypothesis testing under a wide range of site conditions. This study supports the following procedures to achieve the most reliable outcomes from background screening with ISM data from single DUs:

(1) Use hypothesis tests rather than UTL screening because the UTL method has a conceptual limitation in the context of comparing distributions of means, and the sample sizes generally do not afford sufficient power to detect differences that exist.
(2) Apply the standard two-sample *t*-test to the log-transformed replicate results, acknowledging that the test provides insights on differences in the medians (or geometric means if the distributions are in fact lognormal) rather than the means.
(3) Run the standard two-sample *t*-test under the assumption of equal variance for site and background DUs, rather than analogous tests under conditions of unequal variance. The alternative tests will not be able to compensate for the drop in power that can occur when variances differ substantially. Likewise, do not run nonparametric hypothesis tests, which lack sufficient power for most ISM sampling designs.
(4) Calculate the sample CVs, ratios of the sample means, and note whether the sampling designs (number of increments and replicates) and general conditions of the DU are similar. If the sample CV for the BG-DU is greater than that of the PI-DU by more than a factor of two, acknowledge that the test may lack power (and yield false negative results).

These analyses demonstrate that there is minimal advantage to increasing the number of increments unless one is doing UTL comparisons. In contrast, increasing the number of increments lowered the CV, which increases the precision of subsequent testing and conclusions. More importantly, our results indicate that reasonably accurate hypothesis test results can be achieved even when the CVs of both the PI-DU and BG-DU are relatively large as long as the CVs of the two DUs being compared are approximately equal. From a sampling design perspective, this can be achieved by carrying out small pilot studies of potential BG-DUs and the PI-DU to compare variability. Although the variability of the PI-DU is beyond the control of those carrying out the sampling, the variability of the BG-DU can potentially be controlled by controlling the size of this DU and expanding or limiting it to areas with a level of variability that is similar to that of the PI-DU. This strategy of reexamining the BG-DU area is consistent with USEPA guidance on background screening for soil.[1]

As a general rule, hypothesis testing is more reliable than UTL comparisons and in cases where hypothesis testing will not provide accurate results (i.e., when there is a substantial disparity in variability between the BG-DU and PI-DU), we recommend graphical comparisons of the data using side-by-side dot plots (see ITRC[10] for examples).

This simulation study was intended to capture conditions applicable to most sites. However, findings from this study may not apply to certain sites with conditions that are more extreme at site or reference areas, or sampling designs that fall outside the conditions simulated in this study. Also, although the sampling designs and conditions examined in this study explore a variety of questions concerning the performance of ISM for background screening assessment, there are a number of questions yet to be explored. This analysis does not directly address questions regarding options when a historic background data set is based on discrete sampling—Is it possible to "convert" such a sample to an *equivalent* ISM result, even when population parameters are unknown? Also, applications of background screening are of interest for large sites with numerous DUs, some of which may not have sufficient sample sizes to provide reliable estimates of the variance. Are there decision rules that can be established for "threshold" variance estimates that would allow for hypothesis test outcomes to be determined? Questions regarding extrapolation uncertainty, while not unique to ISM, can be readily addressed through numerical simulation, informed by empirical data from site investigations. Practical answers and guidance to these questions may be developed in subsequent simulation studies.

REFERENCES

1. U.S. EPA. Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites. EPA/540/R-01/003. OSWER 9285.7-41. Washington, DC: Office of Emergency and Remedial Response, 2002.
2. Navy. Navy Policy on Use of Background Chemical Levels. Washington, DC: Office of the Chief of Naval Operations, 2004.
3. U.S. EPA. The Role of Screening-Level Risk Assessments and Refining Contaminants of Concern in Baseline Ecological Risk Assessments. Washington, DC: Office of Solid Waste and Emergency Response, 2001.
4. Navy (Department of the Navy). Guidance for Environmental Background Analysis, Volume I: Soil. UG-2049-ENV. Washington, DC: Naval Facilities Engineering Command, 2002.
5. PNNL (Pacific Northwest National Laboratory). Visual Sample Plan—Version 7.0 User's Guide. Prepared for the U.S.

Department of Energy. PNNL-23211. Richland, WA, 2014 March. Available at: http://vsp.pnnl.gov/documentation.stm, Accessed May 30, 2015.
6. U.S. EPA. Performance of Statistical Tests for Site Versus Background Soil Comparisons When Distributional Assumptions are Not Met. Technology Support Center Issue by E.J. Englund. EPA/600/R-07/020. Las Vegas, NV: Office of Research and Development, 2007.
7. U.S. EPA. Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities—Unified Guidance. EPA 530/R-09/0007. Washington, DC: Office of Resource Conservation and Recovery, 2009.
8. U.S. EPA. ProUCL Version 5.0.00 Technical Guide. Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations. EPA/600/R-07/041. Washington, DC: Office of Research and Development, 2013.
9. U.S. EPA. ProUCL Version 5.0.00 User Guide. Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations. EPA/600/R-07/041. Washington, DC: Office of Research and Development, 2013.
10. ITRC (Interstate Technology & Regulatory Council). Incremental Sampling Methodology. ISM-1. Washington, DC: Interstate Technology & Regulatory Council, Incremental Sampling Methodology Team, 2012. Available at: www.itrcweb.org. Accessed October 14, 2012.
11. Pitard F. Pierre Gy's Sampling Theory and Sampling Practice: Heterogeneity, Sampling Correctness, and Statistical Process Control, 2nd ed. Boca Raton, FL: CRC Press, 1993.
12. U.S. EPA. Role of Background in the CERCLA Cleanup Program. OSWER 9285.6-07P. Washington, DC: Office of Emergency and Remedial Response, 2002.
13. U.S. EPA. Guidance on Systematic Planning Using the Data Quality Objectives Process. EPA/240/B-06/001. Washington, DC: Office of Environmental Information, 2006.
14. R Core Team. R: A Language and Environment for Statistical Computing. Austria, Vienna: R Foundation for Statistical Computing, 2016.
15. Warnes GR, Bolker B, Gorjanc G, Brothendieck G, Korosec A, Lumley T, MacQueen D, Magnusson A, Rogers J. gdata: Various R Programming Tools for Data Manipulation. R package version 2.17.0, 2015.
16. Conover WJ. Practical Nonparametric Statistics, 3rd ed. New York: John Wiley & Sons, 1999.
17. Siegel S, Castellan NJ. Nonparametric Statistics for the Behavioral Sciences, 2nd ed. New York: McGraw-Hill, 1988.
18. Gilbert RO. Statistical Methods for Environmental Pollution Monitoring. New York: Van Nostrand Reinhold, 1987.
19. Young DS. Tolerance: An R package for estimating tolerance intervals. J Statistical Software, 2010; 36(5):1–39.
20. Owen DB. Handbook of Statistical Tables. Palo-Alto, CA: Addison-Wesley, 1962.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's website:

The R code used to execute and summarize the simulations is included as an online supplementary file to promote an open exchange of this and future simulation studies among the community of scientists and investigators who are studying ISM performance.