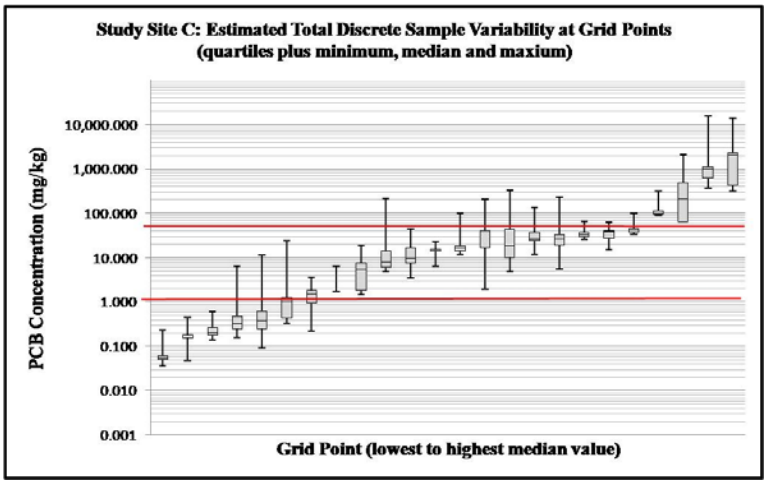
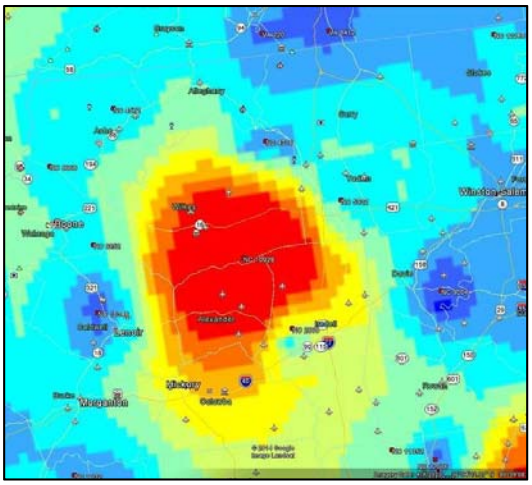
Small-Scale Variability of Discrete Soil Sample Data

Part 2: Causes and Implications for Use in Environmental Investigations





[1]Roger Brewer, [1]John Peard, [1]Jordan Nakayama and [2]Marvin Heskett

Hawai´i Department of Health

[1]Hazard Evaluation and Emergency Response
[2]Element Environmental

June 2015

Forward

This report presents Part 2 of a two-part field-based study of the variability of contaminant concentrations within and between co-located discrete soil samples. Part 1 of the report summarizes the results of the field study and briefly discusses the implications of the study findings on the use of discrete sample data for decision making in environmental investigations. Part 2 of the study evaluates the causes of discrete sample variability and uncertainty and discusses implications for site characterization, risk assessment and remedial design based on discrete sample data in more detail.

This report will be updated and amended as needed in the future. Comments and suggestions are welcome and should be provided to Roger Brewer at roger.brewer@doh.hawaii.gov. Data tables presented in this report will be made available in Excel format along with the report on the Hawai´i Department of Health (HDOH), Hazard Evaluation and Emergency Response (HEER) web page or are available upon request from the above contact.

> *"Perhaps the sentiments contained in the following pages, are not yet sufficiently fashionable to procure them general Favor; a long habit of not thinking a Thing wrong gives it a superficial appearance of being right, and raises at first a formidable outcry in defense of Custom. But the Tumult soon subsides. Time makes more Converts than Reason.*"
>
> Thomas Paine, 1776 (*Common Sense*, on succession and independence of the new United States from Great Britain)

Executive Summary

Part 1 of this study presents the results of a field-based investigation of the variability of contaminant concentrations both within a single discrete sample and between co-located samples around a single grid point. The data indicate that random, small-scale variability unrelated to larger-scale trends normally the target of site investigations is ubiquitous and can lead to significant error in decision making if left unrecognized. The magnitude of potential error, e.g., missing large areas of heavily contaminated soil, increases as the magnitude of small-scale variability around individual discrete sample points increases.

Part 2 of this study briefly reviews the origins of random, small-scale variability of contaminant concentrations in soil. This is followed by a more in-depth review of the implications of this variability with respect to reliance on discrete soil sample data to investigate, assess risk and remediate contaminated sites. Topics discussed include:

- Nature and meaning of "hot spots;"
- Comparison to risk-based screening levels;
- Estimation of the extent of contaminated soil;
- Preparation of isoconcentration maps; and
- Estimation of mean contaminant concentrations.

Additional topics broached in the report include the adequacy of current laboratory protocols for processing and subsampling of discrete soil samples prior to analysis and estimation of contaminant mass for *in situ* remedial actions.

Variability in contaminant concentrations reported for analysis of multiple subsamples from a single discrete sample or for co-located samples around a single grid point primarily arises from small-scale, distributional heterogeneity of contaminants in soils; i.e., heterogeneity at the mass of an individual sample (e.g., 100 to 200 grams) or the mass of soil actually analyzed by the laboratory (e.g., 1 to 30 grams). Variability associated with laboratory analytical error is minimal in comparison to that caused by distributional heterogeneity (see Pitard 1993, 2005, 2009; Minnitt et al 2007; ITRC 2012).

Random variability around a grid point at the scale of a typical, discrete soil sample negates the reliability of comparing data for individual sampling points to risk-based screening levels. The magnitude of potential error is tied in part to the nature of the original release. Small-scale variability can be reasonably low (e.g., +/- 100%) for intentionally applied chemicals (e.g., pesticides) or impacts to fine-grained sediment due to wastewater streams from controlled, industrial processes. The magnitude of random variability increases with increasing chance of small, isolated pockets of contamination or "nuggets" of concentrated material in the soil.

Discrete sampling approaches typically lack adequate mass to capture the inherent, small-scale, random variability of the target contaminated soil and are therefore unreliable for decision

making. High discrete sample variability is most obvious for contaminants such as lead shot and lead paint in soil, but can also be the case for releases of waste oils to soil or for sites that have been driven over or otherwise disturbed since the initial release. Investigations using discrete sampling at these sites lead to a high instance of both "false negatives" (i.e., samples collected from random small clean spots within the overall contaminated area) as well as "false positives" (i.e., samples collected from random, small contaminated spots within the overall contaminated area).

Isoconcentration maps based on discrete soil sample data give a false sense of resolution at the scale of an individual, discrete sample point. Identification of large-scale patterns of soil contamination based on grids of discrete samples is possible if small-scale variability within any given area is reasonably low. False patterns of contaminant concentrations are generally unavoidable, however, due to random variability of contaminant concentrations above and below contour intervals at any given grid point. Such areas are easily identifiable on isoconcentration maps as seemingly isolated "hot spots" and "cold spots" that change locations when co-located or "replicate" samples are collected and tested from the same grid points or from a different set of grid points. Note that the term "replicate" as used in earlier USEPA guidance (e.g., USEPA 1987, 1991) is in a strict sense not applicable, since a single discrete sample cannot be replicated. These problems worsen as the magnitude of random, small-scale, variability increases, with boundaries between truly "clean" and "contaminated" soil becoming even more blurred.

Processing of discrete samples in accordance with incremental sampling methods reduces error associated with intra-sample variability (e.g., air dry, sieve and subsample). Small-scale variability between closely spaced, discrete samples can still be significant, however, and better processing and testing of samples will not solve the field variability issues noted above.

Use of geostatistical methods to estimate mean contaminant concentrations for a targeted area and volume of soil based on a single set of discrete samples can also give a false sense of precision. Statistical evaluation of a single data set only assesses the precision of the estimated mean *in terms of the data set provided and the statistical method employed*. The representativeness of the set of samples in terms of is unknown and unmeasurable in absence of the collection of completely independent, replicate sets of discrete samples. Replicate sets of discrete samples are rarely if ever collected. Even if replicate sets of samples were available, the lack of systematic control of bias in the collection and processing of samples would still introduces considerable uncertainty in the data. Comparison of random sets of discrete soil sample data for the study sites highlights the potential for an unrevealed, lack of representativeness for any given set of samples. In contrast, the collection of replicates to assess field representativeness is a required part of incremental sampling methodologies. This provides a much more effective method to evaluate the overall precision of mean contaminant estimations for risk assessment and other purposes.

Conclusions drawn from this study are quite clear – the reliability of traditional, discrete samples to identify, characterize and remediate contaminated soil is at best uncertain and in many if not most cases quite low. The use of discrete soil samples in environmental soil investigations is founded on an untested and ultimately erroneous assumption in the late 1980s that contaminant concentrations in soil should be generally consistent within and between spill areas, regardless of the mass of soil tested. This was largely based on experience from testing of waste generated from well-controlled industrial processes.

Accepting these assumptions as fact allowed the mass of soil collected for environmental contaminant investigations to be determined by the laboratories and negated the need for the field investigator to demonstrate the reliability and representativeness of the samples submitted. Assuming that small-scale variability was negligible likewise negated the need for serious processing of soil samples, which unlike some liquid wastes or food can be costly and time consuming. This led to an ongoing push to test smaller and smaller subsamples of soil in order to expedite analysis and minimize costs at the laboratory. When tested, discrepancies between co-located field samples or replicate subsamples in the laboratory were often assumed, at least by those who submitted the samples, to reflect laboratory error rather than error in the manner in which the sample had been collected in the field or how the subsample had been selected for testing at the laboratory. (If replicate data were provided by the laboratory, then the highest-reported concentration would be used for decision making.) Precision in terms of data quality and reliability only needed to be evaluated in terms of the laboratory "QA/QC" validation for analytical error, without serious consideration of either field or lab sub-sampling error. In short, if the laboratory equipment was operating properly, then the lab reporting data was typically considered "valid" for decision making by those who submitted the samples for contaminant investigations.

The unreliability of this approach for particulate matter such as soil was recognized decades ago in the mining and agricultural industries. An entirely different method of testing referred to as "incremental sampling" quickly developed once the cause of the errors was realized. The inclusion of these methods in environmental investigations has only recently begun to move forward, as evidence against discrete soil sampling approaches mounted in the field.

The discussions provided in this report document the shortcomings of discrete soil sampling methodologies in use since the 1980s and introduce ideas for further research. The review of the field data collected is not comprehensive. It is hoped that other researchers with a strong understanding of sampling theory and environmental investigations will carry out more detailed reviews of the data to further improve methods for the investigation of contaminated soil, as well as other media. More up-to-date soil sampling methodologies based on our past thirty years of experience are progressively being incorporated into state and federal guidance documents. At the writing of this report, over 3,000 people have joined a two-day webinar on Incremental Sampling Methodologies developed by a one hundred-strong team of regulators, consultants and private entities hosted by the Interstate Technology and Resource Council (refer to ITRC 2012).

A broader-scale understanding of the limitations of 1980s-era, discrete sampling approaches and a demand for higher quality data is still needed from the regulatory agencies, however. The economic and scientific demand to do so will necessarily have to come from those most directly affected. It is hoped that the field study of discrete sample error presented in this report will aid in these discussions and further development of the environmental industry.

TABLE OF CONTENTS

Tables

Figures

1 INTRODUCTION AND STUDY OBJECTIVES

The objective of this study is to address a deceptively simple question – What is the variability of contaminant concentrations in soil around a fixed point *at the scale of a typical, discrete soil sample*? The latter part of the question is important. Decisions regarding the need to address the presence of a potentially toxic chemical in soil are routinely made on a point-by-point basis for investigations based on the collection of discrete soil samples. Such approaches are enshrined in guidance published by the US Environmental agency and other regulatory agencies. With only a few exceptions, however, the reliability of this approach has not been tested in detail by state and federal regulatory agencies who oversee environmental investigations.

Testing of soil for contaminants rose exponentially in the 1980s following passage of federal, environmental legislation such as the Resource Conservation and Recovery act and publishing of associated regulations and guidance. With little soil experience to go by, the authors of the guidance recommended and even required methods for sampling and testing of soil adopted from approaches already in place for water and other liquid wastes. Testing of small subsamples from a relatively small number of samples is common practice and reasonably accurate for these types of media. Unlike a liquid, however, the concentration of a contaminant in soil can vary dramatically both within single, discrete samples and between closely-spaced, co-located samples. This variability, described in terms of *distributional heterogeneity*, requires that greater attention be paid to specific questions being asked in soil investigations as well as the manner in which samples are collected, processed and analyzed.

For example, a single concentration would be reported for a chemical in soil if the entire mass of soil from an area targeted for investigation could be collected, extracted and analyzed as a single sample. The value reported represents the *true mean* concentration of the contaminant for the volume of soil as a whole. The concentration can be expected to vary above and below the overall mean in smaller, subsets of the targeted volume of soil. Variable concentrations would be expected, for example, if a targeted exposure area was divided into four subareas and the entire mass of soil from each area again collected and independently tested.

As demonstrated in Part 1 of this study, this same type of distributional heterogeneity extends down to the scale of an individual, discrete sample, typically a few hundred grams in mass. The concentration of PCBs in separate, laboratory subsamples of single, discrete samples collected less than 0.5 meters apart from Study Site C varied by up to two orders of magnitude (refer to Table 4-17 in Part 1). Variability of reported contaminant concentrations in the targeted volume of soil can be expected to *increase* as the scale of measurement *decreases (e.g. sample mass collected in field decreases and/or mass sub-sampled by the lab decreases).* At the scale of individual soil particles or even one area of a particle, the *maximum* concentration of a contaminant in soil, if present, will necessarily at some point be either 0 mg/kg (0%) or 1,000,000 mg/kg (100%) as single, pure particles of soil or pure particles of the contaminant.

The question “What is the maximum concentration of the contaminant in soil?” at this scale is very straight forward – either 0 mg/kg if completely absent or 1,000,000 mg/kg if present.

Importantly, and as documented in Part 1 of this study, such small-scale variability is most likely *random* at the scale of a discrete soil sample and not related to large-scale trends of interest. The fact that the average concentration of PCBs is higher on one side of a discrete soil sample than another cannot of course not be assumed to indicate an increasing, large-scale trend in concentrations in that direction. The same is true for small-scale variability between co-located discrete samples around an individual sampling point (e.g., see Figure 5-2 in Part 1). At some scale the *mean* contaminant concentration in soil will indeed begin to reflect large-scale trends of interest. This represents the point at which the area and volume of soil represented by the sample is sufficient to capture and overcome smaller-scale, random variability. The scale in the field at which this happens is necessarily site-specific, and is an integral part of the Decision Unit designation process (refer to HDOH 2008).

Concern regarding error associated with the use of traditional, discrete soil sampling methods has been growing for some time (e.g., Hadley and Sedman 1992, Pitard 1993, Ramsey and Hewitt 2005, Hadley and Petrisor 2014). As stated by Hadley and Petrisor (2014):

> It has been clear for some time that the major sources of error in soil sampling for chemical contamination come not from laboratories but from field sampling and subsampling. This situation is—and should be—of concern to environmental forensic scientists. Legal arguments and determinations are based on the prevailing standards of science and practice and often rely on relevant requirements, policies, and guidance from regulatory agencies. Perhaps as a result of deferring to regulatory agencies many of these legal proceedings have focused primarily on the potential for laboratory error rather than on the potential for sampling error.

Close detail and oversight has been paid to the precision of the analytical methods and equipment used by laboratories for testing of environmental soil samples. Whether the resulting data were in fact representative of the sample submitted, and the sample representative of the area from which it was collected in the field has largely been overlooked, however. This is quite surprising, considering the large sums of money spent on environmental investigations over the past three or more decades and the legal liabilities implied in declaring a site either “contaminated” or “clean.”

As discussed in this report, these circumstances arose primarily due to the flood of environmental investigations required under newly passed environmental regulations in the 1980s and early 1990s and the sudden need to investigate potential soil contamination at hundreds of thousands of sites. Extracts from these documents are presented in this report to help understand the rationale behind recommendations for soil sampling ultimately presented. The most expeditious and seemingly appropriate solution at the time was to simply apply

existing methods for sampling and testing of liquid wastes to soil. Waste streams from industrial processes were known to be fairly consistent for any given period of time. It seemed reasonable that concentrations of contaminants in soil upon which the wastes were released would also be relatively uniform, at least over short distances. This limited the number of samples required to characterize an impacted area and allowed interpolation of contaminant concentrations in soil between points that had not been directly tested.

An assumption of uniformity also permitted the use of data for very small masses of soil to be considered representative of very large areas and volumes of soil in the field. The mass of material required to generate presumably representative data was assumed to be equal to or less than the minimum mass required for testing by the laboratory. Test protocols for metals, for example, only require that a gram of material be tested. A maximum of 10 to 30 grams is only required for most other chemical analyses.

In the field, only a relatively small number of samples were likewise assumed to be necessary to establish clean boundaries around large-scale contamination patterns of primary interest. As was the case for "short-range," temporal variability of contaminant concentrations in waste streams, smaller scale, spatial variability of contaminant concentrations in soil was assumed to be negligible. This also allowed for contaminant concentrations in soil that had not been directly tested to be interpolated based on nearby data points.

Guidance called for the reproducibility of both field and laboratory data to be evaluated by testing of replicates at the rate of one test per ten to twenty samples (e.g., USEPA 1987, 1991). Differences in replicate data for the same sample were assumed to be associated with laboratory analytical error. The opposite is likely to be true, with the error primarily associated with sample collection and representativeness and secondarily as a result of lab sub-sampling error (see Pitard 1993, 2005, 2009; Minnitt et al 2007; ITRC 2012). Laboratory protocols were modified to recommend non-specified "homogenization" of soil samples prior to the collection of subsamples for testing. When conflicts in replicate data still arose, the higher concentration was normally selected for decision making. The time and cost of retesting the 95% of samples where replicate subsamples were not collected and tested (i.e., 19 of 20), and lack of push by regulators to do so, negated serious attention to this matter outside of some complaints about "laboratory error."

An important difference between industrial waste and contaminants in soil, unknown or overlooked by the authors of the guidance at the time, was the striking heterogeneity of concentrations in the latter at the scale of the discrete samples being collected and the mass of soil being tested. The latter issue did not go unrecognized by most laboratories, but there was little motivation to change in the absence of regulatory requirements for more reproducible methods of sample processing and subsampling for analysis.

The field investigation of discrete sample variability presented in Part 1 of this study raises significant concerns about the validity of these early assumptions. As discussed by multiple experts in sampling theory, field sampling error and laboratory subsampling error, rather than laboratory analytical error, is the likely culprit in poor data quality and resulting failed investigations and remedial actions. Part 2 of this study looks more closely at the causes of random, small-scale variability of contaminant concentrations in soil and implications for the use of discrete sample data in environmental investigations. Topics discussed include:

- Nature and meaning of "hot spots";
- Comparison of individual, discrete sample data to risk-based screening levels;
- Estimation of the extent of contaminated soil potentially impacted above screening levels;
- Reliability of isoconcentration maps based on discrete sample data; and
- Reliability of mean contaminant concentrations for targeted areas based on geostatistical evaluation of a discrete sample data set.

Additional topics discussed include the adequacy of current laboratory protocols for processing and subsampling of soil samples for analysis and estimation of contaminant mass for *in situ* remedial actions.

Section 2 provides a brief overview of discrete sample variability in terms of sampling theory and contaminant heterogeneity in soil. Section 3 explores the concept of "hot spots" and the importance of scale in environmental investigations and decision making. Section 4 reviews the appropriateness of point-by-point comparisons of discrete soil sample data to risk-based screening levels. Section 5 evaluates the reliability of discrete sample data to estimate the extent of soil contamination above levels that could pose a risk to human health and the environment. Section 6 expands on this topic and explores the effect of random, small-scale heterogeneity on the representativeness of computer-generated isoconcentration maps that rely on discrete sample data.

Section 7 reviews the concepts of "bias" and "precision" in sampling and the use of geostatistical methods to estimate the mean concentration of a chemical in soil. This includes the expected reproducibility of estimated, mean contaminant concentrations as well as treatment of "outlier" data that can confound geostatistical analysis of discrete sample data sets. Section 8 reviews other problems caused by random, small-scale variability of contaminant concentrations in soil that if left unrecognized can cause significant problems in environmental investigations. Section 9 summarizes the results of the study and reviews the advantages of Decision Unit and Multi Increment® site investigation approaches to address problems with discrete soil sample data and generate more reproducible and technically defensible data for decision making. (Multi Increment® is a registered trademark of EnviroStat, Inc.)

2 HETEROGENEITY OF CONTAMINANTS IN SOIL

The variability of contaminant concentrations observed between and within discrete, soil samples is primarily related to three factors: 1) Large-scale differences in the amount of the contaminant released to different parts of the site, 2) Random, small-scale heterogeneity of contaminant distribution in soil at the scale of the sample collected (e.g., a few hundred grams) and 3) Similar, random contaminant distribution within a sample at the scale of the mass of soil analyzed by the laboratory (typically 1 to 30 grams). *Distributional heterogeneit*y is not scale dependent and extends down to variability between individual particles.

Differences in contaminant concentrations reported for samples collected around individual grid points in this study are discussed in terms of "intra-" and "inter-" sample variability (see Part 1). "Intra-sample" variability is applied to data for separate subsamples of soil collected and tested from an individual sample. The term "inter-sample" variability is applied to variability of contaminant concentrations reported for closely spaced samples collected in the immediate vicinity of a grid point (e.g., three foot perimeter used in this study).

2.1 HOT SPOTS VERSUS HOT AREAS

The term "small-scale" variability is used in this report to collectively describe the sum of intra- and inter-sample variability around a single grid point. "Large-scale" variability is used in the report to describe variability between different areas of a study site. Large-scale variability is related to the release of greater amounts of a contaminant in one area. The identification of such areas is the primary objective of most site investigations. Small-scale variability refers to often random changes in contaminant concentrations within larger-scale trends of interest. Although defined in very general terms, these concepts should be recognizable to people involved in environmental field investigations. The term "hot spot" has in the past been confusingly applied to all scales of contaminant concentration variability. The term "hot area" is more appropriate from a risk and environmental investigation perspective. A more detailed review of this issue is provided in Section 3.

The nature and implications of random (i.e., non-reproducible), small-scale variability of contaminant concentrations in soil has only recently begun to be evaluated as part of environmental investigations. The root cause of this variability is quite simple – the mass of soil traditionally collected and tested as a discrete sample is in most cases too small to overcome random, distributional heterogeneity of contaminants in soil. This is the essence of "Fundamental Error" as described in sampling theory (Pitard 1993, 2009; see also USEPA 1999; Ramsey and Hewitt 2005; Minnitt et al 2007). Specific equations have been developed by the mining industry to calculate the mass of soil that must be collected and tested in order to overcome and capture random, scale variability and generate a representative sample. These equations are based on assumptions regarding the size, shape and distribution of particles in the soil matrix, the approximate concentration of the contaminant, and multiple other factors. Proper

collection methods to ensure that samples are of equal size, shape and mass and proper processing for the collection of subsamples are also critical.

The potential that discrete soil samples were too small to overcome random variability of contaminant concentrations in soil was not unknown to authors of early USEPA guidance documents. The USEPA guidance document *A Rationale for the Assessment of Errors in the Sampling of Soils* when discussed the need for "representative sampling" (USEPA 1990b):

> Soils are extremely complex and variable which necessitates a multitude of sampling methods… *A soil sample must satisfy the following*: 1) Provide an adequate amount of soil to meet analytical requirements and *be of sufficiently large volume as to keep short range variability reasonably small… The concentrations measured in an heterogeneous medium such as soil are related to the volume of soil sampled* and the orientation of the sample within the volume of earth that is being studied. The term 'support' is used to describe this concept.

The same document warned that errors in the collection and representativeness of soil samples were likely to far outweigh errors in analysis of the samples at the laboratory (USEPA 1990b):

> During the measurement process, *random errors* will be induced from: sampling; handling, transportation and preparation of the samples for shipment to the laboratory; taking a subsample from the field sample and preparing the subsample for analysis at the laboratory, and analysis of the sample at the laboratory (including data handling errors)… *Typically, errors in the taking of field samples are much greater than preparation, handling, analytical, and data analysis errors*; yet, most of the resources in sampling studies have been devoted to assessing and mitigating laboratory errors.

Addressing errors in the laboratory was and has continued to be "low hanging fruit" that received the greatest focus of attenuation over the past 20 to 30 years (USEPA 1990b):

> It may be that those errors have traditionally been the easiest to identify, assess and control. This document adopts the approaches used in the laboratory, e.g. the use of duplicate, split, spiked, evaluation and calibration samples, to identify, assess and control the errors in the sampling of soils.

The implications of these important ideas in the field were, unfortunately, never fully discussed in guidance documents nor followed up in subsequent guidance documents. Ultimately, confusion over the need to determine the "maximum" contaminant concentration within a targeted area and search for sample-size "hot spots" continued (and still continues) to plague the industry, and reliance on often scant discrete soil sample data for decision making quickly became routine.

2.2 CONCENTRATION VERSUS MASS

The reported concentration of a contaminant in soil is always an average and that the average can vary with respect to the mass of the soil represented by the subsample actually tested by the laboratory (see Pitard 1993; ITRC 2012). Consider for example the jar of colored gumballs depicted in Figure 2-1. Assume each gumball reflects the approximate mass of material that the laboratory will test for the chemical analysis. Assume also that the objective of testing the sample is to determine the *average* color of the gumballs in the jar, similar to determining the average concentration of a contaminant in a soil sample submitted to a laboratory for analysis. Selecting and testing individual, randomly selected gumballs from the sample jar in Figure 1 would result in highly variable "concentrations" of color. Collecting and testing groups of two or three, randomly selected gumballs would yield less dramatic but still variable color concentrations. The number or more specifically mass of gumballs that must be collected to represent the sample is, in this case, dependent only on the size and mass of each "particle" and the total variability of colors within the sample.

In the case of soil, the bulk sample must be processed in a manner that allows a smaller subsample to be reasonably representative of the original sample as a whole for testing. Equal access and probability of selection must be provided to all gumballs (particles) within the sample during subsampling. This typically involves drying and sieving the bulk sample to a known maximum particle size range and then collecting a lab analysis mass adequate to reduce subsampling fundamental error to a reasonable level (e.g., <15%). These approaches are incorporated into, incremental sampling methodologies (refer to Section 4 of the HEER office Technical Guidance Manual; HDOH 2008; see also ITRC 2012). In some cases it may be necessary to grind the sample in order to reduce particle size and associated Fundamental error to an acceptable level. If the example in Figure 2-1 were in fact a soil sample and the lab could test no more than a single, gumball mass of material at a time, then significant reduction in "particle size" would indeed be required in order to obtain a subsample that was representative of the original sample as a whole (ignoring the likelihood that the interior of the gumballs is probably not the same as the outside color).

Random variability of contaminant concentrations within a single soil sample is inherent and unavoidable, as demonstrated in Part 1 of this report. As also demonstrated in the field study, variability of contaminant concentrations in "co-located" discrete samples can also be significant and again random. This is to be expected. Contaminants are rarely released to soil in a systematic manner. Even when they are, as is assumed to be the case for arsenic in soil at Study Site A, chemical changes (e.g., preferential binding of arsenic to small aggregates of iron hydroxides), difference in soil types, subsequent disturbance of the soil and other factors can lead to significant and random small-scale variability of concentrations in soil. Large-scale trends can certainly exist, but on close inspection trends between nearby, small masses of soil are likely to be random and unpredictable (e.g., see surface spill of milk in Figure 5-9 of Part 1). Like a single gumball in the above example, discrete soil samples will in most cases simply be too small

to overcome and capture this small-scale variability and allow larger-scale trends of contaminant concentrations to be accurately established.

2.3 Controlling Bias and Improving Precision

Although Fundamental Error can never be completely eliminated, its effect can be minimized by careful sampling design and ensuring that samples are collected, processed and tested in an unbiased manner (e.g., collection of adequate sample mass from an adequate number of locations). Error associated with random, distributional variability of a contaminant within a sample ("intra-sample variability") can in theory be completely eliminated by use of proper field collection, processing and lab subsampling techniques (Pitard 1993, 2005, 2009; Minnitt et al 2007; see also HDOH 2008; ITRC 2012). Error associated with random, distributional heterogeneity between closely spaced, discrete soil samples cannot be eliminated, since this is an inherent property of the soil under investigation (Pitard 1993, 2005, 2009; Minnitt et al 2007). This error can be minimized through the use of decision unit and incremental sample investigation approaches, as described in the HEER office Technical Guidance Manual and references included in that document (HDOH 2008).

An overview of bias and precision in soil investigations is included in Section 7. A more detailed discussion of these topics can be found in the HDOH Technical Guidance Manual (HDOH 2008) and the Interstate Technology Regulatory Council (ITRC) document *Incremental Sampling Methodology* (ITRC 2012; see also USEPA 2014a). Several authors of this report were members of the ITRC "ISM" work group. One of the shortcomings of that document was access to the type of discrete data provided in Part 1 of this report. Such degrees of small-scale variability are predicted by sampling theory, but outside of munitions-related sites had not been widely studied in the field (e.g., see USACE 2009). Left unrecognized, as is almost always the case in discrete sample investigations, the effects of random variability of contaminant concentrations within a targeted area at the scale of a discrete sample can lead to significant error in decision making regarding the extent and magnitude of contamination present. The implications of these factors, once recognized and acknowledged, are likewise significant.

3 CONCEPT OF HOT SPOTS IN EARLY USEPA GUIDANCE

The most common rationale for the collection of discrete soil samples and a central theme in past soil sampling guidance is the need to identify "hot spots" within a site that is under investigation. This information is then used to guide additional sampling, design remedial actions or carry out a risk assessment. The concept of "hot spots" has become irreparably confused since application of the term to the investigation of contaminated soil in the 1980s, however.

As discussed below, early concepts of "hot spots" correspond with more current concepts of mappable "Spill Area" and "Source Area" Decision Units presented in incremental sampling guidance documents (e.g., HDOH 2011; ITRC 2012). In this sense the objectives of guidance documents are very similar. The extent and implications of random, contaminant concentration variability in soil at the scale of a discrete soil sample had not yet been recognized at the time the early guidance documents were under preparation, however. The authors of the guidance documents went on to make a very critical assumption – that contamination in soil within a given spill area or "hot spot" is relatively "homogeneous." This greatly simplified preparation of the guidance. If true then any sample of soil, no matter how small the mass, could be assumed to be reasonably representative of the immediately surrounding area. As will be shown, this ultimately erroneous assumption was never adequately tested in the field during preparation of early site investigation guidance.

3.1 HOT SPOTS AND SPILL AREA DECISION UNITS

Additional background information on the concept of hot spots in early USEPA guidance in provided in Attachment 1. The lack of a clear definition of a "hot spot" and resulting confusion at the scale that hot spots need to be identified and characterized has long plagued environmental investigations. As stated in the USEPA document *Methods for Evaluating the Attainment of Cleanup Standards* (USEPA 1989):

> There is no universal definition of what constitutes a hot spot… This (guidance) models hot spots as localized elliptical areas with concentrations in excess of the cleanup standard… Hot spots are generally small relative to the area being sampled.

A more recent overview of hot spots in light of the more recent transition to Decision Unit and incremental soil sampling methods is provided in the report *Hot Spots: Incremental Sampling Methodology (ISM) FAQs* prepared by the USEPA Superfund office (USEPA 2014a). Two distinct scales of "spots" are discussed (see also USEPA 1987, 1989, 1991, 1992a): 1) Mappable areas of high contamination representing large-scale contaminant trends (e.g., of large enough size to be depicted on a map of the subject site at a scale of interest); and 2) Much smaller, even sample-size spots within a spill area or exposure area that could pose hypothetical, acute toxicity concerns or otherwise exceed again largely hypothetical "not-to-exceed" screening levels. The first types of "hot spots" are referred to as "Spill Area" Decision Units (DUs) in the HDOH

Technical Guidance Manual (HDOH 2008). The ITRC guidance on Incremental Sampling Methodology refers to these as "Source Area" DUs, but the intent is identical.

Although the term is not specifically used, application of the concept of "hot spots" to what are now referred to as "Spill Area" or "Source Area" DUs is clear in key references used to prepare the USEPA guidance document. For example (Gilbert 1987; emphasis added):

> When choosing a sampling plan, one must know the concentration patterns likely to be present in the target population. Advance information on these patterns is used to design a plan that will estimate population parameters with greater accuracy and less cost than can otherwise be achieved. *An example is to divide a heterogeneous target population into more homogeneous parts* or strata and to select samples independently within each part.

In this example the author's concepts of "concentration patterns" and "heterogeneous target populations" correspond with the concept of isolating known or suspected areas of elevated contamination for independent characterization as part of an environmental investigation whenever possible. This is repeated in other USEPA guidance published at the time, including the document *Guidance for Data Useability in Risk Assessment* (USEPA 1991):

> If a chemical can be shown to have dissimilar distributions of concentration in different areas, then the areas should be subdivided…The definition of separate strata or domains should be investigated if a coefficient of variance is above 50%.

The same recommendation is made in the HDOH and ITRC incremental sampling guidance documents in terms of the need to designated separate "Spill Areas" or "Source Areas" Decision Units for known or suspected spill areas in order to improve the precision of the site investigation (see HDOH 2008; ITRC 2012).

3.2 Characterization of Spill Area "Hot Spots"

The next task was to develop a sampling strategy able to identify and characterize spill area "hot spots." Environmental experts at the time were most familiar with discrete sampling methods used to characterize industrial waste. Industrial waste is typically generated under very uniform facility operation conditions. The nature of the waste in terms contaminant concentrations is also likely to be very uniform, with changes occurring only when the process itself is changed. Under these conditions, a grab sample of limited mass can be assumed to represent the larger-scale waste stream reasonably well.

The applicability of this approach to characterization of contaminated soil was largely taken for granted (Gilbert 1987; notations added):

> Stratified random (discrete) sampling is a useful and flexible design for estimating average environmental pollution concentrations… The method makes use of prior information to divide the target population into subgroups (i.e., DUs) that are *internally homogeneous*.

The assumption of small-scale "homogeneity" within contaminated areas was carried forward in subsequent guidance documents. As stated in the USEPA *Data Quality Objectives* guidance (USEPA 1987; emphasis added):

> The probability of not identifying a contaminated zone is related to the area or volume of the *contaminated zone* and the spatial location of the samples… To apply this method, the following assumptions are required… The shape and size of the contaminated zone must be known at least approximately. This known shape will be termed the target… *Any sample located within the contaminated zone will identify the contamination*. These assumptions are not severe and should be met in practice.

This premature assumption is restated in the followup USEPA document *Methods for Evaluating the Attainment of Cleanup Standards* (USEPA 1989):

> When there is little distance between points it is expected that there will be little variability between points.

Similar assumptions were incorporated into even earlier guidance for the use of grids of discrete soil samples to investigation and cleanup PCB-contaminated soil. As discussed in the USEPA document *Verification of PCB Spill Cleanup by Sampling and Analysis* (USEPA 1985; note and emphasis added):

> The implicit assumption (in the use of grids of discrete soil samples) that *residual contamination is equally likely to be present anywhere within the sampling area* is reasonable, at least as a first approximation.

Although ultimately erroneous, these assumptions greatly simplified the preparation of guidance for the investigation of contaminated soil. All that remained was to determine the grid spacing necessary to identify potentially significant spill area hot spots within a site under investigation. Grid spacings were to be based in part on risk, especially in cases where the location of individual spill areas was uncertain (Gilbert 1987):

> The grid spacings are obtained so that the consumer's (i.e., of soil) risk is held to an acceptable level.

In this sense the authors' use of grid spacing is identical to the use of Exposure Area DUs in incremental sampling guidance documents to establish a minimum resolution of an investigation. For example, a maximum DU area of 5,000ft^2 is recommended for investigation of a site that is intended to be used for residential or other sensitive land uses. Designation of appropriate DU

areas is necessarily site-specific. This is recognized in the USEPA *Data Quality Objectives* guidance (USEPA 1987; see also USEPA 1989a, 1991, 1992a):

> Important decisions facing the site manager are how many samples must be taken to investigate the potentially contaminated area and where the samples will be located… The decision maker must determine… the acceptable probability of not finding an existing contaminated zone in the suspected area. For instance, it might be determined that a 20 percent chance *of missing a 100ft-by-100ft (10,000ft^2) contaminated zone* is acceptable but only a *5* percent chance of *missing a 200ft-by-200ft (40,000ft^2) zone* is acceptable.

In this case, however, the authors are assuming that a single, discrete sample will be adequate to represent contaminant levels within any given "contaminated zone," or DU. This is illustrated in Figure 3-1, taken from the USEPA *Methods for Evaluating the Attainment of Cleanup Standards* guidance (USEPA 1989). The various "hot spots" within figure are intended to reflect potential exposure areas of large enough concern to pose a risk to human health if the representative (i.e., mean) contaminant concentration within that area were to exceed a certain level. The figure is used in the document to illustrate how an excessively large grid spacing might inadvertently miss exposure area-size "hot spots" within the hypothetical site.

"Compositing" of samples collected from the grid area was discouraged due to potential dilution of large-scale areas of contamination with large-scale clean areas (see USEPA 1987, 1989a, 1991, 1992a). As stated in the same guidance document (USEPA 1987):

> Compositing does not allow the spatial variability of data to be determined, so the confidence in a composite value may be impossible to determine. Composite samples should not be used when… a measure of spatial variability is important.

In this sense, however, the guidance documents are again describing the need to segregate and independently sample and characterize separate spill or source areas to the extent known practical.

Initial comparison of early site investigation guidance based on grids of discrete sample data and with more up-to-date, Decision Unit and incremental sampling methodologies may at first seem quite dramatic. The basic concepts are in fact very similar – subdivide the targeted site into areas of known or suspected high and low contamination to the extent feasible and then characterize each area independently in order to assess potential risk to human health and the environment. The risk-based concept of maximum-desired grid spacing matches well with the concept of maximum-desired Decision Unit area under incremental sampling methodologies, with the grid area surrounding a single, discrete sample point representing a DU.

Where the early guidance go astray was in the assumption that contamination within a given "hot spot," now referred to as a "spill" or "source" area, could be can assumed to be reasonably

uniform and that a single, discrete sample could be used to represent this area for decision making. The importance and potential consequences of this assumption were not lost on the authors of the guidance. As stated by Gilbert (1987; notations and emphasis added):

> The methods in this chapter require the following assumptions… The definition of "hot spot" (i.e., "Decision Unit") is clear and unambiguous… The types of measurement and the levels of contamination that constitute a hot spot are clearly defined… *There are no measurement misclassification errors*-that is, no errors are made in deciding when a hot spot has been hit (or missed)."

We now know that random contaminant distribution and concentration in soil at the scale of a discrete sample negates the ability to reliably meet the latter requirement. This problem cannot be overcome by the discrete sampling approach envisioned in the early site investigation guidance (and still in widespread use today). Concern by field workers regarding the reliability of the approach could only be addressed by decreasing the spacing of individual discrete samples to the extent feasible, largely controlled by the available project budget, and the passive acceptance of the approach by regulators.

Heterogeneity was already becoming an issue for the use of discrete sample data in relatively new field of risk assessments for contaminated soil. As discussed in Section 7, guidance was developed to estimate and use mean contaminant concentrations for targeted areas (e.g., Exposure Area DUs) for decision making purposes. The authors understood that large numbers of samples were required to reliably estimate the mean and determine if an area was "clean" or "contaminated" in terms of risk to human health and the environment. As discussed in the USEPA document *Guidance for Data Usability in Risk Assessment* (USEPA 1991):

> Information on frequency of occurrence and coefficient of variation determines the number of samples required to adequately characterize exposure pathways and is essential in designing sampling plans. Low frequencies of occurrence and high coefficients of variation mean that more samples will be required to characterize the exposure pathways of interest.

The use of the mean over a large area for decision making purposes in risk assessments was an improvement over a more simplistic and error-prone use of small numbers of discrete samples or even a single sample to define the extent of contamination in general site investigations. Estimate of mean concentrations from sets of discrete samples is still problematic, however, as discussed in Section 7. The first step in addressing these issues is to step back and rethink the scale of decision making driving the investigation to begin with.

3.3 Scale of Decision Making

The concept of "scale" is an important part of geologic as well as environmental field studies. Structures such as fractures and faults are described in terms of increasingly smaller size (Gosh 1993): 1) Macroscopic scale (i.e., structure is so large that it can be studied as a whole only by

preparing a map), 2) Mesoscopic scale" (i.e., scale of both a hand specimen and a single outcrop) and 3) Microscopic scale (i.e., large enough to be observed under an optical microscope but not by the unaided eye as fractures). The term "macroscopic" is synonymous with the concept of "large scale" in environmental investigations, or features large enough to be identifiable at the primary scale of interest, also referred to by field geologists as being "mappable." The terms "mesoscopic" and "microscopic" are more synonymous with the concept of "small scale," as used in this report.

The significance of a particular feature in terms of importance or risk is based in part on its scale of association. For example, a small-scale fault that is restricted to a single outcrop has different significance in terms of risk than a large-scale fault that extends the length of a continent, such as the San Andreas Fault in California. Small-scale, discontinuous faults may *collectively contribute* to the risk posed by larger-scale systems of faults. The objective in terms of earthquake hazard evaluation lies in characterization of the latter. It is important that this characterization capture and accurately represent the contribution to large-scale risk posed by small-scale features. It is both unnecessary and impractical, however, to identify and characterize each and every single small-scale feature within the large-scale system as a whole.

A similar sense of scale and understanding of risk also applies to the magnitude and distribution of contaminants in soil. The HDOH Technical Guidance Manual discusses designation of large-scale "Decision Units" for characterization. The ITRC *Incremental Sampling Methodology* document goes into more detail and discusses variability of contaminant distribution in soil in terms of "large-scale," "short-scale" and "micro -scale" heterogeneity (ITRC 2012; see also USEPA 2014a). "Micro-scale" refers to heterogeneity within a single sample. This can range from variability in contaminant concentration between two side-by-side particles of soil up to variability between different subsample masses of soil within a single sample. "Short-scale," a term adopted from the mining industry, refers to differences in contaminant concentrations at the scale of co-located samples. Note that this term is used in the mining industry to describe variability of mineral concentrations in a moving train or conveyor belt of crushed ore (Pitard 1993, 2005, 2009; Minnitt et al 2007). "Large-scale" is described as mappable, distinct patterns of contamination of a large enough extent and magnitude to pose potential long-term risks to human health and the environment.

For simplicity and to avoid confusion with terms used by the mining industry, the term "small-scale" in this report is used to describe differences in contaminant concentration and distribution both within a single discrete sample (i.e., "intra-sample variability") and between co-located discrete samples (i.e., "inter-sample" variability). This includes the concept of "micro-scale" and "short-scale" heterogeneity discussed in the ITRC document.

3.4 Hot Areas and Hot Spots at Study Site C

Excellent examples of large- and small-scale "hot spots" are provided by discrete soil sample data for Study Site C. Figure 3-2 depicts a large-scale, six-acre "hot spot" or more appropriately

"hot area" of PCB contaminated soil identified within an 89-acre property formerly used as a government radio transmitter station (see Section 2.3 in Part 1). Study area C is located within the identified area of contamination (see Figure 3-2).

Figure 3-3, in contrast, depicts a "hot spot" of very high PCB concentrations in soil that truly is a "spot." The data depict test results for six discrete soil samples collected within a one-meter square area around Grid Point 24 (refer to Figure 2-10 in Part 1). Samples VOA-24 (A) through VOA-24 (E) were processed and tested at the laboratory in the same manner as done for Multi Increment samples. The data for these five "co-located" discrete samples should, in this case, reliable represent the true, mean concentration of PCBs in the samples. A sixth sample was collected from within the area encompassed by the first five samples, represented by the jars in Figure 3-3. This sample was divided into ten subsamples in the field. Each subsample was then tested individually for PCBs using standard laboratory procedures as part of the "intra-sample" heterogeneity part of the field study.

The results for this particular grid point were especially enlightening. The reported concentration of PCBs in the five, processed discrete samples ranged from 4.9 mg/kg to 91 mg/kg. The mean concentration of PCBs in the sixth sample in terms of the average of the ten subsamples, however, was 2,412 mg/kg. The area represented by this sample is no more than 15cm across, covering perhaps 1,000cm^2. The chance of identifying all such small, isolated spots within an area targeted for investigation using discrete samples, as would presumably be required, is obviously small. The mass of the sample was approximately 500 grams. For comparison, the mass of soil in the upper two inches of the 6,000 ft^2 area is estimated to be approximately 30 metric tons (30,000kg) or 60,000 potential 500 gram masses of soil for testing. The chance of clearing a property of the potential presence of all such small spots using randomly collected discrete samples is likewise negligible. Removal of any such "hot spots" fortuitously found as part of a discrete sample investigation cannot be assumed to have significantly reduced the mean concentration of contaminants in the area as a whole or the overall risk to the human health and the environment, even if "confirmation" samples around the point imply that the "hot spot" was successfully removed. The practice of excavating "sample points" is nonetheless still very common in many parts of the country.

Figure 3-4 depicts what might be called a "micro" hot spot – a possible PCB-infused nugget of "fossilized" and degraded mineral oil in fine-grained soil weighing only a few milligrams. Concentration of PCBs over 50,000 mg/kg (5%) have been reported for discrete samples collected at the site. If the material between soil particles within the nugget could be tested, then a "maximum" concentration of PCBs approaching one million parts-per-million (100%) could, in theory, be identified.

These examples illustrate the futility of attempting to identify the "maximum" concentration of a contaminant in soil. At some small scale, the maximum concentration of a contaminant in soil, if present, will always be 100%. The concept of scale is thus critical for both establishing the

objectives of an environmental investigation and designing an appropriate site investigation plan. As discussed in Attachment 1, a focus on large-scale, spill area “hot spots” is clear in early USEPA site investigation guidance. Subsequent guidance unfortunately confused the ability and necessity to investigate sites in terms of hypothetical, acute toxicity concerns at the scale of an individual, discrete sample (see Section 7.4 and Attachment 1). The investigation of contaminated sites down to the resolution of an individual, discrete soil sample has never been routinely required and indeed from both a cost and technical standpoint is not feasible in field investigations. The habit of collecting discrete soil samples has, however, lived on even though the original justification for doing so no longer exists.

4 COMPARISON TO SOIL SCREENING LEVELS

4.1 RISK-BASED SCREENING LEVELS AND MEAN CONTAMINANT CONCENTRATIONS

Risk-based soil screening levels, including Environmental Action Levels (EALs) published by the HEER office (HDOH 2011) as well as the Regional Screening Levels (RSLs) published by the USEPA (USEPA 2014b), are intended for comparison to the mean concentration of a contaminant within a targeted area of concern or "Decision Unit." This was made clear in early USEPA soil sampling guidance (USEPA 1989, emphasis added; see also USEPA 2014a and Attachment 1):

> *The concentration term in the intake equation is the arithmetic average of the concentration that is contacted over the exposure period.* Although this concentration does not reflect the maximum concentration that could be contacted at any one time, it is regarded as a reasonable estimate of the concentration likely to be contacted over time. This is because in most situations, assuming long-term contact with the maximum concentration is not reasonable.

Screening levels to assess chronic health risks, for example, are designed to consider regular but random exposure to contaminants in soil within a targeted, "exposure area" over many years. Risk is assessed in terms of average daily exposure to contaminants in soil over this time period. The range of contaminant concentrations in soil at the scale of assumed exposure (e.g., 100 to 200 mg/day) is not important, provided that this is accurately represented in the mean contaminant concentration estimated for the subject area and volume of soil.

Risk-based screening levels are *not* designed for direct comparison to individual, discrete sample data, since any given sample point cannot be assumed to represent the average concentration of a contaminant in soil over the exposure area or more specifically the Decision Unit as a whole. This is implied in terms of what would today be referred to as "Exposure Area Decision Units" in the USEPA document *Guidance on Surface Soil Cleanup at Hazardous Waste Sites* (USEPA 2005b):

> For sampling data to accurately represent the exposure concentration, they should generally be representative of the contaminant populations at the same scales as the remediation decisions and the exposures on which those decisions are based.

As discussed in the HEER Technical Guidance Manual (HDOH 2008), grids of discrete data can sometimes be useful for gross separation of contaminated versus clean areas in order to optimize DUs for more intensive, incremental sampling. As discussed in Section 6, the reliability of the discrete data depends in part on the magnitude of small-scale variability of contaminant concentrations in soil with respect to the target screening level.

This is illustrated in Figure 4-1(see also ITRC 2012). Area "A" in the figure represents an area of heavy contamination, where the concentration of a contaminant and the overwhelming majority of discrete samples (and even laboratory subsamples) exceeds the target screening level, even though the small-scale variability of contaminant concentrations in soil might be very high. Under these circumstances, a small number of discrete samples from the area will in most cases accurately identify a potential health risk for the area as a whole. The risk of false negatives, where the reported concentration of a contaminant in a discrete sample falls below a screening level even though the mean concentration for the area as a whole exceeds the screening level, is present but relatively low.

As small-scale variability increases and/or the mean concentration of the contaminant for the targeted area approaches the screening level, the reliability of individual discrete samples to accurately identify areas of potential concern decreases. This is illustrated in Area B of Figure 4-1. Within this area, the small-scale variability of contaminant concentrations in soil now straddles the target screening level. Some samples will fall above the screening level and some below, even though the overall average of the contaminant within the area exceeds the screening level. "False negatives" reflected by individual discrete samples collected within this area are unavoidable. This scenario is highlighted by discrete sample data for lead at Study Site B, where concentrations of lead in discrete samples around the majority of grid points fortuitously fall both above and below the HDOH screening level of 200 mg/kg (refer to Section 6 and Figure 6-1 in Part 1).

Consider, for example, cases where the mean concentration of a contaminant in soil is driven well above the median due to the presence of small, scattered pockets of highly concentrated contamination within an exposure area. The concentration of the contaminant in the majority of discrete samples collected from the area will necessarily fall below the mean concentration for the area as a whole, increasing the risk of false negatives. Unrecognized, this can cause the premature termination of site investigations as sample-size, false negatives are encountered in areas of otherwise unacceptably contaminated soil (refer to Section 5). Such "outlier" pockets of high contamination drive both the overall, mean concentration of the contaminant in soil as well as the direct exposure risk posed to human and ecological receptors. Characterization of the area can only be considered complete when a representative number of higher-concentration areas, as well as lower concentration areas, are included in estimation of the mean (see also USEPA 2014a).

Disregarding data for high-concentration areas of soil within a targeted area is of course inappropriate, even though it is sometimes done as part of a risk assessment in order to force a data set to fit a geostatistical model (refer to Section 7). The statistical basis for doing so can sometimes seem rational. From a field perspective it is illogical, however. Such small-scale hot spots are part of the overall exposure area and play an important role in long-term health risk. Under ideal circumstances the entire volume of soil within a designated exposure area would be submitted to a laboratory for extraction and analysis. This is the basic concept of a "Decision

Unit." Removing scattered "hot spots" from the soil prior to analysis would not be acceptable, given the objective to determine the mean concentration of the contaminant in the soil as a whole. If the concentration of a contaminant in one out of thirty discrete samples is highly elevated in comparison to the other samples, this is more appropriately interpreted to indicate that exposure to contaminants in soil will be highly elevated in one out of thirty exposure events (e.g., one day a month). Ignoring "outliers" within exposure areas in the field would otherwise be equivalent to removing particles with "outlier," high concentrations of a contaminant from of a discrete soil sample prior to analysis. While this might help ensure that laboratory replicate data more closely correlate, it is not justifiable in terms of risk assessment. The inclusion of "outlier" data as part of a risk assessment and relate topics are discussed in more detail in Section 7.

"False positives" in otherwise "clean" areas are likewise to be expected when a site is investigated using discrete soil samples. This is illustrated in Area C of Figure 4-1. The mean concentration of the contaminant in the area is below the target, risk-based screening level. It is inevitable, however, that the concentration of the contaminant in small masses of the soil within the area will fall above the screening level if enough samples were to be collected.

Removal of soil around a discrete sample point where the initial concentration of a contaminant was reported above a screening level as part of a site remediation cannot be assumed to significantly reduce the mean contaminant concentration for the area as a whole, even though this is still routinely done in some states. Doing so is equivalent to removal of a single, randomly selected red cell in Figure 4-1 and assuming that the average concentration of the area as a whole has been significantly reduced. In practice this would be impossible to know without knowledge of every single sample-size mass of soil within the area. Recalculation of a mean, contaminant concentration based removal of "hot spots" identified based on a single or even small group of samples is likewise invalid, since the sample set as a whole has now been biased. This is true even if "confirmation" samples are collected around the excavated sample point. Re-estimation of a mean contaminant concentration for the area as a whole would require collection of a new, independent set of discrete samples from separate and randomly selected points. Even this may not be fully adequate, since the representativeness of any single set of discrete samples is unknown. These issues are discussed in more detail in Section 6, Section 7 and Attachment 1.

4.2 Comparison of Study Site Data to Screening Levels

The effects of small-scale variability on the comparison of discrete sample data to target screening levels are highlighted in data presented in Part 1 of this study. Small-scale variability at the scale of the mass of soil analyzed by the laboratory introduces an important, limiting factor in the direct comparison of discrete sample data points to screening levels. As discussed in Section 2, this point was not lost to authors of early guidance on the investigation of contaminated soil (USEPA 1987):

> To apply this method (it must be assumed that) any sample located within the contaminated zone will identify the contamination.

At the time this was assumed to be the case and existing guidance for testing of water, industrial waste and similar media was adopted for characterization of contaminated soil. This is discussed in more detail in Section 9.

The estimated, median Relative Percent Differences between minimum and maximum contaminant concentrations in soil with respect to the mean around individual grid points (i.e., within 0.5m) are 96% for Study Site A, 650% for Study Site B and 3,082% for Study Site C (Table 4-1; refer to Section 5.1 in Part 1). The magnitude of small-scale variability of contaminant concentrations in soil has significant implications for direct comparison of discrete sample data to screening levels for targeted contaminants.

Consider, for example, a single, hypothetical discrete soil sample collected from a grid point at Study Site B. Assume an RPD for minimum and maximum lead concentrations respect to the mean of +/-650%. The maximum concentration of lead in a discrete sample around a grid point relative to the minimum concentration is therefore predicted as:

Maximum Concentration = Minimum Concentration + 650% Minimum Concentration.

A random, one-gram mass of soil is removed from the sample and tested by the laboratory (standard mass for metals). A concentration of 100 mg/kg lead is subsequently reported. The study data suggest that the concentration of lead in additional, discrete samples collected within 0.5m of the original grid point could be as high as 750 mg/kg (i.e., reported 100 mg/kg = minimum-predicted concentration) or as low as 13 mg/kg (i.e., reported 100 mg/kg = maximum-predicted concentration). Based on a single sample, the range of lead around the grid point can at best be assumed to range from either 100 mg/kg to 750 mg/kg or 13 mg/kg to 100 mg/kg. The true range is unknowable without additional, detailed testing. Comparison of the discrete sample data point to a target screening level involves a high degree of uncertainty, since the data cannot be assumed to reflect the mean concentration of lead in either the sample collected or within the immediate vicinity of the sample collection point.

This is highlighted by a comparison of box plots of intra-sample variability for each study site to hypothetical, target screening levels. Figures 4-2 through 4-4 depict box plots for estimated total variability of contaminants in discrete samples for the three study areas. Total variability was estimated based on adjustment of measured concentrations for each processed sample, assumed to represent the mean, in terms of the RPD of minimum- and maximum-reported concentrations of the contaminant for the correlative, intra-sample variability data at the same grid point (refer to Section 4 of Part 1).

4.2.1 Study Site A Box Plots (arsenic)

Box plots depicted in Figure 4-2 indicate the total, estimated small-scale variability of (total) arsenic concentrations in discrete sample masses of soil around grid points at Study Site A, in order of lower to higher median concentration. Variability appears to increase somewhat with increasing concentrations of total arsenic. This could be due to an increasing number of small "nuggets" of arsenic-rich, iron hydroxide in the soil (refer to Part 1, Section 5.1; see also Cutler 2006, 2011). Note that the intra-sample data, based on XRF analysis, and the inter-sample data based on extraction Method 6010B and used to prepare the box plots are not directly comparable. Consistently higher concentrations of arsenic were reported using the XRF (refer to Section 4 in Part 1). The relative, total variability should be similar, however, regardless of the test method used.

4.2.2 Study Site B Box Plots (lead)

The small-scale variability of lead concentrations around grid points at Study Site B is noticeably higher. Box plots of total, estimated variability depicted in Figure 4-3 coincidentally fall both above and below the HDOH residential soil action level for lead of 200 mg/kg (HDOH 2011) for twenty-three of the twenty-four grid points. Discrete sample concentrations at twenty of the twenty-four grid points similarly fall both above and below the USEPA residential soil screening level of 400 mg/kg (USEPA 2014b). The wide range of estimated concentrations matches well with the assumed, incomplete mixture of lead-contaminated ash and fill soil at the site (refer to Part 1, Section 2.2). Lower concentrations of lead in soil suggest that a small pocket of fill material was tested, with concentrations approaching natural background (upper threshold limit 75 mg/kg; HDOH 2012). Higher concentrations of lead suggest that the small mass tested included a significantly higher proportion of ash. Lead data for ash originally generated at the Waipahu incinerator were not immediately available. Data for ash from H-Power, the currently operating, municipal incinerator on the island, suggest that the concentration of lead in ash typically ranges from 1,000 mg/kg to 4,000 mg/kg (Shulgin 2008).

The implications for Study Site B are significant. Data for discrete soil samples cannot be reliably assumed to represent either the soil immediately surrounding a sample collection point or sample submitted to the laboratory for analysis (samples were "unprocessed", so assumes thorough processing and subsampling was not carried out by the lab).. Direct comparison of data for single grid points could in theory declare the site to be either completely "clean" (i.e., lead concentration $\leq$200 mg/kg) or completely "contaminated" (i.e., lead concentration >200 mg/kg) depending on the gram of soil that happens to be tested around a particular grid point (see also Figure 6-1 of Part 1 report).

This is further illustrated in Figure 4-5. Like-colored gumballs reflect an assumption that discrete samples colored at individual points within the site are "uniform" and that the small mass of soil removed for testing are representative of the samples as a whole. This reflects a key assumption made in early, USEPA site investigation guidance (refer to Section 3). Resulting

"data" for the samples are used to estimate large-scale contaminant distribution across the site, with higher concentrations of contamination seemingly apparent in the lower area of the site (red gumball sample) than the upper area (green gumball sample), separated by an area of intermediate contamination (yellow gumball sample).

This seems reasonable enough, but assume that more detailed testing of the unprocessed "samples" as well as the collection of additional, co-located samples around collection points in fact identifies considerable and consistent variability of lead concentrations in the soil at the scale of a discrete sample and subsample mass (Figure 4-6). The mean concentration of the contaminant in the three samples in this hypothetical example is in fact identical, as it is across the entire site. Patterns implied by the random collection and testing of single, small masses from unprocessed, discrete samples are not "real;" they are simply artifacts of or random, small-scale heterogeneity and inadequate processing and subsampling of the samples prior to analysis. The samples were too small to capture and represent random, small-scale variability both within a single mass of "soil" collected and between co-located samples.

Similar patterns of false, large-scale heterogeneity would be identified by testing of alternative subsample masses, but the locations of apparent "hot spots" and "cold spots" would change. The fictitious "hot spots" and "cold spots" are representative only of the small mass of sample collected in the field and the analysis mass subsampled for testing in the laboratory. These samples would simply reflect the range of small-scale contaminant concentration variability within the area as a whole and the analysis mass tested by the laboratory. These types of artificial, contaminant patterns are common in computer-generated isoconcentration maps, as discussed in Section 6.

In combination, it is possible that a discrete sample data set might be of adequate representativeness to estimate a mean concentration of the contaminant in soil for a targeted area as a whole. As discussed in Section 7, however, it is difficult to ascertain the true representativeness of a data set in the absence of independent replicate sets of data from the same site.

A traditional, discrete sample investigation of Study Site B in the absence of an understanding of small-scale, distributional heterogeneity of lead in the soil would identify apparent but ultimately erroneous and misleading patterns of randomly scattered "hot spots" and "cold spots" of lead-contaminated soil across the site. Apparent "outlier" data points might be excluded from calculation of an area-wide mean in order to force the geostatistical model to produce a result with a "low" margin of error (see Section 7). Attempts to remove soil around individual "hot spots" would likely lead to a need for repeated over-excavation as "confirmation samples" fail in what initially appeared to be clean areas. Estimation of a representative mean concentration of lead following removal of identified "hot spots" and in the absence of a completely new and independent set of samples would be inaccurate and underestimate the mean concentration of lead in the remaining soil. This is because the remaining sample points would no longer be

representative of contaminant distribution in the soil as a whole. Such problems plague cases where discrete data are used to guide site investigation and remedial actions.

4.2.3 STUDY SITE C BOX PLOTS (TOTAL PCBS)

Box plots for data from Study Site C depict the extreme variability of total PCB concentrations in both subsamples of individual discrete samples as well as estimated total variability around individual grid points when data for processed samples is considered (Figure 4-4). The high, small-scale variability highlights an even greater chance for decision error based on comparison of screening levels to individual discrete data. Such comparisons are again highly prone to false negatives and early termination of the investigation. As discussed in Sections 7 and 8, such high variability can also confound estimation of mean PCB concentrations for the study site as a whole and cause confusion over the incorporation of seemingly "outlier" data in these calculations.

These examples highlight the limitations of comparisons of individual, discrete sample data points to screening levels intended for comparisons to mean concentrations over large areas. As discussed in Section 5, the risk for potential "false negatives" and premature termination of the site investigation is also high. "False positives" in otherwise clean areas can lead to equally erroneous remedial decisions. These limitations generally preclude the use of discrete sample data to reliably estimate the extent of contamination in soil that could pose a risk to human health and the environment.

5 ESTIMATION OF EXTENT OF CONTAMINATION

5.1 SMALL-SCALE SPATIAL VARIABILITY AND LARGE-SCALE TRENDS

Random, small-scale variability in contaminant concentrations can significantly affect the use of discrete, soil sample data to estimate the lateral and vertical extent of large-scale, mappable trends of interest. Consider the following text from the 1987 USEPA document (USEPA 1987):

> The magnitude of the difference in contaminant concentrations in samples separated by a fixed distance is a measure of spatial variability. The level of spatial variability is site and contaminant specific. When spatial variability is high, a single sample is likely to be unrepresentative of the average contaminant concentration in the media surrounding the sample. Although it is important to recognize the nature of spatial variability at all times, it is crucial when the properties observed in a single sample will be extrapolated to the surrounding volume.

The authors understood the importance of spatial variability as a controlling factor in the reliability of sample data to identify and accurately map areas of contamination. Upon closer inspection, however, it is apparent that they were discussing large-scale, mappable trends of variability and were unaware of the nature and pervasiveness of random, small-scale variability of the type evaluated in this study. The document goes on to prematurely state that (USEPA 1987; emphasis and notation added):

> Grab samples are discrete aliquots *which are representative* of a specific location at a specific point in time… Grab samples offer the most information regarding (large-scale) contaminant variability.

This assumption is repeated and cemented in subsequent guidance documents, as noted in the previous section for the document *Methods for Evaluating the Attainment of Cleanup Standards* (USEPA 1989). Such assumptions were used to justify the use of individual, discrete sample data points to map the extent of contamination in soil above screening levels, a practice still used in many areas of the country today.

Some guidance documents at the time called for the collection of “co-located ” and “replicate” soil samples in order to assess smaller-scale, spatial variability and assess the precision of estimated mean contaminant concentrations within targeted areas (e.g., USEPA 1987). As the name implies, co-located samples represent closely spaced samples assumed to be representative of the same area. In the terminology of that time “field replicates” are subsamples or “splits” of an initial discrete sample prepared in the field for separate analysis in order to assess “homogeneity” (i.e., small-scale variability) as well as bias introduced in the data by handling, shipping and storage. The guidance anticipated that the variability of contaminant concentrations within a single sample would be minimal in comparison to potential (but presumed very low) variability between co-located samples. Laboratory replicates as defined are subsamples of a

larger sample, independently tested to assess the precision of sample processing and analysis. In practice this is more often done with matrix spikes, rather than sample splits, in order to focus on the precision of the analytical method itself and avoid bias due to subsampling error.

The influence of random, small-scale variability both in the field at the laboratory was anticipated to be minimal, however. This is reflected in the minimal number of co-located and replicate samples recommended in the guidance documents (USEPA 1987; see also USEPA 1990, 1991):

> The following are suggested guidelines for the inclusion of collocated (sic) and replicated samples in field programs… Soil, sediments and solids - one out of every 20 investigative samples should be field replicated or collocated. To estimate sampling error, collocated and field replicated samples should be of the same investigative sample. These samples should be spread out over the sampling event, preferably one per each day of sampling.

In reality the number of co-located and "replicate" samples recommended was far too small to identify and appreciate the true nature of contaminant heterogeneity in soil at the scale of the discrete samples being collected and tested. Even then, difference in data for co-located samples (rarely collected) and laboratory replicates are most conveniently assumed to reflect error associated with laboratory analysis rather than error in sample collection or sample processing at the laboratory (see also ITRC 2012). If doubts are raised, the highest contaminant concentration reported is simply used for decision making purposes, even though it is no more likely to be representative of true field conditions than the lowest concentration reported.

Random, small-scale variability in contrast significantly limits the reliability of discrete samples to identify and map out large-scale patterns of contamination in soil. This would only be possible for instances of extremely low, distributional heterogeneity (i.e., "homogenous" small-scale contaminant distribution). Accurate identification and mapping of large-scale trends is in contrast more efficiently and accurately accomplished by comparing mean contaminant concentrations for well-thought-out, Decision Unit areas of soil of a size sufficiently large to capture and overcome random, small-scale variability (see HDOH 2008).

It is important to understand the economic and political environment of the time that these documents were being published before judging the guidance too harshly. Far reaching, new environmental regulatory requirements were being imposed on industries and businesses. An understanding of the risk posed by long-term, chronic, exposure to very low concentrations of contaminants in soil was just developing. Cost was (and still is) a significant factor in encouraging the private sector to comply with the new regulations and undertake extensive investigations of potential contamination on private and government properties. Protocols for testing of liquids, industrial wastes and similar material that could be assumed to be relatively "homogenous" were well established and understood. An understanding of sampling theory as used in the mining and agricultural industries to accommodate highly heterogeneous media was

all but absent in the environmental community. Time and costs were important factors. Assuming a relative "homogeneity" of contaminant concentrations in soil around a given sample location and consistent, mappable trends between discrete sample points would in theory greatly simplify environmental investigations. In practice this was often not the case, with investigations sometimes drawn out years and repeated and costly efforts required to remediate identified contamination.

5.2 False Negatives and Underestimation of Extent of Contamination

Drawing a line between "contaminated" and "clean" areas of a site is integral to environmental investigations and necessary to design appropriate remedial actions. As described in Section 4, however, this process is not as straightforward as traditional, discrete sample investigation methodologies might otherwise imply. The risk of "false negatives" (and positives) when discrete samples are used to estimate the extent of contamination in soil was recognized in early USEPA guidance documents (USEPA 1992b):

> High coefficients of variation mean that more samples will be required to characterize the exposure pathways of interest. Potential false negatives occur as variability increases and occurrence rates decrease.

Consider, for example, the hypothetical "site" outlined in Figure 5-1. Data for closely-spaced discrete samples are noted by red (above screening level), yellow (above background but below screening level) and green (not detected) dots. Dashed red lines indicate areas interpreted to require soil removal. The approach seems straightforward enough and recognizable by most field practitioners today.

The figure in fact represents the "hotness" of a grid of discrete sample points randomly placed on a cutout of the Jackson Pollock painting discussed in Part 1 of the study report (Figure 5-2). A close up inspection of each grid point was made to estimate the relative amount of dark paint present and a red, yellow or green dot assigned. The resulting map of hypothetical, discrete sample data is quite misleading in terms of the actual extent of "contamination" at the "site." Individual data points do not fully reflect the nature of contamination in the surrounding area and may or may not be reproducible if co-located samples were collected. Shifting the grid a small amount in any direction could result in a significant shift of apparent "hot" and "cold" areas, reflecting the high, "inter-sample" variability of the painting (see also Figure 5-8 in Part 1).

The sizes of the sample points depicted in Figure 5-2 (estimated 5-10cm across) are also larger than the mass of soil likely to be collected in the field. This and the routine lack of rigorous sample processing and the small soil subsample mass analyzed by a laboratory (typically a small pinch to approximate three cubic centimeters, or 1 to 30 grams) contribute to sampling related errors.

The high, distributional heterogeneity observed for total PCBs in soil at Study Site C seems to mimic the discontinuous and random distribution of paint across Pollock's canvas. The release

of PCB oil to dry soil can be expected to form droplets similar to those formed by dripping paint onto a canvas (refer to Section 5.1 in Part 1). This would result in scattered clumps of dried, PCB-concentrated aggregates or nuggets in soil. As observed for soil samples at Study Site C, if small, discrete areas of the canvas were tested for paint then false negatives would quickly be encountered, as the sample fell between areas of heavier "contamination."

The same is true for milk-contaminated soil depicted in Figure 5-9 of Part 1. Assume for example that the milk was present but invisible. The potential for underestimation of the extent of contamination based on small, discrete soil samples would be very high. Accurate estimation of extent of contamination and avoidance of confusion due to false negatives is only possible when the area and volume of the sample collected is large enough to capture and overcome small-scale, random variability.

5.3 Transitional Zones at Study Site C

This type of small-scale variability significantly compromises the use of discrete sample data to establish a reliable, clean boundary around an area of contaminated soil. Failure to recognize zones where the small-scale variability of contaminant concentrations in soil begins to span both above and below a target screening level can lead to the premature termination of site investigations as false negatives are encountered (refer to Area A in Figure 4-1).

Consider implications for estimation of the extent of PCB contaminated soil at Study Site C. Figure 5-3 illustrates the estimated range of PCB concentrations in discrete samples around individual grid points relative to the HDOH residential soil action level of 1.1 mg/kg (see also in Figure 4-4; (refer also to Table 4-20 in Part 1). Grid points where the estimated range of total PCBs in discrete soil samples falls entirely above 1.1 mg/kg are highlighted in red. Points where the estimated range of PCBs falls below 1.1 mg/kg are highlighted in green. Points where the estimated range of PCBs in discrete soil samples falls both above and below the action level of 1.1 mg/kg are highlighted in yellow.

As an initial interpretation of the data, it is reasonable to assume that the mean concentration of PCBs in the area of the site highlighted by consistently red grid points as a whole is indeed likely to fall above the target action level. Points highlighted in yellow represent areas within and along the margin of heavy contamination where concentrations of PCBs in discrete samples fall both above and below the action level of 1.1 mg/kg but the mean concentration of PCBs for the area as a whole is likely to exceed this level. These areas would be highly prone to "false negatives" and "failed confirmation samples" in traditional, discrete sample investigations, with pre-excavation, perimeter samples below action levels and post excavations above action levels. Such transitional zones could characterize the entire site, as is almost the case for Study Site B (see Section 3.2 above), depending on the screening level being used and the nature of small-scale, distributional heterogeneity of contaminants in the soil.

The unreliability of single, discrete sample point data for decision making is even more dramatic in comparison of the total, estimated range of PCB concentrations around grid points to the Toxic Substances Control Act (TSCA) limit of 50 mg/kg for disposal of soil at a municipal landfill (Figure 5-4; USEPA 1998a). Concentrations of PCBs in discrete samples were measured or are estimated to fall both above and below 50 mg/kg around twelve of the twenty-four grid points, rendering the discrete sample data essentially useless to assess this hypothetical concern (refer to Table 4-20 in Part 1). Problems with the use of discrete sample data under TSCA are discussed in more detail in Section 8.3.

Seemingly safe "red" and "green" areas on Figure 5-3 and Figure 5-4 are also prone to this problem. Additional discrete samples from these areas would likely identify the presence of isolated "cold spots" in the former and isolated "hot spots" in the latter. Consider again for example samples collected from Grid Point 24 (see Figure 3.2 and "micro hot spot" discussion in Section 3.3). The concentration of PCBs in five processed, discrete samples collected from the location ranged from 4.9 mg/kg to 91 mg/kg, with four of five samples below the mean concentration of 25 mg/kg (see Table 4-18 in Part 1). In all probability this sample point would be placed outside of an area of soil that exceeds a screening level of 25 mg/kg, used as a cleanup level under TSCA for sites with restricted access (USEPA 1998a).

Concentrations of PCBs in a sixth sample split into ten subsamples for evaluation of intra-sample variability were dramatically higher, ranging from 810 mg/kg to 5,700 mg/kg, increasing the mean PCB concentration for soil around the grid point to over 400 mg/kg. This sample, collected directly within a cluster of processed discrete samples, reflects a "hot spot" in the true sense of the word, with the spot itself being no more than one-foot to two-feet across. Indeed, it is not inconceivable that most if not all of the grid points would fall within the "yellow" transitional category if each, ten-gram mass of soil within the 0.5m radius of the grid point used in the study could be tested (approximately 75,000 grams or 7,500 potential laboratory subsample masses to a depth of four inches). Attempting to accurately map large-scale trends of PCB contamination across the site using single, discrete samples from individual grid points would be highly challenging, at best.

The presence of small, isolated pockets of soil with very high concentrations of PCBs at Study Site C has also been documented in earlier investigation reports for the site (USCG 2011). The boundaries of the Study Area C discussed in this report are noted on the figures (refer also to Figure 5-3). Figure 5-5 and Figure 5-6 depict concentrations of PCBs above 1 mg/kg and 50 mg/kg, respectively, identified in discrete samples collected in the same vicinity as the study area in a 2011 investigation (study area boundaries noted on map). Twenty grams of soil were collected from the surface and again at a depth of two and four feet at each grid point. Each sample was tested for total PCBs using a RaPID immunoassay test kit.

Figure 5-5 depicts concentrations of total PCBs in discrete samples greater than 1.0 mg/kg. Note the occurrence of apparent "false negatives" in the eastern part of the area in comparison to data

for discrete samples for this study (Figure 5-3). Figure 5-6 depicts isolated "hot spots" of discrete samples with concentrations of PCBs greater than 50 mg/kg (compare to Figure 5-4). This almost certainly reflects widespread, small-scale, distributional heterogeneity rather than the presence of fortuitously identified, spot-size areas of higher contamination.

The problem of false negatives and potential premature termination of site investigations is an artifact of the sampling method being used. "False positives" within areas where the mean concentration of contaminants is otherwise below action levels can lead to a waste of resources on additional but unnecessary investigation and remedial actions. The mass of discrete samples is simply inadequate to overcome the random, small-scale, distributional heterogeneity of contaminants in soil under many if not most release scenarios. The same degree of small-scale variability can be anticipated with depth, with a similarly high potential for false negatives as well as false positives (see red and blue cells in Figures 5-4 and 5-5; refer also Schumacher 2000, Feenstra 2003). Random, small-scale variability of contaminant concentrations in soil unrelated to larger-scale, vertical trends again inhibit the reliability of interpolating between individual discrete sample points. This is the primary cause of "failed" confirmation samples and the need for remobilization and over-excavation of soil that was thought from initial, discrete sample data to be clean.

These field-based examples highlight the unreliability of using individual, discrete sample data for decision making purposes when the inherent, small-scale variability of contaminant concentrations is great enough to span the target screening level over very short distances. As is the case for HDOH EALs and USEPA RSLs, cleanup levels and disposal criteria presented in TSCA regulations are tied to long-term, chronic exposure to PCBs in soil and are intended for comparison to the *mean* concentration of total PCBs for pre-designated areas and volumes of soil (see Section 4.1). As discussed in Section 8.3, this point is missed in past and even current USEPA investigation guidance and policy. As discussed in Attachment 1, "Iterative Truncation" methods to surgically remove seemingly isolated "hot spots" defined by individual or small numbers of discrete samples can give a false sense that the overall risk posed by contamination in the area has been significantly reduced. The small scale variability of contaminant concentrations also results in potential errors when relying on isoconcentration maps based on discrete sample data.

6 RELIABILITY OF ISOCONCENTRATION MAPS

The ability of statistical methods to distinguish real from artificial patterns is ultimately tied to the reliability of the data set in question. Isoconcentration maps generated by geostatistical software are a good example. Such maps do not "predict" anything; they tell us exactly what we tell them to tell us in terms of the data provided and the models incorporated into the programs. They will also always give us an answer, regardless of whether the data provided are in fact representative of what we are attempting to draw conclusions on in the field. Errors in maps will only be identified when additional samples are collected.

Isoconcentration maps are a common and useful tool to visualize contaminant plumes in groundwater (e.g., Lu and Wong 2008). Simple maps can be drawn by hand with experience. Geostatistical models can be used for more sophisticated interpolation of contaminant concentrations between input data points, based for example on the weighted average of other points within the same neighborhood. Some degree of error is inevitable due to small-scale, random "noise" in the data. Mapping programs have proven reasonably accurate for prediction of large-scale contaminant trends within areas of available data, especially in downgradient areas away from the immediate source of the release. Wells installed within mapped areas typically identify concentrations of contaminants within a relatively narrow range of error. Surprises tend to be the exception, rather than the rule. In such cases the data are indicating that small-scale, random variability is relatively low at the scale of the samples being collected and tested (typically a few liters). Trends between sample points can indeed be assumed to be linear and predictable based on a relatively small set of sample points.

Are isoconcentration maps generated for soil data similarly reliable, however? Yes and no. When properly carried out and interpreted, tight grids of discrete samples can be useful for identification of large-scale areas of elevated contamination within an area of investigation that could pose direct exposure or other potential environmental concerns (see HEER *Technical Guidance Manual*; HDOH 2008). As discussed in the previous section, however, caution must be taken against over-interpretation of maps generated based on discrete sample data. The ability to recognize and separate artificial noise from true patterns of contaminant distribution is especially critical. Surprisingly, considering the hundreds of thousands of sites investigated over the past 30 years, very few detailed field studies of the reliability of discrete soil data to generate reproducible patterns of contaminant distribution have been carried out.

Use of geostatistical methods to interpolate contaminant concentrations between data points requires several critical assumptions, including (USEPA 1987): 1) The distributional heterogeneity of contaminant concentrations in soil at the scale represented by individual, sample data points is well understood, 2) The trend between points is linear, for example progressively lower to higher, 3) Any sample located within interpolated isopleth contours will identify the contamination. The first point is especially critical and controls whether the latter two criteria can be met for a given set of data. Trends between data points will only be linear and predictable

if the data for an individual point is representative of the large-scale trend of interest. This requires that the sample tested be of sufficient area and volume to capture and overcome random, small-scale variability.

Aelion et al (2009) evaluate the influence of small-scale variability on the reliability of discrete sample data for geostatistical interpolation in terms of random heterogeneity at both the scale of an individual sample (i.e., intra-sample variability) and co-located samples (i.e., inter-sample variability). Not surprisingly, the reliability of isoconcentration maps to predict contaminant concentrations in soil at any given point within an area decreases with increasing random, small-scale variability within samples and around individual grid points.

The authors used sets of six, co-located (within one meter) discrete samples to calculate a coefficient of variance (relative standard deviation) for different metals at five locations within a study area. The small-scale variance of metal concentrations in soil was not consistent between locations, similar to observations for each of the study sites in this project (refer to Part 1). A reasonably good correlation of predicted and measured metal concentrations in soil was reported for cases where the coefficient of variance between co-located discrete samples collected at a site was less than approximately 35% (measured soil concentrations within 25th-75th percentiles of predicted concentration). Whether or not this is in fact adequate for risk assessment or site remediation purposes would be site-specific. Measured concentrations of metals for grid points associated with higher coefficients of variance (e.g., >65%) consistently fell outside of the 5th-95th percentiles of concentrations predicted by the kriging method used. Maps generated by discrete sample data from the site were highly unreliable.

The authors note that high, small-scale variability (i.e., high coefficients of variation) and resulting isoconcentration map error are especially a problem for estimation of metal concentrations in clayey soils in comparison to sandy soils. This is in part dependent on the nature of the contaminant release, however. The relatively high, small-scale variability of lead in discrete samples in the relatively sandy soils of Study Site B in this project is interpreted to be due at least in part to presence of randomly scattered, millimeter-scale nuggets of lead-concentrated ash in the soil. As discussed in Part 1 of this study, soil type and particle-size distribution do not appear to be controlling factors in terms of data representativeness.

An expanded approach similar to that used by Aelion et al (2009) was applied to data for the three study sites. Table 6-1 summarizes the Relative Standard Deviation (RSD; coefficients of variance) measured and estimated around individual grid points at each of the study sites. The results are intriguing. The median intra-sample RSDs for Study Sites A, B and C are 12%, 34% and 57%, respectively. The RSDs vary widely between individual grid points, however ranging from 4.8% to 30% at Study Site A (arsenic), 20% to 96% at Study Site B (lead), and 17% to 277% at Study Site C (total PCBs). Median inter-sample RSDs vary within a similar range (see Table 4-2) but are random in terms of comparability with intra-sample RSDs for the same grid

points (Figures 6-1 through 6-3). The potential error associated with any given point for the generation of isoconcentration maps is, as a result, inconsistent and unpredictable.

These observations raise concerns regarding the reliability of isoconcentration maps beyond gross identification of large-scale areas with relatively high versus relatively low contamination. This has significant implications for the use of isoconcentration maps to determine areas of a site for localized, "hot spot" removal. This includes early concepts of "Iterative Truncation" approaches to site remediation that relied on an assumed uniformity of contaminant concentrations around grid points for accuracy (USEPA 2005b; refer to Attachment 1). The collection of an additional set of samples from grid points might generate reliable, large-scale patterns of contaminant distribution if the variability of contaminant concentrations within these areas is consistently above or below a target screening level. Smaller-scale patterns will unavoidably reflect artifacts of random, small-scale variability, however. This is made clear by a more detailed evaluation of the study site data.

6.1 Study Site Contamination Patterns

Figures 6-4, 6-6 and 6-7 present a series of hypothetical, simplistic "isoconcentration" maps generated for each of the study sites based on the estimated range of contaminant concentrations at individual grid points. The random number generator in Excel was used to assign a concentration of the contaminant to each grid point, based on the estimated minimum to maximum range for that point (refer to Tables 4-7, 4-14 and 4-20 in Part 1). Conditional formatting was used to color the grid point cell with respect to example screening levels for the study site. Eight iterations of maps were generated for each site.

The maps are intended to illustrate the collection of eight independent, random replicate sets of discrete sample data from each of the sites within half-a-meter of the original grid points. The maps generated for a study site should be similar if any given individual, discrete sample collected within the grid point area is representative of the grid point area as a whole, as envisioned in early USEPA sampling guidance (see Section 2.3). This of course is not the case in the field.

6.1.1 Study Site A

Figure 6-4 presents a series of hypothetical maps of independently collected sets of discrete samples collected around grid points for Study Site A. The estimated range of arsenic concentrations for discrete sample concentrations collected within 0.5m of each grid point is summarized in Table 4-7 of Part 1. Small-scale variability of arsenic concentrations in soil is relatively low in comparison to Study Sites B and C. Cells are color coded green, yellow, red and purple to indicate grid point concentrations of <200 mg/kg, >200 mg/kg to <400 mg/kg, and >400 mg/kg to <600 mg/kg and >600 mg/kg total arsenic, respectively. Multi Increment sample data for the study site estimate a mean arsenic concentration of 233 mg/kg ("yellow;" mean of 54-increment triplicates, see Part 1 Table 5-5).

Arsenic distribution patterns depicted in the series of maps may at first seem similar, with most of the cells a mix of green or yellow and scattered "hot spots" of red and purple. Closer inspection reveals that these patterns shift between maps, however. Trends between individual grid points (e.g., lower to higher) are likewise random and dependent on the set of discrete sample data points selected. This is a classic signal of random, small-scale noise in the data that would go undetected in the absence of co-located samples for each grid point. The collection of only one or two, co-located samples as recommended in early USEPA guidance would most likely be explained as laboratory error, rather than error in the field (refer to Section .

This phenomenon is further illustrated in Figure 6-5. Like randomly drawing a single card from 24 decks of playing cards, the data for any single point is an artifact of small-scale heterogeneity and cannot be considered to represent the area around the grid point as a whole. Thirteen cards consisting of the Ace through King of spades are hypothetically placed on each of the 24 grid points. The average "concentration" of card numbers for each point and for the "study area" as a whole is "7," with the values 11, 12 and 13 assigned to the Jack, Queen and King, respectively. A single card is drawn at random for each point. The probability of drawing any given card is equal. Map patterns are generated based on assignment of the colors "green" for cards two through six, "yellow" for cards seven through ten, "red" for face cards and "purple" for Aces. The "true mean" color of each cell of the map and the map as a whole is "yellow." All points are identical; no large-scale patterns are present within the grid area.

Eight iterations of this process are depicted in Figure 6-5. Compare the results to the maps generated for Study Site A in Figure 6-4. The patterns generated in the maps are artifacts of small-scale heterogeneity at the scale of an individual sample (i.e., a single card) and not representative of actual "site" conditions. Apparent clusters of low or high cards are not real or reproducible. Attempting to use the "discrete data" to identify and surgically remove "hot spots" in order to reduce "risk" would be misleading, since "data" for the surrounding cells is likewise not representative of those areas. The same is true for declaring some areas of the site to be "clean" relative to a target screening level based on the random cards selected for those areas in Figure 6-5 or random arsenic concentrations depicted in Figure 6-4 for Study Site A. Removal of contamination around fortuitously identified "hot spot" (e.g., red or purple cells) cannot be assumed to significantly reduce risk for the area as a whole.

6.1.2 STUDY SITE B

Figure 6-6 presents a similar set of random, concentration pattern maps generated for lead at Study Site B. Cells are color coded green, yellow, red and purple to indicate grid point concentrations of <200 mg/kg, ≥200 mg/kg to <400 mg/kg, and ≥400 mg/kg to <800 mg/kg and ≥800 mg/kg total lead, respectively. Multi Increment sample data for the study site suggest a mean lead concentration of 287 mg/kg ("yellow;" 54-increment triplicates, see Part 1, Table 5-5). Eight small-scale patterns of lead distribution generated for data from Study Site B again reflect random assignment of a concentration within the minimum and maximum range estimated for each grid point (see Table 4-14 in Part 1).

Variability between maps is greater than observed for arsenic at Study Site A (see Figure 6-4). This is to be expected, given the greater, relative range of lead concentrations for each grid point. The study site is located within a much larger area of lead-contaminated soil and characterized by a heterogeneous mixture of lead-contaminated ash and native soil (refer to Section 2.2 in Part 1). There is no reason from a standpoint of site history to suspect that one area of the study site is more or less heavily contaminated than another.

None of the map patterns depicted in Figure 6-6 can be considered to be representative of actual site conditions. While it is possible that the mean concentration of randomly selected sets of individual points approximates the true mean, the precision of the estimate for any single set of data cannot be estimated in absence of comparison to independent, replicate sets of data. This issue is further explored in Section 7. (Note that "replicate" samples cannot in practice be collected for individual, discrete samples, since the samples only represent the mass of soil actually collected.)

6.1.3 STUDY SITE C

Figure 6-7 presents a set of randomly generated maps for Study Site C based on the estimated range of total PCB concentrations in soil around individual grid points at the scale of a discrete sample. Cells are color coded green, yellow, red and purple to indicate grid point concentrations of <1.1 mg/kg, ≥1.1 mg/kg to <50 mg/kg, and ≥50 mg/kg to <250 mg/kg and ≥250 mg/kg total PCBs, respectively. Multi Increment sample data for the study site suggest a mean lead concentration of 104 mg/kg ("yellow;" 60-increment triplicates, see Part 1, Table 5-5), although the precision of the replicate data is considered to be very poor.

In this case and unlike Study Sites A and B, two distinct "populations" of contaminated soil were known from past investigations to be present within the study area. Soil in the eastern area of the site (upper portion of the patterns illustrated in Figure 6-7) was known to be significantly more contaminated with PCBs than soil in the western area of the site (lower portion of the patterns illustrated in Figure 6-7; see Section 5; refer also to Section 2-3 in Part 1). Map patterns generated by random selection of discrete sample data for each grid point from this study also consistently suggest higher PCB contamination in the eastern area of the site (see Figure 6-7). Smaller-scale patterns within these areas again cannot be assumed to be real and are most defensibly interpreted to be artifacts of random, small-scale variability. Any attempt to draw a boundary between the two areas would necessarily be broad-stroked and require more detailed, followup testing using DU and incremental sampling methods.

The randomness of data for any given grid point in the maps again highlights the limitations of using discrete sample data to identify and surgically remove small "hot spots" from larger areas of contamination (see Section 4). This problem is explored in more detail for PCBs in Section 8.

6.2 HAKALAU PESTICIDE MIXING AREA

The potential presence of artificial "hot spots" and "cold spots" related to random, small-scale variability of contaminant concentrations in soil is readily apparent in most soil isoconcentration maps. Consider, for example, the nine-acre site on the island of Hawai´i depicted in Figure 6-8. The site was formerly used to mix and store arsenic-based herbicides and is now being considered for residential redevelopment. Previous Multi Increment samples identified arsenic-contaminated soil within an area of suspected, past herbicide mixing. Designation of additional DUs to identify the boundaries of contamination in a followup investigation was, however, unclear. Remediation of contaminated soil was also likely to be expensive and time consuming. Optimal DU size and placement was desired to control costs and expedite cleanup.

A decision was made to screen the site using a tight grid of discrete sample points and a portable XRF to help identify large-scale contamination patterns within the site (ERM 2008). Discrete, surface soil samples (approximately 200 grams) were collected at a fifty-foot spacing across the site, with additional samples collected at a twenty five-foot spacing in areas where heavy contamination was initially identified. Samples were dried and hand-mixed prior to testing with a portable XRF. The mass of soil tested by a single, XRF reading was estimated to be less than one gram. Testing of multiple points within a sample suggested that "intra-sample" variability was reasonably low (similar to intra-sample variability observed for Study Site A, located on the same island). Multiple readings were made for each sample and used to estimate a mean arsenic concentration for the sample as a whole.

Figure 6-19 depicts an isoconcentration map generated from the discrete data grid points. A large area of heavy contamination at the northern edge of the site is clearly apparent from the discrete sample data. An assumed, background threshold value of 24 mg/kg was used to screen the site, with red shades in excess of this level (HDOH 2012).

Three large-scale zones of arsenic concentrations in soil are apparent on the map (Figure 6-10; not discussed in original report). The variability of discrete sample data within each zone is depicted in the boxes to the right of the map in the figure (hypothetical, for illustrative purposes). In Zone A, the overwhelming majority of discrete data points fall above the screening level of 24 mg/kg (default upper bound of natural background; HDOH 2011). In Zone B, concentrations of arsenic in discrete samples fall both above and below the action level. In Zone C, the overwhelming majority of discrete data points are consistently below the screening level.

The dashed lines on the map denote areas of the site where the mean concentration of total arsenic over any given 5,000ft^2 DU area could exceed the screening level. The boundaries between these zones are necessarily blurred, given the uncertainty surrounding the mean concentration of arsenic around any single grid point. Whether the mean concentration of arsenic within the middle zone exceeds the target action level is unknown, and can only be determined by more detailed sampling. For example, lot-size, "Perimeter DUs" could be designated within this area and characterized through the collection of Multi Increment samples.

Zone B is best interpreted to reflect the area of the site where the concentration of arsenic in discrete soil samples begins to range both above and below the target screening level. The numerous, seemingly isolated "hot spots" and "cold spots" tens of feet across within this zone generated by the software most reasonably reflect small-scale variability of arsenic concentrations in soil rather than mappable and reproducible spots of higher or lower contamination. As demonstrated by the "inter-sample" variability data presented in Part 1 of the study, if the grid was shifted one or two feet in any direction and discrete samples recollected, then a similar, large-scale pattern of contamination can be expected to appear. Similar, small scale patterns would also appear, but sample-size "hot spots" and "cold spots" would be located in different areas.

Compare Zone B to data for Study Site B, where lead concentrations for discrete samples are estimated to fall both above and below the screening level of 200 mg/kg at 23 of 24 grid points fall (see Figure 6-1 in Part 1). In this case, all of Study Site B is "Zone B," with smaller-scale but still mappable "hot areas" within the site entirely absent.

6.3 Background Metals in US Soils

Once recognized for what they are, these same, artificial and random patterns that reflect small-scale variability are readily apparent on other isoconcentration maps. Excellent examples of random, small-scale variability are depicted in isoconcentration maps of background metals in soil recently published by the U.S. Geological Survey (USGS 2014). The maps were generated based on the collection of several thousand, composite samples from one-meter square points across the country. Samples were ground and subsampled for testing. Data for any given sample can be considered to be reasonably accurate.

The isoconcentration map for arsenic concentrations in surface soil is presented in Figure 6-11. Compare the large- and small-scale patterns of apparent arsenic distribution to those in Figure 6-9 for the nine-acre site in Hawai´i. Large-scale patterns are again most likely real and appear to correspond to well-studied, geologic terranes (Figure 6-12; USGS 2004). A detailed evaluation of the distribution of arsenic in soil in terms of geology has not been carried out. Clear correlations are not necessarily evident in some areas. Elevated levels of arsenic in the upper, Mississippi flood plain could, for example, simply reflect deposition of fine sediment from more arsenic-rich, metamorphic and igneous geologic terranes in the upper watershed of the river.

Scattered and seemingly isolated "hot spots" and "cold spots" within and along the boundaries of larger-scale areas are most defensibly interpreted to reflect random, small-scale variability within larger-scale patterns. For example note the numerous red (hot) spots and yellow (cool) spots scattered through the map. On closer review these spots are based on data for one or two, one meter-square sample extrapolated by the mapping software to an area that covers hundreds or even thousands of square kilometers. Figure 6-13, for example, depicts what is in all likelihood an artificial, 2,500km^2 arsenic "hot spot" based primarily on data for a single sample collected from a one-meter square area of soil. The geology of the area is characterized by highly

metamorphosed and structurally complex metamorphic rock that includes narrow, highly mineralized zones (Cat Square or Newton Window geologic terrane; Merschat and Hatcher 2007). It is highly possible that the sample was collected from soil developed on a local, mineralized zone.

The USGS was well aware of this problem and cautions users against over-interpretation of the maps. As stated in the USGS report (USGS 2014):

> The resulting data sets are not appropriate for the accurate estimation of the concentration of a given element or mineral at a site where a sample was not collected… The data isn't so fine that it will tell you what lies in your backyard...

A decision was made by the authors of the report, however, to intentionally use a high power function to generate the background metal maps in order to illustrate the magnitude of random, small-scale variability within the larger-scale patterns. Figure 6-14, for example, suggests that the background concentration of arsenic in the eastern one third of Oklahoma is somewhat higher than in the western area of the state. The small-scale patterns within the larger-scale areas are again not "real" in the sense that they are unlikely to be reproducible if a sample were collected in the same general area. The data suggest, however, that background concentrations of arsenic in soil in the eastern area of Oklahoma ranges between 5 mg/kg and 15 mg/kg at the scale of a one meter-square soil sample, while the background concentration in the western area of the state ranges between 0.5 mg/kg and 5 mg/kg. The collection and testing of composite or incremental samples from larger areas, for example one square- or even one hundred-square kilometer would most likely result in a narrower range of arsenic concentrations between sample points and be more reflective of mappable trends within and between large-scale patterns.

Samples collected over a larger area can be expected to reflect a mean of these concentrations. A hypothetical map of arsenic distribution in soil across the US with random, small-scale noise in the USGS data removed is depicted in Figure 6-15. The figure is for illustration purposes only. A thorough comparison of geologic provinces to the USGS soil data was not carried out. Each area could be considered to represent a "Decision Unit," with color indicating the mean concentration of arsenic in soil if all of the soil within the province could be collected and tested as a single sample. Lines between provinces are necessarily dashed. On the ground, the dashed lines might be tens of miles wide to denote the uncertainty in placement of the boundaries.

The accuracy of smaller patterns depicted by the USGS data decreases as the sample support for a given area decreases. A single, one-meter square sample point almost certainly cannot be used to extrapolate arsenic concentrations over an area larger than the sample area itself. Trends between individual data points cannot be assumed to be linear. The number of data points required to accurately estimate the mean concentration of arsenic (or any other metal) depends on the magnitude of small-scale variability within that region. Potential error associated with short-scale, inter-sample variability not related to large-scale, mappable patterns will remain

largely unknown in the absence of larger-scale, detailed sampling (e.g., incremental samples collected over square miles of territory). How this problem can be addressed in software used to generate isoconcentration maps with discrete sample data in general is uncertain but is a topic worthy of further research. The key is to start "big" and work towards a progressively smaller and more detailed resolution as technically defensible by available data and warranted the objectives of the investigation.

6.4 ISOCONCENTRATION MAP POWER FUNCTIONS

Isoconcentration mapping programs typically utilize an "inverse distance weighting (IDW)" method to interpolate concentrations between data points and generate contours (see Lu and Wong 2008). An important parameter in the IDW method is the "Power Function" employed to generate the maps. Geostatistical models reflected by higher power function values assign greater influence to data points closest to the interpolated area (e.g., up to a Power Function of 16 in Groundswell software). This results in isoconcentration maps with tighter and wider ranging contours. In theory this provides greater detail. A Power Function of 5 is most commonly used for isoconcentration maps. This is assumed to be the case for the example maps presented in Sections 6.2 and 6.3. The use of a Power Function of 5 is based in part on "professional judgment" although in the case of soil replicate sets of data are rarely collected to test the accuracy of the maps. Perhaps it is most correct to suggest that a Power Function of 5 generates the scale of resolution that the investigator hopes to achieve, whether or not this has been accomplished in reality.

Maps based on high power functions are subject to over-interpretation of individual data points, with each point turned into a seemingly isolated "hot spot" or "cold spot." Lower power function values assign less influence to nearby data points and greater consideration of more distant data points. This generates isoconcentration maps with more widely spaced contours and fewer and less dramatic, isolated hot spots and cold spots. Lower power functions are used for less-than-optimal data sets that could be biased due to small-scale variability unrelated to larger-scale trends of interest.

Minimizing the power function used to generate an isoconcentration map can help reduce but not completely eliminate error associated with random, small-scale variability and noise in discrete sample data sets. This is exemplified in a series of isoconcentration maps for Study Site A based on separate groupings of data for the "A," "B," "C," "D" and "E" processed, discrete samples for each of the twenty-four grid points , as depicted in Figure 6-16. The maps were generated using software developed by Groundswell Technologies (Groundswell Technologies 2013).

A Power Function of 5 was used to generate the isoconcentration maps in Figure 6-17. This is typical for isoconcentration maps for contaminated soil. The center map depicts isoconcentration contours based on use of the "Sample A" data set for each grid point. The upper left-hand, upper right-and, lower left-hand and lower right-hand maps depict isoconcentration contours based on use of the "Sample B," "Sample C," "Sample D" and

"Sample E" data sets, respectively, and similar to the pattern of sample collection in the field (see Figure 6-16; see also Figure 5-2 in Part 1). Note the changing locations of "hot spots" and "cold spots" within the study area depending on which data set is used to generate the map. This is again a classic sign of noise in the data due to small-scale heterogeneity. The individual spots are not real in the sense that they represent actual map patterns. They instead reflect small-scale variability inherent to the soil in the study area as a whole. The variability between processed, 200 gram discrete soil samples is real but the map patterns generated from the data are not.

A similar pattern of seemingly isolated hot spots and cold spots is generated using all five data points for each grid point (Figure 6-18, IDW distance decay parameter of 5 used). These small-scale patterns are, again, best interpreted as artifacts of the computer program and again not real. Testing of additional discrete soil samples within the patterns is unlikely to reliably reflect a concentration within the enclosing contours. A shift of the grid points several feet in any direction and the collection of a new set of samples would likely produce similar, overall patterns but in different locations (see Section 6).

The problem is less acute when a lower distance decay parameter that places less emphasis on individual grid points is used. A distance decay parameter of 1 was used to generate the arsenic isoconcentration map in Figure 6-19. The software is still unable to fully overcome the small-scale variability of arsenic concentrations around and between individual grid points, however, and presumed artificial small-scale patterns are still generated on the map. In general, the map utilizing the lowest distance decay parameter depicts what is likely to be a more accurate picture of mappable, larger-scale variability of mean contaminant concentrations within the study area, with the slightly higher concentration of arsenic in the upper third of the map likely to be "real" (Figure 6-19). The use of alternative software programs has not been evaluated in detail.

7 USE OF DISCRETE SAMPLE DATA IN RISK ASSESSMENTS

Estimation of the mean contaminant concentration for a targeted area and volume of soil is a key step in a human health or ecological risk assessment (USEPA 1987, 1988, 1989a,b, 1991, 1992a, 2011; see also USEPA 2014a). The USEPA's *Risk Assessment Guidance for Superfund* (RAGS), written in 1989 soon after implementation of a host of new environmental laws and regulations, still serves as the primary guidance for human health risk assessments (USEPA 1989b,c). The accuracy of the estimated mean in terms of bias and precision is a function of multiple factors, including (see Pitard 1993, 2009; Minette 2007): 1) The representativeness of the sample(s) in terms of the targeted area and volume of soil from which it was collected, 2) The representativeness of the subsamples removed for analysis and 3) The representativeness of data generated by the laboratory analytical method in terms of the subsample mass tested.

Quantitative evaluation of the representativeness or "precision" of discrete sample data, often referred to as "data validation," most commonly focuses just on the performance of the laboratory analysis factors, noting if the lab analysis precision does or does not fall within a specified "quality assurance" range (e.g. < 20%). In addition, the site-specific lab data results are not reported as a range of values based on the lab analysis precision determinations, so this range would need to be calculated from the reported lab quality assurance data in order to consider it when comparing site data to relevant contaminant action levels and for final site determinations. Although reported lab quality assurance data may include "batch" subsampling duplicates for discrete samples (e.g. 1 per 10 or 20 samples), it is generally not clear if adequate sub-sampling masses are routinely collected for digestion and/or analysis based on the maximum particle sizes in the bulk discrete samples, or if representative sub-sampling methods are utilized. Quantitative evaluation of true, field replicate sets of discrete samples are rarely if ever carried out (co-located discrete samples would typically not be considered field replicates). Replicate sets of samples would need to be collected from separate systematic or stratified random locations in the same designated "decision unit" being evaluated in an unbiased manner and in accordance with sampling theory.

This section focuses on the reliability of a single set of discrete soil samples to estimate mean contamination concentrations in terms of field representativeness. One of the most common rationales for the use of discrete soil sample data over incremental sample data in soil investigations is the need to determine the small-scale variability of contaminant concentrations within an exposure area and better assess the precision of the estimated mean (USEPA 1987). The measured variance can also be used to calculate a more conservative estimate of a mean exposure point concentration for a specified confidence level (e.g., 95% Upper Confidence Level or "UCL"). A set of discrete samples is considered usable when the coefficient of variance for the data set is relatively low, usually defined as <20-50% (e.g., USEPA 1991). The number of samples required to represent a targeted area can in theory be calculated based on a target a coefficient of variation, a required confidence level or certainty, a required statistical power, and

a “Minimum Relative Detectable Difference” (USPA 1991, 2013). In practice these approaches are rarely used due to the large numbers of samples required by the resulting calculations.

The implied precision of these statistical approaches in terms of field representativeness can be misleading, however. A low coefficient of variance does not necessarily imply that the discrete sample data set is representative of the targeted area. As discussed in this section, statistical evaluation of a single discrete data set only assesses the precision of the estimated mean *in terms of the data set provided and the statistical method employed*. The precision of the data set in terms of representing the true mean of the targeted area is unknown. The *field precision* of a given set of discrete sample data can only be evaluated by comparison to the mean concentrations estimated for completely independent, replicate set of discrete samples that were collected in exactly the same manner as the original sample set (USEPA 1987). The collection of replicates to assess data precision in terms of field representativeness is required under incremental sampling methodologies (HDOH 2008; ITRC 2012). Assessment of field representativeness is rarely if ever undertaken for discrete sampling methodologies, however, due to the time and cost involved. This is also due in part to a general misunderstanding by nonstatisticians about the capability (“performance”) of geostatistical methods to account for inadequacies in discrete sample data sets. This limitation is highlighted by comparisons of random sets of discrete sample data for the three study sites evaluated in Part 1 of this report.

7.1 Accuracy, Bias and Precision

The “accuracy” of an estimated mean concentration of a contaminant within a designated area and volume of soil is best described in terms of “bias” and “precision” (USEPA 1989, 1992; Pitard 1993; see also ITRC 2012). “Accuracy” refers to the correctness of an estimated value in terms of the true concentration. In the case of soil testing the true concentration is not known and the accuracy of an estimated exposure area concentration likewise cannot be directly determined. The terms “bias” and “precision” are instead applied to assess the likely representativeness of the data.

7.1.1 Bias

Examples of “precision” and “bias” are presented in Figure 7-1. In order for an estimated mean to be accurate the data set must be both unbiased and precise. Bias occurs when the sampling method employed is not representative of the targeted media, resulting in a systematic over or under estimation of the mean (see Figure 7-1). Bias can be controlled but not eliminated. Examples include the collection of an unrepresentative proportion of discrete samples from either clean or contaminated areas within a larger-scale exposure area, or collecting samples in a manner that biases the mass of soil to a specific depth interval rather than equally representing the targeted zone (e.g., wedge versus core-shaped sample from upper six inches of soil; refer to HDOH 2008; ITRC 2012). Samples (or increments) collected from different parts of the targeted area (DU) must be of similar size, orientation and mass in order to avoid bias. Samples must then be processed at the laboratory in a manner that produces a representative subsample

for analysis. In practice, a systematic approach to obtaining a representative subsample is often neglected for discrete samples (see Section 8.1). Discrete samples are oftentimes collected in a haphazard manner in the field, and rarely adequately processed and subsampled for analysis.

7.1.2 PRECISION

Precisionis a measure of the reproducibility of data for a given sample point or for a given set of samples, evaluated in terms of variability and uncertainty (see Figure 7-1). This includes the inherent heterogeneity of contaminant distribution in soil as well as differences in exposure parameter values incorporated into the risk assessment. Natural variability cannot be reduced, only better understood and captured as part of representative sampling. This improves the precision of the mean estimated for the data.

The precision of laboratory analytical methods is well-studied and controlled though Standard Operating Protocols (SOPs) published by the USEPA and other entities for different suites of chemicals (USEPA 1998b). The performance of the extraction process and the analytical equipment is evaluated, for example, through repeated testing of samples of known concentration. A precision of +/- 35% (or better) is generally assumed to be adequate and within the capabilities of current analytical methods, as well as acceptable for risk-based decision making (see ITRC 2012).

The precision of laboratory subsampling protocols is, in contrast, often poorly assessed or controlled at the laboratory, with many analytical method SOPs simply calling for non-specific "homogenization", and the digestion/analytical mass selected not tied to the maximum particle sizes of the bulk samples received. Differences in true subsample replicate data, when obtained, are largely ignored or the maximum value referred to for decision making, with the variability optimistically assumed to reflect laboratory analytical error rather than bias and error in the process as a whole (refer to Section 8; see also ITRC 2012). The potential for decision error based on data for inadequately processed and subsampled discrete soil samples is high, as illustrated by "intra-sample" variability data discussed in Part 1 of this study. Labs might also use matrix spikes to test analytical precision rather than testing of multiple subsamples from a single sample. This can further mask problems with data representativeness and introduce significant but hidden uncertainty in decisions based on the data.

7.1.3 STATISTICAL PRECISION AND FIELD REPRESENTATIVENESS

The representativeness of a single set of discrete soil samples for a targeted area cannot be assessed in terms of geostatistical analysis alone. A suggestion that it can could be misleadingly drawn from language in the USEPA document *Supplemental Guidance to RAGS: Calculating the Concentration Term* (USEPA 1992b):

> The key to any statistical sampling plan is the use of the variation within the sample set to test hypotheses about the population and to determine the precision or reliability of the data set.

This statement only applies to the precision of the statistical method employed to estimate a mean contaminant concentration for a specific set of discrete sample data provided. No direct information is provided regarding the precision of the mean in terms of *field reproducibility*. This can be a very important drawback for the use of discrete samples to estimate mean contaminant concentrations in soil.

Reconsider, for example, the classic examples of "bias" and "precision" in Figure 7-1. The hypothetical targets include multiple data points. The tightness of these data points, measured in terms of variance, defines the precision of the data for the parameter of interest. In this case the soil the parameter of interest is the mean. Only a single mean is estimated from a single set of discrete samples, however.

A more appropriate illustration of precision for a single discrete mean value is presented in Figure 7-2, with a single point noted. In theory bias could be controlled in the collection and testing of a discrete sample data set in the same way it is controlled under incremental sampling (e.g., number, location, size, mass, processing and subsampling, etc.). For a discrete sample set, each individual increment would be tested and the mean calculated. In practice, of course, this is rarely if ever done. Assuming, however, bias iscontrolled, the mean estimated from a single set of discrete samples would be represented by one of the two "unbiased" targets on the left of the figure.

An estimate of the precision of the estimated mean based on geostatistical analysis of the single discrete data set would most appropriately be illustrated by the size of the (mean) point, with a larger-size point reflecting decreased precision in the measurement. The surrounding gray area represents additional uncertainty in terms of the reproducibility of the data set and the estimated mean. It is not possible, with the single data point, to directly estimate the full magnitude of this uncertainty, as represented by the question marks in Figure 7-2. As briefly discussed in the following section, assumptions can be made regarding the anticipated precision of the data set in terms of the number of samples collected and past studies of reproducibility. Poor precision for replicate sets of incremental sample data that represent what would be considered a very large number of sample (increment) locations within a targeted area can illustrate the unreliability of such assumptions on a site-specific basis, however.

The precision of the mean estimated from a single set of discrete samples (or a single incremental sample) can only be tested by comparison of replicate sets of independently collected samples (USEPA 1987, 1990, 1992; see also HDOH 2008; ITRC 2012). If a significant difference between the estimated mean is identified, then the most likely cause is imprecise field sampling resulting from site-specific contaminant heterogeneity (see Pitard 1993, 2005, 2009; Minnitt et al 2007; ITRC 2012). The collection of replicate samples is a required feature of incremental sampling methodologies (see HDOH 2008; ITRC 2012). This provides a quality control test of the mean and a method to quantitatively estimate the true precision of the data. This is illustrated in Figure 7-3, with each point representing the mean generated for a

single incremental sample. The outer boundary of the gray area of uncertainty can now be estimated, for example based on the Relative Standard Deviation or the 95% Upper Confidence Level for the data set.

The need to collect and evaluate replicate sets of discrete samples to assess precision in terms of field representativeness is discussed in early USEPA guidance (e.g., USEPA 1987). Generic recommendations for the collection of co-located or "replicate" discrete samples at an interval of one per twenty samples are far too inadequate to assess the representativeness of an individual data set, however (see Section 5.1). These recommendations were most likely based on batch tests of the precision of analytical methods and equipment calibration, rather than sample processing or more importantly field error. This problem is recognized in the USEPA document *A Rationale for the Assessment of Errors in the Sampling of Soils* (USEPA 1990b):

> Previous EPA guidance for the number of quality assessment samples has been one for every 20 field samples (e.g., USEPA 1987). However, such rules of thumb are oversimplifications and should be treated with great caution… The number of field duplicates to be obtained in the study should be dictated by how precise one wants that estimate of the total measurement variance to be… The number of samples required to detect random bias will depend on the distribution of the biasing errors, and this distribution will generally be unknown… Unique characteristics of a particular site may require an increased number of quality assessment samples to measure performance against stated data quality objectives… The importance of pilot studies to the overall monitoring effort cannot be stressed enough.

The "random bias" that the guidance document warns against is the random, small-scale variability highlighted in Part 1 of this study. The 1990 USEPA document unfortunately does not delve into this issue in detail. Had the authors collected and tested multiple, completely independent sets of discrete samples as part of the study then the ubiquity of random variability at the mass of the samples being collected and associated problems with soil sampling methods being developed at the time would have become immediately apparent.

The lack of control for bias in the collection, processing and testing of discrete soil samples combined with the lack of a test of the reproducibility of the data would ensure that a mean estimated by incremental sample replicates will always be superior in quality (less biased and better precision) to a mean estimated from a single set of discrete samples. The mean estimated by a single, incremental sample is more reliable given systematic efforts to control bias and (in most cases) inclusion of a larger number of increment locations within the decision unit as part of the bulk sample collected.

7.2 Estimate of Mean Exposure Area Concentrations

Statistical evaluation of discrete sample data sets is discussed in a series of USEPA guidance documents prepared in the late 1980s and early 1990s, including the following:

- *Data Quality Objectives for Remedial Response Activities* (USEPA 1987);
- *Superfund Exposure Assessment Manual* (USEPA 1988);
- *Methods for Evaluating the Attainment of Cleanup Standards, Volume 1: Soils and Solid Media* (USEPA 1989a);
- *Risk Assessment Guidance for Superfund, Volume I, Human Health Evaluation Manual, Part A* (USEPA 1989b);
- *Risk Assessment Guidance for Superfund, Volume II, Environmental Evaluation Manual* (USEPA 1989c);
- *Guidance for Data Usability in Risk Assessment Part A* (USEPA 1991);
- *Supplemental Guidance to RAGS: Calculating the Concentration Term* (USEPA 1992b).

These and related documents recommend that a 95% Upper Confidence Level on the arithmetic mean be used as the exposure area concentration in a risk assessment. As discussed in the *Data Quality Objectives* document (USEPA 1987):

> At sites where contamination is known to exist, a parameter of interest is the mean contaminant concentration over the contaminated area. Confidence limits can, theoretically, be placed on any quantity calculated from a data set. When the sample mean is calculated from a set of data, it is unlikely that the actual or population mean will equal the sample mean. The sample mean for a fixed number of data is a random variable whose value will fluctuate depending on the specific data collected. Confidence intervals are a method of quantifying the likely range of fluctuation of the sample mean. Confidence intervals are defined as follows: if the *95* percent confidence interval is set for the sample mean after each repetition of an experiment and the experiment is performed 100 times, the population mean is expected to fall between confidence limits *95* times.

The rationale behind use of a 95% UCL of the mean rather than an arithmetic mean is further expanded in the USEPA *Supplemental Guidance to RAGS: Calculating the Concentration Term* (USEPA 1992b) document:

> For Superfund assessments, the concentration term (C) in the intake equation is an estimate of the arithmetic average concentration for a contaminant based on a set of site sampling results… While an individual may not actually exhibit a truly random pattern of movement across an exposure area, the assumption of equal time spent in different parts of the area is a simple but reasonable approach… Because of the uncertainty associated with estimating the true average concentration at a site, the 95 percent upper confidence limit (UCL) of the arithmetic mean should be used for this variable… ”

The document suggests that a minimum of 20 to 30 discrete soil samples is required to reliably estimate the mean concentration of a contaminant in soil for a targeted area (USEPA 1992b):

> Sampling data from Superfund sites have shown that… data sets with 20 to 30 samples provide fairly consistent estimates of the mean (i.e., there is a small difference between the sample mean and the 95 percent UCL)…

A reference for this conclusion is not provided, but appears to be related to an evaluation of coefficients of variation for data collected at Superfund sites included in the document *Guidance for Data Useability in Risk Assessment* (USEPA 1991; refer to Exhibit 23 in document). The number of samples included in the data sets reviewed are not provided in this document, however.

Based on data collected as part of the HEER field study (see Part 1), small-scale variability at the Superfund sites evaluated was low to moderate and similar to Study Sites A and B for arsenic and lead. Multi Increment data at Study Sites B and C, as well as other similar sites reviewed by the HEER office, clearly indicate more than 30 samples (or increments) are required to adequately capture and represent mean contaminant concentrations with moderate to high small- and large-scale spatial variability. Note also that decisions regarding "clean boundaries" around contaminated soil are often made on a very small number of discrete samples, even a single sample. Although an integral part of incremental sampling methodologies, the collection and statistical evaluation of large sets of discrete samples (or increments) for confirmation of clean "DU" areas bordering known contamination is rarely if ever required by regulatory agencies.

The importance of the collection of replicate sets of data to test the precision of the original data set is only briefly hinted to in early sampling guidance documents (USEPA 1992b):

> There is a tendency on the part of many investigators who sample soil to believe that grab, purposive, biased, or judgmental sampling is all that is needed to arrive at a decision about a particular site that is under investigation. Without the input of some form of statistical control, there is no means of determining the reliability of the data or of making a valid decision about the action needed at the site.

The document further discusses apparent conflicts between the maximum-reported concentration in the discrete soil sample set and the estimated 95% UCL (USEPA 1992b):

> The UCL can be greater than the highest measured…concentration. In these cases, if additional data cannot practicably be obtained, the highest measured value could be used as the concentration term…"

This would not be acceptable under HEER guidance if discrete samples were used to characterize an area, since the maximum concentration cannot be assumed to be representative of the mean. This is noted but the USEPA in the same document but not addressed (USEPA 1992b):

> Note, however, that the true mean still may be higher than this maximum value (i.e., the 95 percent UCL indicates a higher mean is possible), especially if the most contaminated portion of the site has not been sampled.

Geostatistical analysis of environmental data requires that the data be representative of the targeted population. Potential problems with the representativeness of discrete sample sets is further recognized, but not fully explored, in more recent USEPA guidance (USEPA 2005b):

> It is important to note that geostatistical techniques are not a substitute for collecting sample data; the reliability of the results depends on adequate sampling data… *Extrapolating the results of a small number of samples to a large area can be misleading unless the contaminant distribution across the large area is uniform.* Clearly, for areas with heterogeneous distribution of contamination (e.g., scattered or dumped), the more extensive the sampling data the more representative they will be of the exposure concentration…uncertainty associated with sampling error can be very large, particularly at sites where there is significant spatial heterogeneity in contaminant concentrations.

Well-thought-out protocols for the collection of representative samples have been an essential part of the mining and agriculture industry for decades (Pitard 1993). Failure to collect representative samples and make accurate decisions for subsequent actions is readily apparent in the form of failed mining operations (e.g., overestimation of reserves present) or failed crops (e.g., poorly optimized use of expensive fertilizers), to the financial detriment of the industry and the material detriment of the end users. The pressure to collect science-based, reproducible samples in order to make accurate and defensible decisions is intense.

The same has not been the case for the environmental industry in the past. The majority of the USEPA risk assessment documents noted above were written before an adequate appreciation of the small-scale, distributional heterogeneity of contaminants in complex, soil matrices had been gained. Understanding the precision of estimated mean contaminant concentrations in terms of the representativeness of the field data set is important but at this point in time was largely overlooked as a source of potential error in environmental investigations. These issues are highlighted by a detailed review of the study site data.

7.3 Field Precision of Estimated Means for Study Sites

The precision of random sets of discrete sample data to estimate mean contaminant concentrations for targeted areas can be tested by a closer evaluation of discrete sample data sets from the study areas presented in Part 1 of this report. Three approaches are taken. The USEPA ProUCL software is used to generate a 95% UCL of the arithmetic mean for data set (USEPA 2013).

The first approach compares estimated mean contaminant concentrations for completely random groupings of ten grid points for each study site, with one of ten intra-sample data points randomly assigned to each grid point. Twenty iterations are carried out for each set of grid

points. This illustrates the precision of (non-stratified) random sets of discrete soil samples to represent a targeted area of soil for comparison to the stratified, random grid designs utilized in the other two approaches.

In the second approach two sets of twelve grid points are generated for each study site, with grid points assigned in different stratified random locations for each set. One of ten intra-sample data points is randomly selected and assigned to each grid point. Twenty iterations are carried out for each set of grid points. The range of estimated means is compared within each set of grid points and between each set of grid points. Mean contaminant concentrations are then compared to Multi Increment data for the site to further evaluate the likely field precision of any one set of random grid points. Multi Increment data are considered to be most representative of the study sites, due to both the mass of soil collected and the number of points represented by the samples (refer to Section 5.4 and Table 5-6 in Part 1).

The third approach compares the variability of mean contaminant concentrations predicted for a given study area based on random assignment of one of ten intra-sample data points for each of 24 stratified random grid points (refer to Section 5.1 in Part 1). Twenty iterations of this approach are again carried out and evaluated. The results are used to evaluate the overall field precision for the random collection of a discrete soil sample from a pre-established set of grid points. The range of mean contaminant concentrations calculated is compared to Multi Increment sample data for the same site to evaluate the likely representativeness of the twenty-four grid points to the true, although still unknown mean for the site as a whole.

7.3.1 Precision of Random, Ten-Point Data Sets

Geostatistical analyses of random groupings of ten, discrete data points for Study Site A (arsenic), Study Site B (lead) and Study Site C (total PCBs) are presented below. Refer to Part 1 of this study for summaries of intra-sample data collected for individual grid points at each site. The results are used to evaluate the potential *field precision* of the calculated mean for any given, ten-point set of discrete samples collected from the 24 grid points used for the studies. The precision of the estimated means is further evaluated by comparison to Multi Increment data collected for the study site.

Study Site A

Intra-sample arsenic data for Study Site A are provided in Table 4-1 of Part 1. Grid point locations are depicted in Figure 2-4 in Part 1 (see also Figure 6-4 in this report). Table 7-1 summarizes the range of mean contaminant concentrations predicted for the area based on each of 20 random groupings of ten grid points and grid point data.

The results correspond well with the low, small-scale variability identified for Study Site A in Part 1 of the study report (refer to Table 4-7 and Table 5-1 in Part 1). Calculated 95% UCL arsenic concentrations for random, ten-point groupings of discrete sample data range from 403 mg/kg to 776 mg/kg, with a median of 463 mg/kg and a mean of 481 mg/kg. The ProUCL

software recommended use of the Student's-t UCL for all data groups. The relative standard deviation for the groupings ranges from 34% to 67%, with a median of 46% and a mean of 48%. This suggests a fair to poor precision in the mean concentration of arsenic estimated for any given set of discrete samples with respect to the statistical model employed.

The range of estimated means and 95% UCLs for the set of grid point groupings as a whole is moderately narrow (see Table 7-1). This suggests that no distinct, large-scale patterns of elevated arsenic concentrations are present within the study site, since random groupings of data yield very similar means. Calculated, average arsenic concentration range from 316 mg/kg to 512 mg/kg, with a median of 370 mg/kg and a mean of 372 mg/kg and a Relative Standard Deviation of 12%. The relative tightness of the median and mean values suggests a near normal distribution of calculated means. Downward adjustment of the XRF-based, intra-sample data by 31% to estimate equivalent 6010B data yields a range of mean arsenic concentrations for the ten-point data sets of 218 mg/kg to 353 mg/kg, with a mean of 255 mg/kg (refer to Section 4.1 in Part 1).

The RSD of calculated means for the 20 iterations of random groupings as a whole implies that the overall *field precision* of a mean arsenic concentration estimated from any given, ten-point set of discrete samples is in fact potentially strong. This is further supported by comparison of the adjusted, estimated mean for the random groupings to a mean arsenic concentration of 233 mg/kg estimated for the study area based on Multi Increment sample replicate data (refer to Table 5-5 in Part 1; 95% UCL 259 mg/kg).

Study Site B

Intra-sample lead data for Study Site B are provided in Table 4-9 of Part 1. Grid point locations are depicted in Figure 2-7 in Part 1 (see also Figure 6-6 in this report). Table 7-4 summarizes the range of mean contaminant concentrations predicted for the area based on each of 20 random groupings of ten grid points and grid point data.

The results correspond well with the moderate, small-scale variability identified for Study Site B in Part 1 of the study report (refer to Table 4-14 and Table 5-1 in Part 1). Calculated 95% UCL lead concentrations range for the random groupings of grid points range from 201 mg/kg to 439 mg/kg, with a median of 345 mg/kg and a mean of 343 mg/kg. The ProUCL software again recommended use of the Student's-t UCL for all data groups, even though the range of relative standard deviations was higher than for Study Site A.

The variability of estimated means and 95% UCLs for lead is somewhat higher than calculated for random groupings of grid points for Study Site A. The relative standard deviations for individual grid point groupings display an increased range of 20% to 86%, with a median of 63% and a mean of 61%. This reflects the higher, small-scale variability identified for this site and overall poor precision of estimated mean lead concentrations for random, ten-point groupings of discrete samples with respect to the statistical model employed.

The average lead concentration calculated for the 20 sets of random, ten-point groupings ranges from 159 mg/kg to 333 mg/kg, with a median of 248 mg/kg and a mean of 249 mg/kg and a Relative Standard Deviation of 19% (Table 7-4). The relative tightness of the median and mean values for the 20 iterations of grid point groupings suggests a near normal distribution of calculated means and again suggests that the large-scale distribution of lead within the study area is relatively even. Upward adjustment of the XRF-based, intra-sample data by 6.8% to estimate equivalent 6010B data yields a range of mean lead concentrations for the ten-point data sets of 170 mg/kg to 356 mg/kg, with a mean of 266 mg/kg (refer to Section 4.2 in Part 1).

The RSD of calculated means for the 20 iterations as a whole implies that the overall *field precision* of a mean lead concentration estimated from any given, ten-point set of discrete samples is again potentially good. This is further supported by comparison of the adjusted, estimated mean for the random groupings to a mean lead concentration of 287 mg/kg estimated for the study area based on Multi Increment sample replicate data (refer to Table 5-5 in Part 1; 95% UCL 383 mg/kg).

Note that the calculated range of estimated, mean lead concentrations in the soil again spans both below and above the HDOH action level of 200 mg/kg. Recall that the estimated total range of lead concentrations for discrete samples around a grid point spanned both below and above this screening level for 23 of the 24 grid points (refer to Section 5.3 and Figure 6-1 in Part 1). The calculated mean for twenty-percent of the discrete sample groupings fell below this level (4 of 20). This suggests a moderately high rate of potential *decision error* for random, ten-point sets of discrete samples collected from the site, even though the field precision of the data might be considered reasonably good. This is true even if 95% UCL concentrations of lead are used for some of the data sets. The minimum calculated 95% UCL of 201 mg/kg could lead to a false decision that the site was clean. If the higher, USEPA screening level for lead in residential soil of 400 mg/kg was applied (USEPA 2014) then several of the random data sets could lead to the false conclusion that the site was contaminated.

Study Site C

Intra-sample arsenic data for Study Site C are provided in Table 4-16 of Part 1. Grid point locations are depicted in Figure 2-10 in Part 1 (see also Figure 6-7 in this report). Table 7-3 summarizes the range of mean contaminant concentrations predicted for the area based on each of 20 random groupings of ten grid points and grid point data.

The results reflect the high, combined small- and large-scale variability of total PCB concentrations in the soil identified for the study area in Part 1 (refer to Table 4-20 and Table 5-1 in Part 1). The variability of estimated means and 95% UCLs for random groupings of ten data points within Study Site C is significantly higher than that calculated for Study Sites A and B. Calculated 95% UCL PCB concentrations range, rather spectacularly, from 9.4 mg/kg to over 1,000,000 mg/kg, with a median of 730 mg/kg and a mean of 52,522 mg/kg. The ProUCL software recommended use of the Student's-t UCL for only one grouping of data. The Adjusted

Gamma UCL method was recommended for 13 of the groupings, with the remaining six groupings split between use of the Hall's Bootstrap UCL and the Chebyshev UCL.

The relative standard deviations for the data point groupings are similarly high, with a range of 124% to 315%, a median of 216% and a mean of 221%. This suggests a consistently very poor precision in the mean concentration of PCBs estimated for any given, ten-point set of discrete samples with respect to the statistical models employed.

The average PCB concentration calculated for the 20 sets of random grid point groupings ranges from 5.5 mg/kg to 1,025 mg/kg, with a median of 134 mg/kg and a mean of 313 mg/kg and a Relative Standard Deviation of 116% (Table 7-3). This suggests that several, distinct and mappable populations of PCB-contaminated soil could be present at the site, a fact supported by both discrete sample data and field observations (refer to Section 2.3 in Part 1). Unlike the results for Study Areas A and B, the RSD for the 20 data sets generated for Study Site C show that the *field precision* of mean total PCB concentrations estimated for random, ten-point groupings of data is very poor. The mean of the 20 data sets (313 mg/kg) is higher than the mean of 104 mg/kg for triplicate, Multi Increment samples collected from the study area but similar to the 95% UCL of 346 mg/kg (refer to Table 5-5 in Part 1).

Adequacy of 10-point Discrete Sample Data Sets

The results of the geostatistical evaluation suggest that single sets of ten randomly located, discrete samples are not reliable for estimation of exposure area concentrations at any of the study sites when both data precision and target screening levels are taken into consideration. Error increases with increasing small-scale variability, with estimates of 95% UCLs for the mean PCB concentration at Study Site C especially unreliable.

Are 20 to 30, discrete sample points as suggested in early USEPA guidance documents (USEPA 1992b), representing testing of a few tens to at most few hundred grams of soil, truly sufficient to represent any given targeted area? Evaluation of random combinations of data for the full set of 24 grid points at each study site in conjunction with Multi Increment replicate data sheds some light on this topic.

7.3.2 Precision of Random, Twelve-Point Data Sets

The study site design allows for two sets of 12-point systematic random grid points to be evaluated separate from the original 24-point data set (Figure 7-4). The data simulate shifting of a systematic random, twelve-point grid across the site. Twenty iterations of random assignment of intra-sample data to each grid point were again carried out.

Tables 7-4a and 7-4b and Tables 7-5a and 7-5b summarize geostatistical analyses of random groupings of intra-sample discrete data for Study Sites A and B. The range of calculated arithmetic mean and 95% UCL mean concentrations for data groups for Study Sites A (Tables 7-4a&b) and B (Tables 7-5a&b) are similar between the two sets of data for each site. The RSDs for the two Study Site A data sets are 3.8% and 6.1%, respectively, with correlative means for

the 20 iterations for each data set of 344 mg/kg and 375 mg/kg. The RSDs for the two Study Site B data sets are 8.6% and 8.1%, respectively, with correlative means for the 20 iterations for each data set of 344 mg/kg and 375 mg/kg. Although only two sets of grid points were evaluated, this suggests that the field precision of twelve-point data sets designated in a systematic, random fashion is reasonably good for these two study sites.

Tables 7-6a and 7-6b summarize geostatistical analyses of random groupings of intra-sample discrete data for Study Site C. In this case the difference in average means and 95% UCLs between the two data sets is dramatic. The mean PCB concentration for the 20 iterations of the first data set is 981 mg/kg, with an RSD of 46% (Table 7-6a). The mean for the 20 iterations of the second data set is significantly lower, at just 86 mg/kg, with a lower RSD of 25%. This is due to the chance inclusion of two, small-scale, "hot spots" in the first data set (Grid Points 12 and 24).

7.3.3 Precision of Random, Twenty Four-Point Data Sets

As discussed in Part 1, Section 5.4, data for Multi Increment samples collected at each of the study sites is considered to provide the most precise estimate of the true mean of the target contamination. This is due to the significantly greater sample support represented by the MIS data, including the systematic control of bias during sample collection, the increased number of points within each area represented and the significantly higher mass of soil represented by the data (refer to Part 1, Table 5-6). These data are used to evaluate estimates of mean contaminant concentrations based on random combinations of discrete sample results for each of the 24 grid points at the three study sites.

Tables 7-7, 7-8 and 7-9 summarize geostatistical analyses of random groupings of intra-sample discrete data for each of the 24 grid points at Study Site A (arsenic), Study Site B (lead) and Study Site C (total PCBs), respectively. Discrete and MIS data for Study Site A (arsenic) and Study Site B (lead) are not directly comparable due to the use of a portable XRF to test the discrete samples and a laboratory extraction method to test the MIS samples (refer to Sections 4.1 and 4.2 in Part 1). Adjustment of the discrete data to reflect the average difference of the sample type results allows for a more useful comparison of the data, however.

Study Site A

Random combinations of discrete intra-sample data for the 24 grid points at Study Site A yield a fairly tight distribution of estimated mean, arsenic concentrations, ranging from 345 mg/kg to 383 mg/kg with a mean of 364 mg/kg and a Relative Standard Deviation of 3.0% (Table 7-7). The RSD suggests that the precision of the data sets in total to represent the collective, mean concentration of arsenic for the 24 grid points themselves is strong. As discussed above, however, the RSD in itself cannot be used to fully assess the precision of the 24 grid points to represent the mean concentration of arsenic for the larger study area as a whole.

The RSD values for given individual sets of discrete sample data ranges from 39% to 54% (see Table 7-7). This suggests that the precision of any single set of discrete sample data (or small number of discrete samples) to estimate a mean concentration of arsenic in soil at the site is moderate to somewhat poor. The estimated precision is still greater than observed at the other study sites, however.

Replicate MIS data for the study area yielded arsenic concentrations of 220 mg/kg, 230 mg/kg and 250 mg/kg, with a mean of 233 mg/kg (Table 7-10; see also Part 1, Table 5-5). The RSD for the MIS data is 6.5%, indicating a very good precision of the data to estimate the mean arsenic concentration for the study areas as a whole (see Table 7-11). Adjusting the mean of the discrete sample data sets of 364 mg/kg downward to reflect an average +31% bias in XRF data compared to Method 6010B data for MI samples (see Section 4.1 in Part 1) yields an Method 6010B equivalent of 251 mg/kg.

Comparison of the MIS and discrete data suggest that the 24, random points represented by the latter over-estimate the true concentration of arsenic for the study area as a whole by a factor of 20%. This is a reasonably good correlation in terms or relative error, and given the lower quality of sample support represented by the discrete data sets in comparison to the MIS data. The average, 95% UCL calculated for the random sets of discrete data of 426 mg/kg (see Table 7-7) over estimates the Student's t 95% UCL calculated for the MIS replicates adjusted for XRF of 339 mg/kg by a factor of 26% (See Table 7-10).

Study Site B

The variability of mean lead concentrations for random combinations of discrete sample data for Study Site B is slightly higher. Estimates of mean concentration of lead for the soil range from 235 mg/kg to 281 mg/kg, with a mean of 260 mg/kg and a Relative Standard Deviation of 5.5% (Table 7-8). The RSD suggests that the precision of the data sets in total to represent the collective, mean concentration of lead for the 24 grid points themselves is again strong.

The RSD values for individual sets of discrete sample data ranges from 49% to 80% (see Table 7-8). This suggests that the precision of any single set of discrete sample data (or small number of discrete samples) to estimate a mean concentration of lead in soil at the site is poor.

Replicate MIS data for the study area yielded lead concentrations of 240 mg/kg, 270 mg/kg and 350 mg/kg, with a mean concentration of 287 mg/kg (see Table 7-10; see also Part 1, Table 5-5). The RSD for the MIS data is 20%, indicating a reasonably good precision of the data to estimate the mean lead concentration for the study areas as a whole (see Table 7-11). Adjusting the mean of the discrete sample data sets of 260 mg/kg upward to reflect an average -6.8% bias in XRF data compared to Method 6010B data for MI samples (see Section 4.2 in Part 1) yields an Method 6010B equivalent of 278 mg/kg.

Comparison of the MIS and discrete data suggest that the 24, random points represented by the latter under estimate the true concentration of lead for the study area as a whole by a factor of

only 3.0%. This is somewhat remarkable, considering the significantly higher small-scale variability of lead concentrations in discrete samples at the site (median RSD 650% vs 96% at Study Site A; refer to Table 5-1 in Part 1). This could be interpreted to suggest that while the small-scale variability of lead concentrations in soil at Study Site B is high, the overall range of lead concentrations within any given area of the site is relatively similar. For comparison, concentrations of arsenic in soil within Study Site A appear to be significantly more variable from point to point, increasing the possibility that the true mean within the area could be significantly underestimated or overestimated based on a relatively small number of discrete soil samples.

The average, 95% UCL calculated for the random sets of discrete data of 328 mg/kg (see Table 7-8) underestimates the Student's t 95% UCL calculated for the MIS replicate data adjusted for XRF of 357 mg/kg by only 8% (see Table 7-10). This again demonstrates very good correlation with the MIS data even given the significantly lower quality of sample support (see Part 1, Table 5-5).

Study Site C

Evaluation of random sets of discrete sample data for Study Site C is especially interesting, given the exceedingly high, small-scale variability of total PCB concentrations in soil within and around the 24 grid points as well as the apparent, larger-scale variability of PCB concentrations across the site. Estimates of the mean concentration of PCBs range from 131 mg/kg to 972 mg/kg, with a mean of 534 mg/kg and a Relative Standard Deviation of 42% for the twenty, random groupings of data (Table 7-9). The RSD suggests that a moderate precision of the data sets to represent the collective, mean concentration of PCBs for the 24 grid points as a whole. Based on the MIS replicate samples collected from the same area, however, the grid points do not appear to be representative of the larger study area as a whole.

The RSD values for given sets of discrete sample data ranges from 251% to 434% (see Table 7-9). This suggests that the precision of any single set of discrete sample data to estimate a mean concentration of PCBs in soil at the site is exceptionally unreliable.

Replicate MIS data for the study area yielded PCB concentrations of 19 mg/kg, 24 mg/kg and 270 mg/kg with a mean of 104 mg/kg (see Table 7-10; see also Part 1, Table 5-5). The RSD for the MIS data is 138% (see Table 7-11), indicating a very poor precision of the data to estimate the mean PCB concentration for the study areas as a whole. Even so, the level of sample support for the MIS data is significantly higher than for the discrete sample data sets (e.g., see Table 5-6 in Part 1).

Comparison of the MIS and discrete data suggest that the 24, random points represented by the latter significantly *overestimate* the true concentration of total PCBs within the study area. The average, 95% UCL calculated for the random sets of discrete data of 4,399 mg/kg (see Table 7-9) is dramatically higher than the Chebyshev 95% UCL calculated for the MIS replicate data of 467

mg/kg (Table 7-10; see also Part 1, Table 5-5). The comparison of the data suggest that the discrete sample data set over represents small-scale "hot spots" of very elevated PCB concentrations within the study area. The unreliability of the 24-point discrete data set is further highlighted by the high RSD of the 60-point MI samples, which suggest that testing of greater than 60 points within the study area is required to adequately capture and represent the distributional heterogeneity of PCBs in the soil as a whole. Testing of few samples points (or increments) could either underestimate or, in this case, overestimate the likely true mean concentration of PCBs for the area.

Summary

This brief review further illustrates the potential low *field precision* of random grids of discrete samples for the site, even when the statistical precision of data for a given set of grid points is relatively good. While it is entirely possible that a small number of discrete soil samples, 20 or 30 or even less (USEPA 1992b), might adequately represent the mean contaminant concentration for a targeted area, whether in fact the sample set provided is indeed representative can only be known if replicate and completely independent sets of discrete samples are collected and tested.

The problem is particularly acute when using a small number of discrete samples, and even a single discrete sample, to determine the boundaries of contaminated soil that could pose a potential risk to human health and the environment. The collection and independent testing of large numbers of discrete samples to verify "contaminated" and "clean" areas is unlikely to be economically feasible, however. It is also unnecessary from a sampling theory perspective. The use of Decision Unit and incremental sampling methodologies, combined with the collection or replicate samples, is a far more efficient and effective means to collect high quality data for decision making. This was realized decades ago by the mining, agriculture and food industries but is only now beginning to be understood by the environmental industry.

7.4 Acute Toxicity

Several USEPA guidance documents mention the concept of using discrete soil samples to determine the presence or absence of very small but unspecified "hot spots," that could pose "acute" toxicity risks (i.e., health effects within minutes or a few days; USEPA 2011a), with data to be compared to as yet undeveloped acute toxicity or "not-to-exceed" screening level (e.g., USEPA 1989a, 1992a; see also Attachment 1). This concept was made prominent in the USEPA document *Guidance on Surface Soil Cleanup at Hazardous Waste Sites: Implementing Cleanup Levels,* with such criteria referred to as "Remedial Action Levels" (USEPA 2005; annotations added; note that this document is a Peer Review Draft and to our knowledge has not been finalized):

> Because soils with contaminant concentrations exceeding the cleanup level will be left onsite, it is important to ensure that those concentrations are not so high that they pose acute or subchronic health risks if exposure to them occurs. Therefore…, the (project

manager) should conduct a separate assessment of potential acute effects to determine the contaminant concentration at which acute effects are likely to occur.

To those unfamiliar with risk assessment or sampling theory this may at first seem reasonable and feasible. The potential for and evaluation of "acute" toxicity risk is in fact entirely hypothetical. Acute or not-to-exceed soil screening levels have never, to the authors' knowledge, been published by the USEPA. It is worth noting that none of the documents provide guidance on the calculation of either "acute" or "not-to-exceed" screening levels, nor do they provide guidance on sampling methods to establish with any degree of reliability the presence or absence of contaminated soil that could pose such concerns.

Acute toxicity would in theory need to be tied to the masses of soil as small as ten grams, the default mass of soil assumed to be ingested by a pica child (USEPA 2011b). Each ten-gram mass of soil at a site then becomes an individual "Decision Unit." Were acute toxicity factors and screening levels in fact available, the level of effort to prove beyond a reasonable doubt that no single, ten-gram mass of soil poses acute toxicity risks for even relatively small areas would be enormous and not feasible from either a technical or financial standpoint.

In practice regulators do not routinely require that sites be investigated to evaluate potential short-term, acute toxicity concerns posed by small-scale hot spots. Decision making is instead made in terms of potential long-term, chronic health risk to much lower concentrations of contaminants in soil, as discussed above. This more reasonable and feasibly requires comparison of the *mean* contaminant concentrations for large-scale, spill areas or exposure areas to risk-based screening levels for long-term exposure. The investigation and remediation of contaminated soil to meet significantly lower, risk-based screening levels for long-term, chronic risk based on conservative exposure assumptions, DU designations and incremental sampling data methods can reasonably be assumed to address short-term, acute exposure to theoretical and unidentifiable "hot spots" of contamination within these areas. If acute health risks are indeed a concern at a site, for example the incidental ingestion of lead-based paint chips or lead shot randomly scattered in soil, then the area should be remediated (e.g., scraped or capped) and confirmation, Multi Increment soil samples collected to evaluate any remaining chronic exposure risk (refer to HDOH 2011, and updates). Soil samples could also be ground to help assess the potential presence of large nuggets of targeted contaminants.

7.5 Outlier Data

Perhaps no other issue leads to more debate and confusion in the environmental industry than the interpretation and use of apparent "outlier" discrete sample data. In the mining industry, randomly located "outlier" veins or pockets of target mineral concentrations may make or break the economic viability of an ore deposit. Sampling protocols for ore bodies are carefully designed to capture and represent the smaller-scale variability of the targeted mineral within the body in order to make sound decisions (see Pitard 1993). Over representation of such "hot spots" can lead to over estimates of the mass of the targeted mineral present and subsequent

economic failure of the venture. Under representation of "hot spots" (or over representation of "cold spots") can lead to the missed discovery of materials critical to the success of a mining venture.

The same concepts apply to the investigation of contaminants in soil. An equivalent appreciation of the importance of "outliers" and "distributional heterogeneity" of targeted analytes has until very recently, however, been lacking in most environmental sampling guidance. The importance of understanding the implications "outlier" data is recognized in the USEPA guidance document *Methods for Evaluating the Attainment of Cleanup Standards* (USEPA 1989):

> This document recommends that all data not known to be in error should be considered valid… High concentrations are of particular concern for their potential health and environmental impact.

Such data can cause significant problems with the precision of geostatistical models, however. Consider, for example, this statement in the USEPA ProUCL document (USEPA 2013; see also following section):

> The inclusion of outliers in the computation of the various decision statistics tends to yield inflated values of those decision statistics, which can lead to incorrect decisions. Often inflated statistics computed using a few outliers tend to represent those outliers rather than representing the main dominant population of interest (e.g., reference area).
>
> Outliers represent observations coming from populations different from the main dominant population represented by the majority of the data set. *Outliers distort most statistics (e.g., mean, UCLs, UPLs, test statistics) of interest* (emphasis added). Therefore, it is desirable to compute decisions statistics based upon data sets representing the main dominant population and not to compute distorted statistics by accommodating a few low probability outliers (e.g., by using a lognormal distribution).

While perhaps true in some sampling scenarios, the suggestion that outliers "distort" estimation of the mean and should therefore not be "accommodated" in geostatistical analysis of a soil sample data set is misleading. Concentration is a function of the volume tested, or otherwise directly represented by the lab subsample (see ITRC 2012). The true mean, for example, of a one cubic-meter volume of soil is a composite of every particle of soil within that volume. Removal of small, "outlier hot spots" from the volume prior to testing, for example chips of lead-based paint, would yield erroneous data and conclusions.

Removal of so-called outlier data from individual, discrete sample data sets prior to statistical evaluation masks the imprecision of the calculated mean and further reduces the reliability of the data set. Such an approach, for example, would certainly not be acceptable for evaluation of an ore deposit. It is the presence of such outliers, including non-mineralized areas of the deposit ("cold spots") as well as isolated veins and randomly scattered pockets of concentrated

mineralization ("hot spots"), which control the mean concentration and overall economic viability of a deposit. The specific locations and numbers of small, concentrated accumulations of targeted analytes is not important, since the ore must be crushed and processed in its entirety in order to extract the mineral. The mean, not the mode or the median, is the objective. This is not a modern concept; the same issue was no doubt discussed by budding geologists and engineers of the stone, iron and bronze ages of ancient history.

The same process applies to environmental risk assessments. Characterization of the "population" of discrete sample-size masses of soil within an overall, targeted area and volume of soil is not the objective of a site investigation. The concentration of a contaminant within any given, discrete sample-size mass of soil within the targeted body of soil is inconsequential in terms of risk or estimation of the total mass of contaminant present. Estimation of the mean is the objective, not estimation of the mode of discrete samples collected from the soil (i.e., the concentration of the contaminant in "the main dominant population"). Exposure "Outliers" don't "distort" the health risk posed by contaminants in soil or the total mass of contaminant in the soil; they drive risk and they drive mass. Their identification and inclusion in the dataset in a representative manner is critical to accurate and technically defensible decision making.

The difficultly in dealing with outlier data in statistical analysis is real, but this is an artifact of the sampling and statistical methods used rather than an "error" in the contaminant distribution in the targeted area and volume of soil. Intentionally inducing error in the data in order to accommodate shortcomings in sampling approaches and statistical methods used to evaluate discrete sample data is unacceptable from a science standpoint. Inappropriate manipulation of a dataset should be considered a fifth source of error in decision making, in addition to sources of error associated with the physical nature of contaminants in soil and analytical error discussed in the following section.

Apparent outliers in discrete soil sample data are often tied to the presence of concentrated clumps or "nuggets" of a contaminant within the soil at a size that approaches the nominal particle size of the soil itself. As somewhat bluntly stated by Pitard (1993):

> As samples (i.e., laboratory subsamples) become too small, the probability of having one of these grains present in one selected sample diminishes drastically; furthermore, when one grain is present, the estimator … of the true unknown average… becomes so high that it is often considered as an outlier by the unexperienced operator.

Pitard repeatedly emphasizes the need for sampling methods that accurately represent all parts of the investigation area:

> All the constituents of the lot to be sampled must be given an equal probability… of being selected and preserved as part of the sample (and estimation of the mean; Pitard 2005).

> A common error has been to reject "outliers" that cannot be made to fit the Gaussian model or some modification of it as the popular lognormal model. The tendency, used by some geostatisticians, has been to make the data fit a preconceived model instead of searching for a model that fits the data… It is now apparent that outliers are often the most important data points in a given data-set (Pitard 2009).
>
> …the above sampling protocol (i.e., improper sample mass, sample collection, sample processing, etc.) introduces an enormous fundamental error (in the data set), resulting in a huge artificial nugget effect that confuses the interpretation of the data, subsequent geostatistical studies, and even the feasibility of the project (Pitard 1993).

As clearly demonstrated in the data collected as part of this study, "outliers" can appear or disappear based simply on the mass of soil randomly selected from a (unprocessed) sample for analysis or by moving the sample collection point over a seemingly insignificant distance (e.g., a few inches or feet).

The recommendation in the ProUCL guidance to similarly ignore "non-detect (ND)" results in the statistical evaluation of data set is similarly inappropriate for soil data (USEPA 2013). The document correctly calls out the same problem with the inclusion of ND results in statistical evaluation of data sets, stating that the statistical models employed "…do not perform well even when the percentage of ND observations is low." This again implies a failure of the approach being employed to estimate a mean from both a field and statistical standpoint, rather than an error in the data provided.

Unlike mining, agriculture, and most other industries, these problems have largely remained hidden or misunderstood in the environmental industry, largely due to the fact that the reproducibility of discrete sample data is not routinely tested. Far from being "distortions," the ability to capture and represent "outliers" in sample data is what makes ore deposits economic, crop yield predictions reliable and, as discussed in the following section, estimates of risk to human health and the environment technically defensible. Like an uncalibrated instrument that produces the wrong reading, the inappropriate interpretation of discrete sample data can lead to significant mistakes in decision making.

The opposite is true of incremental sampling approaches, where the objective is to collect a representative sample of a well-defined, targeted area and volume of soil. In contrast to the recommendations in ProUCL, the incorporation of "outliers" and "NDs" in correct proportions is a requirement for the collection of representative samples and defensible decision making under incremental sampling approaches (HDOH 2008; ITRC 2012). Such practices have survived in discrete sampling methodologies primarily due to the mistaken assumption that the variability of contaminant concentrations at the (arbitrary) scale of a discrete sample must be determined in order to estimate a defensible mean and perhaps more importantly to assess the "maximum"

concentration of a contaminant in the soil as part of the risk assessment process (refer to Section 3).

7.6 Implications for Use of Discrete Sample Data in Risk Assessments

The implications of the above observations on the use of single sets of discrete sample data to estimate mean contaminant concentrations for targeted exposure or source areas are significant. Estimation of the mean concentration of a contaminant for a targeted DU area and volume of soil can in theory be accomplished by testing of (processed or unprocessed) discrete soil samples. Variability due to random, small-scale distributional heterogeneity can be expected to increase as the mass of soil tested decreases, however (USEPA 2003; refer to Section 2). Greater variability lessens confidence in the precision of the geostatistical method employed to estimate a reliable mean (USEPA 2013).

The difficulty of using discrete soil sample data to estimate mean contaminant concentrations at sites where small-scale variability is high is acknowledged in the USEPA RAGS guidance (USEPA 1989b):

> If there is great variability in measured or modeled concentration values (such as when too few samples are taken or when model inputs are uncertain), the upper confidence limit on the average concentration will be high, and conceivably could be above the maximum detected or modeled value.

The authors were in all likelihood unfamiliar with the theory of sampling (Pitard 1993) at the time the document was prepared. In absence of an alternative approach, and apparently under the assumption that additional discrete data could not or would not be collected in most cases, they default to use of the maximum concentration detected as the exposure concentration for the targeted area (USEPA 1989b; emphasis added).

> In these cases, the *maximum* detected or modeled value should be used to estimate exposure concentrations. This could be regarded by some as too conservative an estimate, but given the uncertainty in the data in these situations, this approach is regarded as reasonable.

Recall that laboratories may only test one to ten grams of soil for certain contaminants of concern (maximum tested typically 30 grams). The upper 10cm (four inches) of a relatively small, 100m^2 (1,000ft^2) exposure area includes roughly ten metric tons of soil - 10,000,000 one-gram masses or 1,000,000, ten-gram masses for potential analysis. The potential for a small number (e.g., <20-30; USEPA 1992b) of samples collected from this total population to identify the true maximum concentration present is slim. This raises the question of what the maximum concentration of a contaminant reported for a small set of samples in fact represents. In truth it is highly unlikely to represent either the true “maximum” or the mean concentration of the contaminant present for the area as a whole and is of little use for investigation or risk

assessment purposes. It reflects the maximum concentration of the contaminant in an arbitrary, laboratory subsample mass from the specified set of discrete samples, nothing more.

Perhaps even more important in terms of uncertainty is the representativeness of the data set itself for the targeted exposure area as a whole, even in cases where variability between individual data points is relatively low. Use of a 95% UCL to estimate an exposure area concentration for use in a risk assessment addresses only the precision of the method used to assess the data set provided, not the representativeness of the data set itself. The precision of a data set can only be assessed through the collection and comparison of additional, independent sets of data to the original data. Measuring the precision of the data in terms of representativeness is a core part of incremental sampling but is rarely if ever evaluated as part of discrete sampling approaches.

This issue is not explored in the RAGs document or related risk assessment guidance documents, including classic guidance for ecological risk assessment (USEPA 1989c). As demonstrated above using data from the study sites investigated as part of this report, uncertainty in the representativeness of a random set of discrete sample data collected from a targeted area seems difficult if not impossible to measure. The use of "block kriging" and similar techniques to estimate mean contaminant concentrations for targeted areas faces the same set of problems regarding data representativeness and reliability.

Block kriging techniques combine isoconcentration mapping programs with geostatistical methods used to estimate mean contaminant concentrations for targeted areas of a site (USEPA 1992a):

> The investigator or RPM at a site desires to know not the concentration at a particular point in space but the average concentration over a block of soil that represents either an actual or potential risk to a human population or the environment.

The USEPA document notes potential limitations of this approach (USEPA 1992a; notations added):

> "(The technique assumes)... that contaminant concentrations are devoid of any spatial structure or correlation, and that the sampling is unbiased and accurately represents exposure concentrations… If the sample soil concentrations display spatial structures or correlations, or if the samples do not accurately represent exposure or are collected in a biased way (e.g., oversampling of areas thought to have high concentrations), then application of non-spatial statistical techniques should result in unreliable (estimates of mean contaminant concentrations)."

As discussed above, the validity of these assumptions is unknown in the absence of replicate sets of data for targeted areas to test precision in terms of the representativeness of a data set as a whole. The use of smaller, subsets of the original discrete sample data set to estimate mean

contaminant concentration for subareas of the site investigated further decreases the reliability of the conclusions. Additional evaluation of this issue and acceptability of the continued use of discrete sample data in environmental risk assessments is warranted but beyond the scope of this report at this time.

8 OTHER DISCRETE SAMPLE ISSUES

As described in earlier sections, recommendations for discrete soil sampling methods are incorporated in numerous USEPA (and state) guidance documents written in the early 1980s and 1990s. This reflects our limited understanding of contaminant heterogeneity in soil at the time. State and even federal environmental regulators are largely free to move toward more advanced, incremental sampling methods once the need is realized and training of staff and consultants has been undertaken. Training of regulators and consultants has surged since publication of ITRCs *Incremental Sampling Methodology* guidance (ITRC 2012). Efforts to update laboratory protocols for processing and subsampling of soil samples have been underway for more that ten years. The most challenging effort ahead perhaps applies to revisit regulatory requirements for the collection of discrete soil samples to characterize and remediate sites. As discussed below, the multitude of sampling requirements for PCBs embedded in regulations prepared under the Toxic Substances Control Act (TSCA) are a notorious example.

8.1 LABORATORY PROCESSING AND TESTING PROTOCOLS

The study clearly demonstrates that the concentration of a contaminant in soil can be highly variable at the scale of a typical, laboratory subsample mass, for example one gram for most metals and ten to thirty grams for other chemicals (Figure 8-1). In the absence of adequate processing and collection of the subsample mass to be tested, the representativeness of the resulting laboratory data for the sample as a whole cannot be assumed. This requires that the objectives of a site investigation and the ability of discrete sample data to meet these objectives be carefully reviewed.

If the objective of the site investigation requires that a reliable mean concentration of a contaminant be determined for individual discrete samples then appropriate processing and sub-sampling of the discrete samples is required. Uncertainty can be reduced by processing and collecting subsamples in a manner that captures small-scale variability within the sample submitted for analysis (see USEPA 2003). Most laboratory analysis methodologies call for "homogenization" of samples prior to the collection of a subsample mass for testing (e.g., Figure 8-2). Specific methods to accomplish this are lacking in method-specific protocols, however, with the exception of Method 8330B for explosives (USEPA 2006).

Chapter 3 of the USEPA SW846 guidance for inorganic calls for samples to be "well-mixed and homogenized" prior to the collection of a subsample for analysis (USEPA 2007). Grinding of samples is recommended if data for replicate subsamples are significantly different. In practice grinding of samples is rare due to both the time and expense involved and concerns about the representativeness of the resulting data for use in risk assessment, since grinding can unnaturally increase the bioavailable fraction of a contaminant in soil. Replicate laboratory subsample testing from each "batch" of samples to assess sub-sampling data representativeness is relatively common (typically 1 replicate for each 10-20 samples; see also USEPA 1991, 1992b). The

higher of reported contaminant concentrations reported for the replicate data is typically used for decision making, however, with the discrepancy assumed to be "laboratory error."

Based on discussions with commercial laboratories, in the absence of method-specific directions "homogenization" of a sample, if carried out at all, typically consists of mechanical mixing of the sample prior to the collection of the mass required for digestion and analysis. While this can reduce "intra-sample" variability under some circumstances, mechanical mixing can also increase heterogeneity due to separation of fine and coarse particle fractions. The resulting laboratory replicate data can be highly variable (refer to USEPA 2003).

Processing of discrete soil samples in the same manner as carried out for Multi Increment samples is recommended. For non-volatile chemicals, this includes air drying, sieving to eliminate sticks/stones and determine the maximum particle size that will be analyzed (generally <2mm for many soil contaminants, however some methods call for fines analyses (<0.25mm), and representative subsampling for the collection of the minimum digestion/analysis mass required by the maximum particle size of the sample and the laboratory method (Figure 8-3; refer to Section 4 of the HEER office Technical Guidance Manual, HDOH 2008; see also ITRC 2012). This minimizes the effect of "intra-sample" variability on the representativeness of the resulting laboratory data for the sample as a whole as well as between samples within a data set. As discussed in Section 7, whether the resulting data set is in fact representative for the targeted area and volume of soil as a whole may still be questionable.

Error associated with intra-sample variability can be overcome by properly processing and sub-sampling the sample, and variability is typically reduced by analyzing a larger mass of soil. For < 2mm-sized particulate samples, a digestion/analysis mass above ten grams for metals and thirty grams for most other chemicals is not common, given cost and technical constraints of the laboratory (HDOH 2008). Correct processing and subsampling with incremental sampling methodology was employed for the second part of this study in order to evaluate distributional heterogeneity between closely spaced discrete samples collected around individual grid points that were processed prior to analysis.

The presence of a chemical in soil as scattered nuggets rather than finely disseminated particles can also be expected to lead to significant variability in laboratory replicates. When the mass of subsamples from an unprocessed sample becomes too small, variability in the number and overall mass of contaminant nuggets within any given subsample increases dramatically, (see Pitard 1993). This reflects the intrinsic heterogeneity of contaminant distribution within the sample and not laboratory analytical error, as might be otherwise assumed. The probability of having a representative number of nuggets in a laboratory subsample mass diminishes with decreasing mass collected. This supports the need for the extraction of larger subsample masses at commercial laboratories, the exact opposite of the current trend of smaller and smaller masses.

8.2 Estimation of Contaminant Mass for *In Situ* Remediation

Discussions and field data presented as part of this study focus on the investigation of surface soils. The same unavoidable errors regarding the true extent and magnitude of contamination decision making likewise apply to subsurface soil investigations. Consider again the pattern of the milk release presented in Figure 5-9 in Part 1 one this report (reproduced in Figure 8-4). In this case, the release of milk followed "preferential pathways" related to low lying areas of the ground surface to create highly disjointed fingers of "contamination" downgradient of the source area.

The mechanics of infiltration has been widely studied for rainfall in hydrogeologic studies but less so for releases of hazardous liquids to the ground surface. As stated in Freeze and Cherry's classic book on groundwater (Freeze and Cherry 1979): "Small differences in the hydrologic properties of similar field soils can account for large differences in their reaction to the same hydrologic event (i.e., surface runoff versus infiltration)." Similar, heterogeneous patterns are also likely to characterize the vertical migration of released liquids, due to both the initial dispersion pattern at the surface and minor variability in the permeability of subsurface soils.

Such heterogeneity can lead to a significant underestimation of the lateral and vertical extent of contamination as well as the mass of contaminants present. In many cases the target chemical of concern is a petroleum fuel, a chlorinated solvent or other highly mobile and volatile chemical. Under a traditional, discrete sample investigation, very small, five- to ten-gram plugs of soil are typically removed from set depth intervals in cores from a small number of borings. Even when sample point locations are biased to apparent, higher concentration areas of soil within a given core (e.g., using a Photoionization Detector), it is highly unlikely that the resulting data will be representative of the subsurface area beyond other than a very gross, screening level. As is often the case for surface soils, true "hot spots" are likely to be missed and the average concentration and mass of the chemical present underestimated.

Remedial experts are well aware of this dilemma and often compensate by assuming that up to an order of magnitude greater mass of contaminant could be present when designing *in situ* remedial actions. Even then, significant underestimation of contamination and/or failure to treat the full area can be common place. As discussed in the HEER office Technical Guidance Manual, Multi Increment sampling methodologies are far superior to traditional discrete sample methodologies for both in situ and ex situ remedial efforts (HDOH 2008). Under this approach Decision Unit layers are designated and fully subsampled based on the resolution required to optimize remediation (Figure 8-5 and Figure 8-6). Decisions are then made based on combined increments from large numbers of cores and/or based on testing of individual, targeted layers within single cores (e.g., in order to estimate the depth of contamination at a single location).

8.3 Toxic Substances Control Act Regulations

One of the most indoctrinated uses of discrete sampling methodologies can be found in regulations and guidance for the investigation of PCB-contaminated soils under the Toxic

Substances Control Act (USEPA 1985, 1986, 1990, 1998). Perhaps hundreds of thousands of sites across the US were instantly affected when the regulations were first passed. Releases of small volumes of PCB containing transformer oil were common place. The new regulations required that impacted soils be expeditiously tested for contamination and either removed or capped.

Guidance being developed by the USEPA, largely under contracts from outside consultants, again adopted sampling methods already in use for water and industrial waste streams for use in the investigation of PCB contaminated soil dating back to the 1970s. As stated in the introduction to the document *Verification of PCB Spill Cleanup by Sampling and Analysis* (USEPA 1985):

> The U.S. Environmental Protection Agency under the authority of the Toxic Substances Control Act (TSCA) Section 6(e) and 40 CFR Section 761.60(d), has determined that polychlorinated biphenyl (PCB) spills must be controlled and cleaned up. The Office of Toxic Substances has been requested to provide written guidelines for cleaning up PCB spills, with particular emphasis on the sampling design and sampling and analysis methods to be used for the cleanup of PCB spills.

Much of the emphasis in early guidance was placed on laboratory test methods, with sample collection methods secondary. Although prepared by different groups of consultants and project managers within the USEPA, the groups were no doubt familiar with each other's work.

The 1985 guidance incorporates two key misunderstandings of related guidance being developed at the time: 1) Risk-based screening levels being developed in conjunction with the guidance apply to any given, discrete-size mass of soil within an impacted area and 2) A single, discrete sample (of unspecified mass) can be assumed to be representative of the immediate surrounding soil. The most important error in the 1985 document is that risk-based screening levels for PCBs being developed at the time apply to any given, testable mass of soil within a potentially contaminated area (USEPA 1985):

> …the goal of the analysis effort is to *determine whether at least one sample* has a PCB concentration above the allowable limit. This sampling plan assumes the entire spill area will be recleaned if a single sample contaminated above the limit is found. Thus, it is not important to determine precisely which samples are contaminated or even exactly how many.

This simple statement subsequently controls how a suspected release area must be investigated. The guidance proceeds to present similar, ultimately misguided statements regarding the need to divide the screening level by the number of individual samples included in "composite" sampling strategy, with the maximum number of samples allowable based on division of the screening level by the laboratory reporting limit for PCBs at the time (USEPA 1985):

> *If the PCB level in the composite is sufficiently high, one can conclude that a contaminated sample is present*… The samples from which the composite was constructed must (therefore) be analyzed individually to make a determination… Do not form a composite with more than 10 samples, since in some situations compositing a greater number of samples may lead to such low PCB levels in the composite that the recommended analytical method approaches its limit of detection and becomes less reliable.

The maximum number of discrete samples that can be included in a composite sample in order to ensure that no single sample exceeds the target screening level is determined by dividing the screening level by the laboratory detection limit. For example, a maximum of ten discrete samples per composite is set based on dividing a screening level of 10 mg/kg by the then detection limit for PCBs in soil of 1 mg/kg.

Methods for establishing grids of adequate size to ensure that "hot spots" are not missed are subsequently presented. The guidance was specifically designed to address relatively small areas of contamination, based on a reported median affected area of 249 ft^2 reported for PCB spills from capacitors, with spills assumed to rarely affect an area <1,000ft^2. An elaborate statistical evaluation of the probability of failure, i.e., missing a discrete sample size mass of soil with a mean PCB concentration that exceeds the screening level, is likewise presented in the guidance. This is based on a second, key assumption that a single, discrete sample (mass not specified) is representative of a surrounding, circular area of soil defined by the grid spacing, with one-half of the grid spacing representing the radius of the circle (USEPA 1985; emphasis added):

> The implicit assumption that *residual contamination is equally likely to be present anywhere within the sampling area* is reasonable, at least as a first approximation… The detection problem was modeled as follows: *try to detect a circular area of uniform residual contamination* whose center is randomly placed within the sampling circle.

Importantly, further delineation of the lateral (and vertical) extent of contamination is presumed to be unnecessary once the reported concentration of PCBs in a single sample falls below the target screening level (USEPA 1985):

> …it is important to note that not all samples collected will need to be analyzed. The calculations… show that… no more than 8 analyses will usually be required to reach a decision.

This conclusion is based on an assumption that no more than 37 discrete samples will be required for release areas <1,000ft^2 in size and the grid will be designed under the assumption that the outer points fall in presumed clean areas. This mistaken assumption is the root cause of the preponderance of "false negatives" in PCB soil investigations.

The guidance then goes on to discuss the use of hexagonal grids of discrete samples to determine the presence or absence of PCB contamination associated with a given discrete sample point, with the assumption that the smaller the grid point spacing "…the smaller the residual contamination area which can be detected with a given probability." An elaborate review of "hot spot" detection probability is then provided. Based on the area of contamination reported to be associated with typical releases from capacitors, the guidance in essence recommends target "hot spot" areas for different size releases that range from $13ft^2$ to $32ft^2$ based on grid spacings of 4 to 6.4 feet and presumed, impacted areas that range in size from $50ft^2$ to $1{,}000ft^2$. Note that these areas are not based on risk (e.g., assumed exposure area). They are simply artifacts of the use of a hexagonal grid spacing approach for a default, circular spill area with a ten-foot diameter, slightly modified for different, assumed sizes of releases.

Assuming that the distribution of PCBs in soil within a spill area is "uniform," like a spot of paint spilled onto a piece of wood, greatly simplifies the investigation process. Sample collection and testing can cease once a "clean" spot below the target screening level is identified, both laterally and vertically. Just as important, since any size mass of soil is presumed to contain roughly the same concentration of PCBs, a sample of any mass can be assumed to be representative of the surrounding soil as a whole. This means, critically, that the mass of soil to be collected as a discrete sample only need meet the mass required by the laboratory for analysis, including quality control (default 100 grams per sample recommended; USEPA 1987). The concept of "data quality" could then be shifted to the comfortable confines of the laboratory with the main source of error presumed to be associated with analytical error. Testing of replicate samples to assess the precision of the resulting data were specifically focused on the precision of the analysis, rather than the precision of sample representativeness in the field.

Subsequent USEPA guidance documents (e.g., USEPA 1987, 1990) and regulations simply expand on the above themes to justify the use and outright requirement of discrete sampling approaches for the investigation of PCB-contaminated soil. Data quality is emphasized, but only after the sample has been collected and submitted to the laboratory for analysis (USEPA 1987):

> Quality assurance (QA) and quality control (QC) must be an integral part of any sampling scheme... Some of the requirements of quality control are discussed in this report, including field blanks, sampling without cross-contamination, sample custody, and documentation of the field sampling activities.

Field validation of the proposed, discrete sample grid approach was carried out at some point between June 1985 and May 1986 (refer to Section 10 of the 1986 USEPA manual). The validation, however, focused only the degree of difficulty in laying out the grid design, which was deemed to be acceptable. One cannot help but wonder how the direction of environmental soil investigations might have changed if the authors of the guidance had included validation of the critical assumption that the distribution of PCBs in soil was indeed "uniform" at the scale of a discrete sample.

Somewhat ironically, given what we are now beginning to realize about the magnitude and implications of random, small-scale variability of contaminant concentrations in soil and the implications for discrete sample investigations, the 1985 document concludes with this statement (USEPA 1985):

> …because an enforcement finding of noncompliance must be legally defensible; that is, a violator must not be able to claim that the sampling results could easily have been obtained by chance alone.

As demonstrated in this study random chance does, in fact, control the concentration of PCBs reported for any given discrete soil sample.

9 SUMMARY AND CONCLUSIONS

This field study set out to help answer three deceptively simple questions: 1) "How variable is the concentration of a contaminant within an unprocessed, discrete sample with respect to the mass of soil typically used for laboratory digestion/analysis (e.g., 1 to 30 grams)?", 2) "How variable is the concentration of a contaminant between co-located, discrete samples collected within a short distance of a given sample point?" and 3) "What are the implications for the reliable use of discrete soil samples in environmental investigations?" The results of the field study suggest that contaminant concentration variability at the scale of a discrete sample can be very high both within a discrete sample and between closely-spaced samples. The resulting implications are likewise significant

Discrete soil samples are not routinely processed to ensure representative sub-sampling prior to analysis. The data provided by the laboratory cannot, therefore, be assumed to be representative of the average concentration of the contaminant in the bulk sample originally submitted. The concentration of a contaminant reported for a discrete sample cannot be assumed to be representative of soil within the immediate vicinity of the point from which it was collected. In total, concentrations of a contaminant in "co-located" discrete samples can be expected to vary by at least a factor of two. Examples include the discharge of wastewater from well-controlled industrial processes or application of water-based pesticides to fine-grained soil, as evaluated for Study Site A. Concentrations of contaminants in co-located samples for mixtures of original clean soil with contaminated media can be expected to vary both within single samples and over short distances by at least an order of magnitude. Examples include mixtures of lead-contaminated ash from municipal incinerators with fill material investigated at Study Site B. Even more dramatic variability at the scale of a discrete sample should be expected for soil contaminated by releases of small particles or releases of waste oils, with variability within and between discrete samples reaching two orders of magnitude or more. Beading of waste transformer oil upon release to soil and the formation of PCB-infused "tar balls" is a likely contributor to the exceptionally high, small-scale variability at Study Site C.

This variability is random and cannot be assumed to be representative of larger-scale trends of contaminant distribution across a site. Implications for the reliable use of discrete soil samples as a routine part of environmental investigations are significant. Comparison of individual discrete soil sample data to risk-based, soil screening levels is highly prone to premature termination of a site investigation as the inherent variability of contaminant concentrations begins to fluctuate both above and below the screening level. This is the primary cause of "false negatives" in otherwise contaminated areas and the need for repeated remobilizations and sample collection following initial removal of contaminated soil.

"False positives," represented by seemingly isolated, sample-size "hot spots" are likewise an artifact of random variability. Surgically removing such spots cannot be assumed to have significantly reduced the mean concentration of the contaminant in the targeted area.

Artificial, seemingly mappable patterns of higher and lower concentrations generated within isoconcentration maps are likewise artifacts of random, small-scale variability and unlikely to be reproducible. Grids of discrete samples can be useful to help identify gross contamination patterns within a large area and designated decision units for more intensive, incremental sampling. Care must be taken, however, not to over-interpret the discrete sample data. Isoconcentration maps give a very false impression that discrete samples can be used to assess contaminant concentrations down to the resolution of a single sample mass (roughly 100-200 grams) or area (<100cm^2) for any given spot of soil within the target area. In reality such a fine resolution is both technically and economically impossible with the resources typically available as well as unnecessary from a human health and environmental risk perspective. Investigations should be carried out at the scale of well-thought-out "Decision Units" representing known or suspect "spill areas" or "exposure areas" with characterization of mean contaminant concentrations accomplished by intensive, incremental sampling methodologies (refer to HDOH 2008, ITRC 2012).

Evaluation of the total precision of discrete soil samples for estimation of mean contaminant concentrations with targeted spill areas or exposure areas is also problematic. It is possible to adequately capture small-scale heterogeneity of contaminant concentrations within an area using discrete samples and estimate a reliable, unbiased mean. Statistical analysis of a single data set only evaluates precision of the estimated mean in terms of the data set provided and the statistical method used. The *field precision* of the estimated mean remains unknown. The precision of the data set and estimated mean with respect to field representativeness of the targeted area can only be evaluated by the collection of independent, replicate sets of discrete samples from the same area for comparison to the original data set. In practice this is rarely if ever carried out. The collection of true field "replicates" to fully evaluate the precision of estimated mean contaminant concentrations is, in contrast, a required part of incremental sampling methodologies (see HDOH 2011; ITRC 2012).

The importance of replicate sampling did not go unnoticed by authors of early, USEPA site investigation guidance. In discussing the need to test the reproducibility of discrete sample data, the USEPA document *Data Quality Objectives for Remedial Response Activities* states the following (USEPA 1987; emphasis added):

> Collocated samples can be used to estimate the overall precision of a data collection activity. Sampling error can be estimated by the inclusion of collocated and replicated versions of the same sample. If a significant difference in precision between the two subsets is found, it may be attributed to sampling error. *As a data base on field sampling error is accumulated, the magnitude of sampling error can be determined.*

The fact that it has taken over 25 years to begin to realize and accept the magnitude of sampling error associated with the use of discrete soil samples is attributable to multiple factors, including: 1) The lack of field studies to verify the reliability of discrete sampling methodologies when site

investigation guidance was initially being prepared in the 1980s and 1990s, 2) The lack of a requirement for the collection of detailed sets of field replicate samples as a routine part of soil investigations, 3) The lack of market forces on the side of regulatory authorities to ensure high data quality for decision making and 4) The lack of training of regulators and consultants in sampling theory, based on experience from the mining and agricultural industries.

Sampling practices in the mining and agricultural industries quickly advanced due to market forces as ore bodies once thought to be profitable led to bankruptcy or crops failed due to inadequate understand of soil properties and nutrient needs. Such market forces are not a standard part of the government-directed, environmental industry. The demand for higher data quality and more cost-effective decision making is instead being led by those paying for investigations and cleanup. Innovative Decision Unit and incremental sampling methodologies offer significant time and cost savings for "responsible parties" and decrease uncertainty regarding future financial uncertainty for inadvertently missed contamination.

Such methodologies are well established for soil investigation projects in Hawai'i and are being expanded to sediment as well as surface water studies. The need for continued change is currently most urgent for the standalone use of incremental sampling methodologies at PCB sites that come under the oversight of USEPA TSCA offices. Requirements for the collection of discrete soil samples at PCB sites initially being investigated using Decision Unit and incremental sampling methodologies have in some cases added tens and hundreds of thousands of dollars to projects in Hawai'i, with no added benefit to human health and the environment and with significant disruption to the investigation and cleanup of the sites. Incremental sampling methods can be implemented under "risk-based" options in TSCA regulations, however, without the need for revisions to the regulations themselves. Staff with HEER office and USEPA were working to formalize this process at the time of this report.

References

ADEC, 2009, *Guidance on Multi Increment Soil Sampling*: Alaska Department of Environmental Conservation, Division of Spill Prevention and Response, March 2009.

Aelion, C.M ., Davis, H.T., Liu, Y., Lawson, A.B. and S. McDermott, 2009, Validation of Bayesian Kriging of Arsenic, Chromium, Lead, and Mercury Surface Soil Concentrations Based on Internode Sampling: Environmental Science and Technology, 43, pp. 4432–4438.

CDCP, 2012, Low Level Exposure Harms Children: A Renewed Call for Primary Prevention: Center for Disease Control and Prevention, Advisory Committee on Childhood Lead Poisoning Prevention, January 4, 2012.

Cutler, W.C., Hue, N., Ortiz-Escobar, M.E., and T. Martin, 2006, Approaches to Reduce Bioaccessible As in Hawaii Soils: Proceeding of Fifth International Conference on Remediation of Chlorinated and Recalcitrant Compounds, Monterey, CA, May 2006.

Cutler, W.C., 2011, Bioaccessible Arsenic in Soils of the Island of Hawaii: University of Hawai´i-Manoa, Department of Geology and Geophysics, PhD Dissertation, May 2011, 136p.

ERM, 2008, Sampling and Analysis Plan Amendment Former Pepe´ekeo Sugar Company Property Hakalau, Hawaii: Environmental Resources Management, 30 September 2008.

Freeze, R.A. and J.A. Cherry, 1979, *Groundwater*: Prentice Hall, London, England.

Gilbert, R.O., 1987, *Statistical Methods for Environmental Pollution Monitoring*: Van Nostrand Reinhold Company Inc., New York, New York.

Gosh, S.K., 1993, *Structural Geology, Fundamentals and Modern Developments*: Pergamon Press, New York.

Groundswell Technologies, 2013, Groundswell Technologies Web Application and API: Groundswell Technologies, Inc.

Hadley, P.W. and R.M. Sedman, 1992, How Hot Is That Spot?: Journal of Soil Contamination, 3, pp 217-225.

Hadley, P.W., Crapps, E. and A.D. Hewitt, 2011, Time for a Change of Scene: Environmental Forensics, 12, pp 312-318.

Hadley, P.W. and S.D. Mueller, 2012, Evaluating "Hot Spots" of Soil Contamination: Soil and Sediment Contamination, 21, pp 335-350.

Hadley. P.W. and I.G. Petrisor, 2013, Incremental Sampling: Challenges and Opportunities for Environmental Forensics: Environmental Forensics, 14, pp 109–120.

Hadley, P.S. and Bruce, M.L., 2014, On Representativeness: Environmental Forensics, 15:1, pp1-3.

HDOH, 2008, *Technical Guidance Manual* (2008 and updates): Hawai‘i Department of Health, Office of Hazard Evaluation and Emergency Response.

HIDOH, 2011, *Screening for Environmental Concerns at Sites with Contaminated Soil and Groundwater* (Fall 2011 and updates): Hawai'i Department of Health, Office of Hazard Evaluation and Emergency Response.

HDOH, 2012, *Hawaiian Islands Soil Metal Background Evaluation*: Hawai‘i Department of Health, Hazard Evaluation and Emergency Response (May 2012).

ITRC, 2012, Incremental Sampling Methodology: Interstate Technology Regulatory Council, February 2012.

Lu, G.Y. and D.W. Wong, 2008, An adaptive inverse-distance weighting spatial interpolation technique for computers and Geosciences, Vol. 34 (9), pp 1044-1055.

Minnitt, R.C.A., Rice, P.M. and C. Spangenberg, 2007, Part 1: Understanding the components of the fundamental sampling error: a key to good sampling practice: The Journal of the Southern African Institute of Mining and Metallurgy, August 2007, Vol. 107.

Pitard, F., F., 1993, Pierre Gy's Sampling Theory and Sampling Practice: CRC Press, New York, NY.

Pitard, F.F., 2005, Sampling Correctness - A Comprehensive Guideline: Sampling and Blending Conference, Sunshine Coast, Queensland, Australia, May 9-12, 2005.

Pitard, F.F., 2009, Theoretical, practical and economic difficulties in sampling for trace constituents: Fourth World Conference on Sampling & Blending, The Southern African Institute of Mining and Metallurgy, 2009.

Ramsey, C. A. and A.D. Hewitt, 2005, A Methodology for Assessing Sample Representativeness: Environmental Forensics, 6:71–75, 2005.

Roberts, S.M., Munson, J.W., Lowney, Y.W. and M.V. V. Ruby, 2007, Relative Oral Bioavailability of Arsenic from Contaminated Soils Measured in the Cynomolgus Monkey: Toxicological Sciences, Vol. 95(1), pp 281–288.

Shulgin, A.I., 2008, Evaluation of HMATM Treatment to Allow for Reuse of Waste Ash as Landfill Daily Cover: Microganics, LLC, October 30, 2008.

Silver, N., 2012, *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't*. New York: The Penguin Press.

USACE, 2009, Interim Guidance 09-02 Implementation of Incremental Sampling (IS) of Soil for the Military Munitions Response Program: U.S. Army Corps of Engineers, Environmental and Munitions Center of Expertise, July 20, 2009, Interim Guidance 09-02.

USEPA, 1985, *Verification of PCB Spill Cleanup by Sampling and Analysis*: U.S. Environmental Protection Agency, Office of Toxic Substances, EPA-560/5-85·026, August 1985, Washington DC.

USEPA, 1986, *Field Manual for Grid Sampling of PCB Spill Sites to Verify Cleanups*: U.S. Environmental Protection Agency, Office of Toxic Substances, EPA-560/5-86·017, May 1986, Washington DC.

USEPA, 1987, *Data Quality Objectives for Remedial Response Activities*: U.S. Environmental Protection Agency, Office of Emergency and Remedial Response, EPA/540/G-87/003, March 1987, Washington DC.

USEPA, 1988, *Superfund Exposure Assessment Manual*: U.S. Environmental Protection Agency, Office of Remedial Response, EPA/540/1-881001, April 1988, Washington DC.

USEPA, 1989a, *Methods for -Evaluating the Attainment of Cleanup Standards, Volume 1: Soils and Solid Media*: U.S. Environmental Protection Agency, Office of Policy, Planning, and Evaluation, EPA 230, U2-89-042, February 1989, Washington DC.

USEPA, 1989b, *Risk Assessment Guidance for Superfund, Volume I, Human Health Evaluation Manual (Part A):* U.S. Environmental Protection Agency, Office of Policy, Planning, and Evaluation, EPA/540/1-89/002, December 1989, Washington DC.

USEPA, 1989c, *Risk Assessment Guidance for Superfund, Volume II, Environmental Evaluation Manual:* U.S. Environmental Protection Agency, Office of Policy, Planning, and Evaluation, EPA/540/1-89/001, March 1989, Washington DC.

USEPA, 1990a, *Guidance on Remedial Actions for Superfund Sites with PCB Contamination:* U.S. Environmental Protection Agency, Office of Emergency and Remedial Response, EPA/540/G-90/007, August 1990, Washington DC.

USEPA, 1990b, *A Rationale for the Assessment of Errors in the Sampling of Soils*: U.S. Environmental Protection Agency, Environmental Monitoring Systems Laboratory, EPA/800/4-90/013, May 1990, Washington DC.

USEPA, 1991, *Guidance for Data Usability in Risk Assessment (Part A)*: Environmental Protection Agency, Office of Research and Development, EPA/540/R-92/003, December 1991, Washington DC.

USEPA 1992a, *Preparation of Soil Sampling Protocols: Sampling Techniques and Strategies*: Environmental Protection Agency, Office of Research and Development, EPA/600/R-92/128, July 1992, Washington DC.

USEPA, 1992b, *A Supplemental Guidance to RAGS: Calculating the Concentration Term*: U.S. Environmental Protection Agency, Office of Solid Waste and Emergency Response, EPA 9285.7-081, May 1992, Washington DC.

USEPA, 1998a, 40 CFR Part 761.61, PCB Remediation Waste: U.S. Environmental Protection Agency, Code of Federal Regulations.

USEPA, 1998b, *Test Methods for Evaluating Solid Waste, Physical/Chemical Methods (Revision 5)*: U.S. Environmental Protection Agency, Office of Solid Waste, SW-846 Manual, Washington, D.C., April 1998 (and updates).

USEPA, 1999. Correct Sampling Using the Theories of Pierre Gy: U.S. Environmental Protection Agency, National Exposure Research Laboratory, Environmental Sciences Division, Technology Support Center, Fact Sheet 197CMB98.FS-14, March 1999.

USEPA 2002, *Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites*: Office of Emergency and Remedial Response, OSWER 9285.6-10, December 2002, Washington DC.

USEPA, 2003, Guidance for Obtaining Representative Laboratory Analytical Subsamples from Particulate Laboratory Samples: U.S. Environmental Protection Agency, Office of Research and Development, EPA/600/R-03/027, November 2003, Washington DC.

USEPA, 2005a, Polychlorinated Biphenyls (PCBs) Manufacturing, Processing, Distribution in Commerce and Use Prohibitions**:** U.S. Environmental Protection Agency, Code of Federal Regulations, Chapter 40 Section 261.

USEPA, 2005b, *Guidance on Surface Soil Cleanup at Hazardous Waste Sites: Implementing Cleanup Levels (Peer Review Draft)*: U.S. Environmental Protection Agency, Office of Emergency and Remedial Response, EPA 9355.0-91, April 2005.

USEPA, 2007, SW846, Chapter Three, Inorganic Analytes: U.S. Environmental Protection Agency, February 2007, Washington DC.

USEPA, 2011a, IRIS Glossary: U.S. Environmental Protection Agency, National Center for Environmental Assessment, Integrated Risk Information System, last updated August 31, 2011.

USEPA, 2011b, *Exposure Factors Handbook*: U.S. Environmental Protection Agency, National Center for Environmental Assessment, Office of Research and Development, September 2011, EPA/600/R-09/052F.

USEPA, 2013, ProUCL Version 5.0.00, User Guide: U.S. Environmental Protection Agency, Office of Research and Development, EPA/600/R-07/041, September 2013, Washington DC.

USEPA, 2014a, Hot Spots: Incremental Sampling Methodology (ISM) FAQs: U.S. Environmental Protection Agency, Superfund, March 27, 2014.

USEPA, 2014b, *Screening Levels for Chemical Contaminants*: U.S. Environmental Protection Agency, (2014), prepared by Oak Ridge National Laboratories, http://www.epa.gov/region09/waste/sfund/prg/

USGS, 2004, Geologic Provinces of the United States: Records of an Active Earth: U.S. Geological Survey, USGS Geology in the Parks, http://geomaps.wr.usgs.gov/parks/index.html

USGS, 2014, Geochemical and Mineralogical Maps for Soils of the Conterminous United States: U.S. Geological Survey, USGS Open-File Report 2014-1082.

Table 3-1. Summary of estimated, relative percent difference between minimum and maximum concentration of contaminant in soil within a 0.5m radius of a grid point at each study site.

Study Site	[1]Estimated Total RPD			
	Minimum RPD	**Maximum RPD**	**Median RPD**	**Mean RPD**
Site A (arsenic)	29%	308%	96%	112%
Site B (lead)	205%	4050%	650%	879%
Site C (PCBs)	524%	115916%	3802%	19550%

1. Minimum, maximum, median and mean RPDs calculated for individual grid points at each study site. Refer to Table 4-7 (Site A), Table 4-14 (Site B) and Table 4-20 (Site C) in Part 1 of the study report. Collection of additional discrete sample data around grid points would likely identify greater variability and a wider range of RPDs.

Table 4-1. Summary of estimated, total variability of contaminant concentrations in soil within a 0.5m radius of a grid point at each study site.

Study Site	[1]Median Total Variability	[1]Range Total Variability	[2]Median RPD	[2]Range RPD
Site A (arsenic)	2.0	1.3 to 4.1	96%	29% to 308%
Site B (lead)	7.5	3.2 to 42	650%	205% to 4,050%
Site C (PCBs)	39	6.2 to 1,160	3802%	524% to 115,916%

1. Variability measured as ratio of maximum to minimum-reported concentration of the contaminant within (intra-sample) and between co-located (inter-sample) discrete samples collected around grid points. Refer to summary tables in Part 1 for noted study site.
2. Estimated, median Relative Percent Difference between estimated minimum and maximum concentrations of discrete samples collected within a 0.5m radius of a grid point.

Table 6-1. Range and sum of intra-sample and inter-sample Relative Standard Deviation (RSD) of discrete sample variability around individual grid points.

	[1]Intra-Sample Data			[2]Inter-Sample Data			[3]Sum Intra- and Inter-Sample Data		
Study Site	**Mean RSD**	**Median RSD**	**Range RSD**	**Mean RSD**	**Median RSD**	**Range RSD**	**Mean RSD**	**Median RSD**	**Range RSD**
Site A (arsenic)	12%	12%	4.8% to 30%	14%	12%	1.5% to 38%	27%	24%	9%-52%
Site B (lead)	40%	34%	20% to 96%	30%	26%	11% to 81%	70%	64%	44%-139%
Site C (PCBs)	72%	57%	17% to 277%	72%	56%	14% to 151%	144%	126%	58%-336%

1. Refer to Tables 4-3, 4-10 and 4-17 in Part 1.
2. Refer to Tables 4-6, 4-13 and 4-19 in Part 1.
3. Refer to Tables 4-7, 4-14 and 4-20 in Part 1.

Table 7-1. Statistical analysis of twenty sets of random arsenic data for each of ten randomly selected grid points at Study Site A. One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	174	559	355	137	0.4	39%	435	C
2	191	677	378	145	0.4	38%	462	C
3	172	528	324	136	3.2	42%	403	C
4	144	677	326	188	2.2	58%	435	C
5	144	572	403	161	2.9	40%	496	C
6	158	719	379	201	2.5	53%	495	C
7	207	740	374	157	1.7	42%	495	C
8	165	695	369	201	1.7	55%	486	C
9	185	691	367	162	2.1	44%	461	C
10	144	674	341	176	1.3	51%	443	C
11	204	591	382	129	1.9	34%	457	C
12	161	884	420	257	2.4	61%	569	C
13	165	563	316	154	2.7	49%	405	C
14	169	695	364	172	3.0	47%	463	C
15	169	615	372	162	3.0	44%	466	C
16	161	633	332	152	1.7	46%	420	C
17	201	1412	512	341	2.2	67%	776	C
18	193	567	339	136	3.1	40%	418	C
19	165	815	406	225	2.6	55%	537	C
20	161	691	388	194	1.7	50%	501	C
Minimum:			316			34%	403	
Maximum:			512			67%	776	
Median:			370			46%	463	
Mean:			372			48%	481	
SD:			44					
RSD:			12%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-2. Statistical analysis of twenty sets of random lead data for each of ten randomly selected grid points at Study Site B. One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	67	642	241	162	0.6	67%	335	C
2	101	387	242	105	0.4	43%	303	C
3	85	525	228	130	3.2	57%	304	C
4	139	677	333	160	2.2	48%	425	C
5	188	349	260	52	2.9	20%	290	C
6	89	654	321	155	2.5	48%	411	C
7	38	578	233	181	1.7	78%	338	C
8	60	734	245	197	1.7	81%	359	C
9	56	734	266	184	2.1	69%	438	C
10	84	679	327	193	1.3	59%	439	C
11	56	291	177	81	1.9	46%	224	C
12	19	799	252	214	2.4	85%	376	C
13	32	253	159	72	2.7	45%	201	C
14	32	525	254	169	3.0	67%	352	C
15	32	677	263	203	3.0	77%	381	C
16	19	337	195	98	1.7	50%	252	C
17	86	812	262	205	2.2	79%	425	C
18	168	677	311	171	3.1	55%	410	C
19	40	703	222	191	2.6	86%	333	C
20	19	353	194	131	1.7	67%	270	C
Minimum:			159			20%	201	
Maximum:			333			86%	439	
Median:			248			63%	345	
Mean:			249			61%	343	
SD:			48					
RSD:			19%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-3. Statistical analysis of twenty sets of random PCB data for each of ten randomly selected grid points at Study Site C. One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	0.14	96.00	23	35	1.5	152%	103	A
2	0.07	100	14	31	2.1	214%	58	A
3	0.18	10,000	1,003	3,161	3.2	315%	1,006,112	B
4	0.08	1,400	200	435	2.2	217%	1,333	A
5	0.18	6,700	720	2,102	2.9	292%	5,310	A
6	0.25	3,100	388	970	2.5	250%	2,421	A
7	0.02	130	23	41	1.7	174%	79	D
8	0.24	480	84	147	1.7	175%	352	A
9	0.43	1,000	149	308	2.1	207%	695	A
10	0.02	180	47	58	1.3	124%	80	D
11	0.15	110	17	33	1.9	194%	67	A
12	0.10	920	120	287	2.4	240%	766	A
13	0.21	21	5.5	7	2.7	124%	9.4	C
14	0.17	10,000	1,025	3,154	3.0	308%	10,948	D
15	0.08	6,700	707	2,107	3.0	298%	5,638	A
16	0.18	140	27	46	1.7	171%	111	A
17	0.17	1,200	166	373	2.2	224%	923	A
18	0.02	6,800	688	2,147	3.1	312%	3,648	D
19	0.22	6,700	810	2,091	2.6	258%	11,625	B
20	0.32	230	43	73	1.7	171%	155	A
Minimum:			5.5			124%	9.4	
Maximum:			1,025			315%	1,006,112	
Median:			134			216%	730	
Mean:			313			221%	52,522	
SD:			364					
RSD:			116%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-4a. Statistical analysis of twenty sets of random arsenic data for each of the twelve grid points at Study Site A (Set A). One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	177	615	354	144	0.4	41%	428	C
2	176	633	339	134	0.4	39%	408	C
3	177	656	349	156	0.4	45%	430	C
4	176	587	333	132	0.4	40%	401	C
5	144	633	344	153	0.4	44%	423	C
6	177	633	349	142	0.4	41%	423	C
7	181	546	349	132	0.4	38%	417	C
8	180	601	362	151	0.4	42%	440	C
9	144	642	328	150	0.5	46%	405	C
10	168	528	326	117	0.4	36%	387	C
11	144	615	340	160	0.5	47%	423	C
12	195	546	338	124	0.4	37%	402	C
13	180	656	331	141	0.4	43%	405	C
14	178	721	373	186	0.5	50%	470	C
15	186	587	339	134	0.4	40%	409	C
16	180	587	361	147	0.4	41%	437	C
17	178	633	356	141	0.4	39%	429	C
18	155	572	344	133	0.4	39%	413	C
19	155	601	325	138	0.4	42%	396	C
20	178	591	335	133	0.4	40%	403	C
Minimum:			325			36%	387	
Maximum:			373			50%	470	
Median:			342			41%	415	
Mean:			344			41%	417	
SD:			13					
RSD:			3.8%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-4b. Statistical analysis of twenty sets of random arsenic data for each of the twelve grid points at Study Site A (Set B). One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	172	765	373	196	0.5	53%	475	C
2	165	873	394	213	0.5	54%	504	C
3	144	884	392	227	0.6	58%	510	C
4	158	733	377	200	0.5	53%	481	C
5	169	873	408	214	0.5	53%	519	C
6	176	765	371	169	0.5	46%	459	C
7	165	695	369	190	0.5	52%	467	C
8	165	815	402	208	0.5	52%	510	C
9	172	554	356	138	0.4	39%	427	C
10	158	876	404	252	0.6	62%	535	C
11	165	815	389	206	0.5	53%	495	C
12	172	740	378	201	0.5	53%	482	C
13	172	740	385	209	0.5	54%	540	C
14	144	719	350	172	0.5	49%	440	C
15	172	554	338	134	0.4	40%	407	C
16	175	884	377	210	0.6	56%	486	C
17	165	637	343	151	0.4	44%	421	C
18	158	554	338	157	0.5	47%	419	C
19	144	873	407	242	0.6	59%	533	C
20	144	674	356	173	0.5	49%	445	C
Minimum:			338			39%	407	
Maximum:			408			62%	540	
Median:			377			53%	482	
Mean:			375			51%	478	
SD:			23					
RSD:			6.1%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-5a. Statistical analysis of twenty sets of random lead data for each of the twelve grid points at Study Site B (Set A). One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	101	298	201	53	0.3	26%	228	C
2	55	387	218	96	0.4	44%	268	C
3	164	320	238	57	0.2	24%	267	C
4	38	457	257	100	0.4	39%	309	C
5	41	578	260	150	0.6	58%	337	C
6	65	578	247	126	0.5	51%	312	C
7	101	412	249	90	0.4	36%	296	C
8	84	539	228	130	0.6	57%	296	C
9	85	578	261	129	0.5	49%	328	C
10	85	539	253	119	0.5	47%	315	C
11	164	539	266	111	0.4	42%	337	A
12	55	431	278	111	0.4	40%	335	C
13	65	431	267	95	0.4	35%	316	C
14	41	396	212	103	0.5	48%	266	C
15	84	598	260	141	0.5	54%	333	C
16	55	431	239	112	0.5	47%	297	C
17	113	457	287	93	0.3	32%	335	C
18	101	558	261	125	0.5	48%	326	C
19	94	686	253	159	0.6	63%	335	C
20	55	539	250	135	0.5	54%	321	C
Minimum:			201			24%	228	
Maximum:			287			63%	337	
Median:			253			47%	316	
Mean:			249			45%	308	
SD:			22					
RSD:			8.6%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-5b. Statistical analysis of twenty sets of random lead data for each of the twelve grid points at Study Site B (Set B). One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	60	596	268	169	0.6	63%	356	C
2	40	596	246	156	0.6	64%	327	C
3	103	799	292	226	0.8	77%	442	B
4	62	681	254	187	0.7	74%	351	C
5	32	642	266	206	0.8	77%	373	C
6	67	659	295	197	0.7	67%	397	C
7	40	654	256	204	0.8	80%	362	C
8	64	734	276	220	0.8	80%	444	C
9	56	681	265	187	0.7	70%	362	C
10	40	1,014	300	304	1.0	102%	567	C
11	40	703	270	214	0.8	79%	444	A
12	62	723	238	203	0.8	85%	402	A
13	63	679	256	166	0.6	65%	342	C
14	67	659	270	173	0.6	64%	360	C
15	67	350	220	96	0.4	44%	270	C
16	55	799	276	241	0.9	87%	401	C
17	67	723	276	199	0.7	72%	379	C
18	56	654	286	157	0.5	55%	368	C
19	55	703	310	202	0.7	65%	474	A
20	30	681	291	207	0.7	71%	398	C
Minimum:			220			44%	270	
Maximum:			310			102%	567	
Median:			270			72%	376	
Mean:			271			72%	391	
SD:			22					
RSD:			8.1%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-6a. Statistical analysis of twenty sets of random total PCB data for each of the twelve grid points at Study Site C (Set A). One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	0.09	5,700	716	1,736	2.4	242%	23,140	B
2	0.14	19,000	1,826	5,463	3.0	299%	233,119	B
3	0.3	110	1,058	2,437	2.3	230%	82,102	B
4	0.25	11,000	1,197	3,212	2.7	268%	73,759	B
5	0.27	2,600	355	811	2.3	228%	2,684	D
6	0.25	5,700	825	1,897	2.3	230%	5,090	A
7	0.19	10,000	918	2,869	3.1	313%	2,405	C
8	0.15	6,800	920	2,061	2.2	224%	8,222	B
9	0.33	10,000	1,003	2,862	2.9	285%	20,117	B
10	0.26	10,000	927	2,867	3.1	309%	42,432	B
11	0.21	1,400	172	396	2.3	230%	735	A
12	0.15	10,000	961	2,861	3.0	298%	23,879	B
13	0.14	6,700	841	2,046	2.4	243%	30,765	B
14	0.09	3,900	426	1,118	2.6	263%	2,263	A
15	0.14	19,000	1,680	5,459	3.2	325%	116,807	B
16	0.21	3,100	544	1,195	2.2	220%	12,532	B
17	0.26	19,000	1,876	5,469	2.9	292%	148,540	B
18	0.17	11,000	1,152	3,189	2.8	277%	74,954	B
19	0.27	11,000	1,223	3,174	2.6	260%	14,671	B
20	0.19	11,000	1,009	3,157	3.1	313%	72,261	B
Minimum:			172			220%	735	
Maximum:			1,876			325%	233,119	
Median:			944			266%	23,510	
Mean:			981			267%	49,524	
SD:			448					
RSD:			46%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-6b. Statistical analysis of twenty sets of random total PCB data for each of the twelve grid points at Study Site C (Set B). One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	0.04	790	78	225	2.9	289%	360	A
2	0.07	590	65	167	2.6	258%	151	C
3	0.1	920	96	261	2.7	271%	464	A
4	0.07	770	75	219	2.9	292%	704	D
5	0.02	660	71	187	2.6	264%	324	A
6	0.02	790	78	225	2.9	290%	379	A
7	0.04	660	67	187	2.8	280%	300	A
8	0.08	890	91	253	2.8	278%	422	A
9	0.02	960	95	273	2.9	289%	473	A
10	0.08	940	88	269	3.1	307%	859	D
11	0.08	890	90	253	2.8	283%	418	A
12	0.04	590	64	167	2.6	261%	279	A
13	0.04	770	80	219	2.7	274%	371	A
14	0.04	590	68	166	2.4	245%	294	A
15	0.05	960	100	273	2.7	274%	466	A
16	0.04	920	91	262	2.9	287%	443	A
17	0.10	660	68	187	2.8	276%	309	A
18	0.02	960	105	272	2.6	259%	483	A
19	0.04	890	90	253	2.8	280%	411	A
20	0.08	1,500	160	427	2.7	267%	774	A
Minimum:			64			245%	151	
Maximum:			160			307%	859	
Median:			84			277%	415	
Mean:			86			276%	434	
SD:			21					
RSD:			25%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-7. Statistical analysis of twenty sets of random arsenic data for each of the twenty-four grid points at Study Site A. One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	172	765	367	170	0.5	46%	427	C
2	165	873	376	180	0.5	48%	439	C
3	144	884	372	193	0.5	52%	452	A
4	158	733	361	169	0.5	47%	420	C
5	144	873	379	186	0.5	49%	444	C
6	176	765	371	159	0.4	43%	426	C
7	165	695	361	161	0.4	45%	417	C
8	165	815	383	180	0.5	47%	446	C
9	144	642	348	148	0.4	42%	400	C
10	158	876	372	200	0.5	54%	442	C
11	144	815	366	182	0.5	50%	430	C
12	172	740	363	169	0.5	46%	422	C
13	172	740	368	181	0.5	49%	442	C
14	144	721	366	178	0.5	49%	429	C
15	172	587	345	134	0.4	39%	400	C
16	175	884	371	178	0.5	48%	433	C
17	165	637	351	144	0.4	41%	402	C
18	155	572	345	143	0.4	42%	395	C
19	144	873	370	198	0.5	54%	439	C
20	144	674	352	157	0.4	45%	407	C
Minimum:			345			39%	395	
Maximum:			383			54%	452	
Median:			367			47%	428	
Mean:			364			47%	426	
SD:			11					
RSD:			3.0%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-8. Statistical analysis of twenty sets of random lead data for each of the twenty-four grid points at Study Site B. One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	60	596	235	127	0.5	54%	347	D
2	40	596	236	125	0.5	53%	280	C
3	103	799	263	165	0.6	63%	321	C
4	38	681	257	146	0.6	57%	308	C
5	32	642	261	177	0.7	68%	348	A
6	65	659	273	164	0.6	60%	330	C
7	40	654	250	156	0.6	62%	322	A
8	64	734	248	180	0.7	73%	321	C
9	56	681	267	153	0.6	57%	335	A
10	40	1,014	281	224	0.8	80%	374	A
11	40	703	264	169	0.6	64%	335	A
12	55	723	255	163	0.6	64%	312	D
13	63	679	262	132	0.5	50%	309	C
14	41	659	247	138	0.6	56%	295	C
15	67	598	242	119	0.5	49%	284	C
16	55	799	255	186	0.7	73%	320	C
17	67	723	277	155	0.6	56%	331	C
18	56	654	277	137	0.5	50%	337	A
19	55	703	281	180	0.6	64%	357	A
20	30	681	270	173	0.6	64%	394	A
Minimum:			235			49%	280	
Maximum:			281			80%	394	
Median:			262			61%	326	
Mean:			260			61%	328	
SD:			14					
RSD:			5.5%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-9. Statistical analysis of twenty sets of random PCB data for each of the twenty-four grid points at Study Site C. One of ten "intra-sample" data points randomly selected for each grid point.

Sample Set	Min (mg/kg)	Max (mg/kg)	Mean (mg/kg)	SD	Coeff Var	RSD	95% UCL (mg/kg)	UCL Type
1	0.04	5,700	397	1,254	3.3	316%	2,943	D
2	0.07	19,000	945	3,885	4.1	411%	8,837	D
3	0.1	920	577	1,764	3.1	306%	4,161	D
4	0.07	11,000	636	2,299	3.6	362%	5,305	D
5	0.02	2,600	213	593	2.8	279%	1,418	D
6	0.02	5,700	451	1,375	3.0	305%	3,244	D
7	0.04	10,000	492	2,035	4.1	413%	4,226	D
8	0.08	6,800	505	1,497	3.0	296%	3,546	D
9	0.02	10,000	549	2,042	3.7	372%	4,696	D
10	0.08	10,000	507	2,037	4.0	402%	4,644	D
11	0.08	1,400	131	328	2.5	251%	797	D
12	0.04	10,000	512	2,034	4.0	397%	4,644	D
13	0.04	6,700	460	1,475	3.2	320%	3,456	D
14	0.04	3,900	247	803	3.3	325%	652	A
15	0.05	19,000	890	3,865	4.3	434%	8,740	D
16	0.04	3,100	317	877	2.8	276%	2,099	D
17	0.10	19,000	972	3,896	4.0	401%	8,884	D
18	0.02	11,000	629	2,277	3.6	362%	5,253	D
19	0.04	11,000	657	2,277	3.5	347%	5,281	D
20	0.08	11,000	584	2,245	3.8	384%	5,145	D
Minimum:			131			251%	652	
Maximum:			972			434%	8,884	
Median:			510			354%	4,435	
Mean:			534			348%	4,399	
SD:			224					
RSD:			42%					

SD: Standard Deviation; Coeff Var: Coefficient of Variation; RSD: Relative Standard Deviation; UCL: Upper Confidence Level.

UCL Types

A	95% Adjusted Gamma UCL
B	95% Hall's Bootstrap UCL
C	95% Student's-t UCL
D	95% Chebyshev (Mean, Sd) UCL

Table 7-10. Range of mean contaminant concentration calculated for random sets of discrete samples at each study site.

Study Site	[1]Ten-Point Data Sets			[2]Twenty Four-Point Data Sets			[3]MIS Triplicate data		
	Average Mean (mg/kg)	**Median Mean (mg/kg)**	**Range Mean (mg/kg)**	**Average Mean (mg/kg)**	**Median Mean (mg/kg)**	**Range Mean (mg/kg)**	**Mean (mg/kg)**	**Range (mg/kg)**	**95% UCL (mg/kg)**
Site A (arsenic)	372	370	316-512	364	367	345-383	233 (302)	220-250 (288-328)	259 (339)
Site B (lead)	249	248	159-333	260	262	235-281	287 (268)	240-350 (223-326)	383 (357)
Site C (PCBs)	313	114	5.5-1,025	534	510	131-972	104	19-270	467

1. Refer to Tables 7-1, 7-2 and 7-3.

2. Refer to Tables 7-7, 7-8 and 7-9.

3. Refer to Part 1, Table 5-5. Adjustment of MIS Method 6810B data to reflect average increase (Study Site arsenic, +31%) or decrease (Study Site B lead, -6.8%) in concentrations reported for discrete sample XRF data noted in parentheses. Student's t 95% UCL indicated for Study Site A (RSD 6.5%) and study Site B (RSD 20%); Chebyshev 95% UCL indicated for Study Site C (RSD 138%).

Table 7-11. Range of Relative Standard Deviation (RSD) of calculated mean for random sets of discrete samples at each study site.

Study Site	[1]Ten-Point Data Sets			[2]Twenty Four-Point Data Sets			[3]MIS data
	Mean RSD	**Median RSD**	**Range RSD**	**Mean RSD**	**Median RSD**	**Range RSD**	**RSD**
Site A (arsenic)	48%	46%	34%-67%	47%	47%	39%-54%	6.50%
Site B (lead)	61%	63%	20%-86%	61%	61%	49%-80%	20%
Site C (PCBs)	221%	216%	124%-315%	348%	354%	251%-434%	138%

1. Refer to Tables 7-1, 7-2 and 7-3.
2. Refer to Tables 7-7, 7-8 and 7-9.
3. Refer to Part 1, Table 5-5.

Figure 2-1. Random, small-scale, distributional heterogeneity in a jar of colored gumballs.

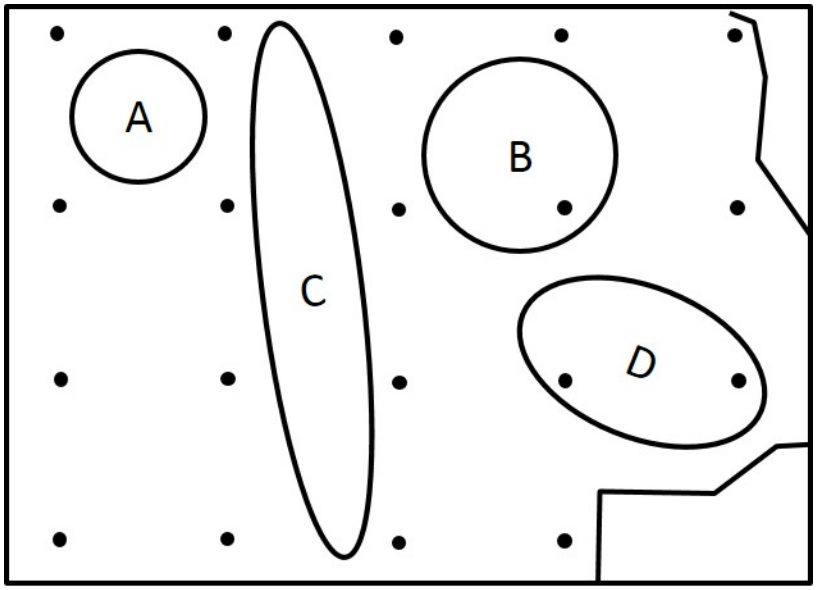


Figure 3-1. Discrete sampling grid designated for a site under investigation overlain with hypothetical, “hot spots” superimposed (USEPA 1989). Under this approach a single, discrete soil sample was assumed to be adequate to identify large areas of contamination above potential levels of concern.

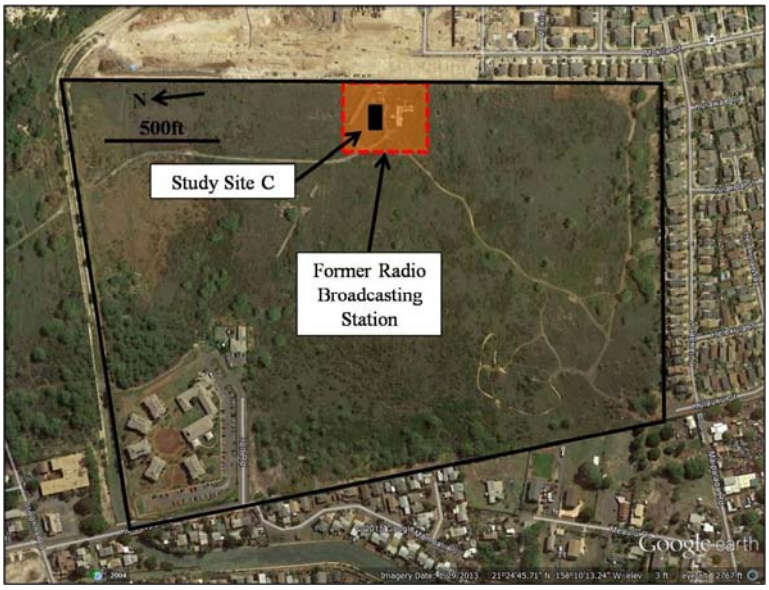


Figure 3-2. Large-scale area of PCB contaminated soil at Study Site C identified within the 89-acre site using decision unit and Multi Increment investigation methods. Additional sampling underway to provide better resolution of PCB distribution within identified "hot area."

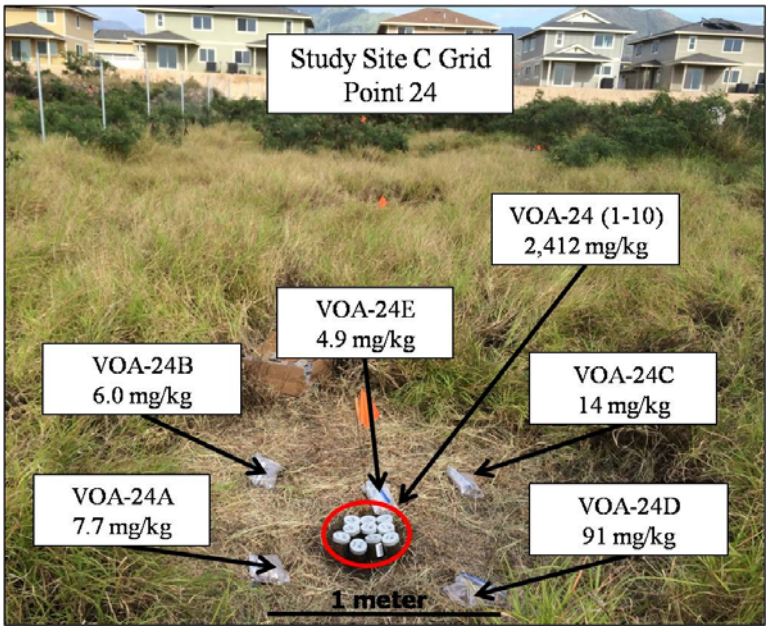


Figure 3-3. Random, small-scale "hot spot" identified within the one meter-square area of Grid Point 24 at Study Site C (see Figure 3-2; refer also to Table 4-16 and Table 4-18 in Part 1). Actual grid point 24 not depicted in figure. Concentration trends between co-located samples around grid points are random and not related to larger-scale trends of interest.

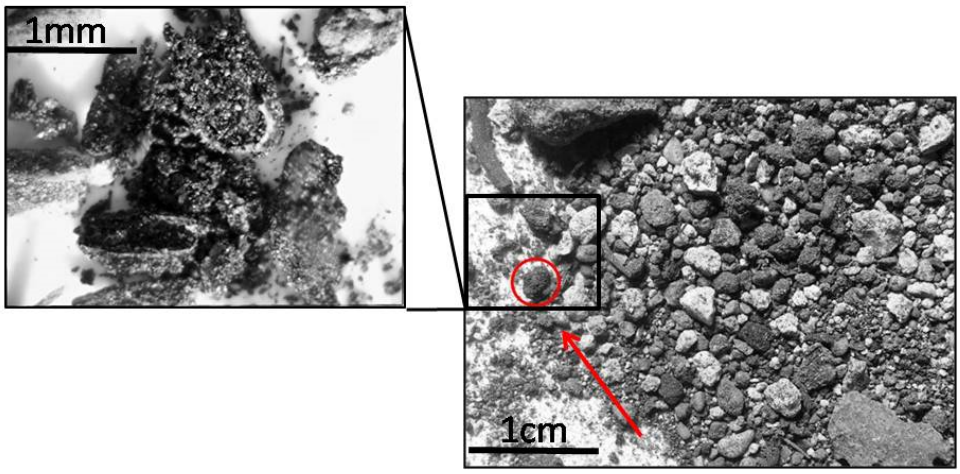


Figure 3-4. Suspect micro-scale “hot spot” of PCB-infused, tarry nugget within Sample VOA-8-12 (8) from Study Site C (see also Figure 5-6 and Figure 5-7 in Part 1). Nugget depicted in photomicrograph on the left is not the nugget shown in the larger-scale photo but is from the same sample.

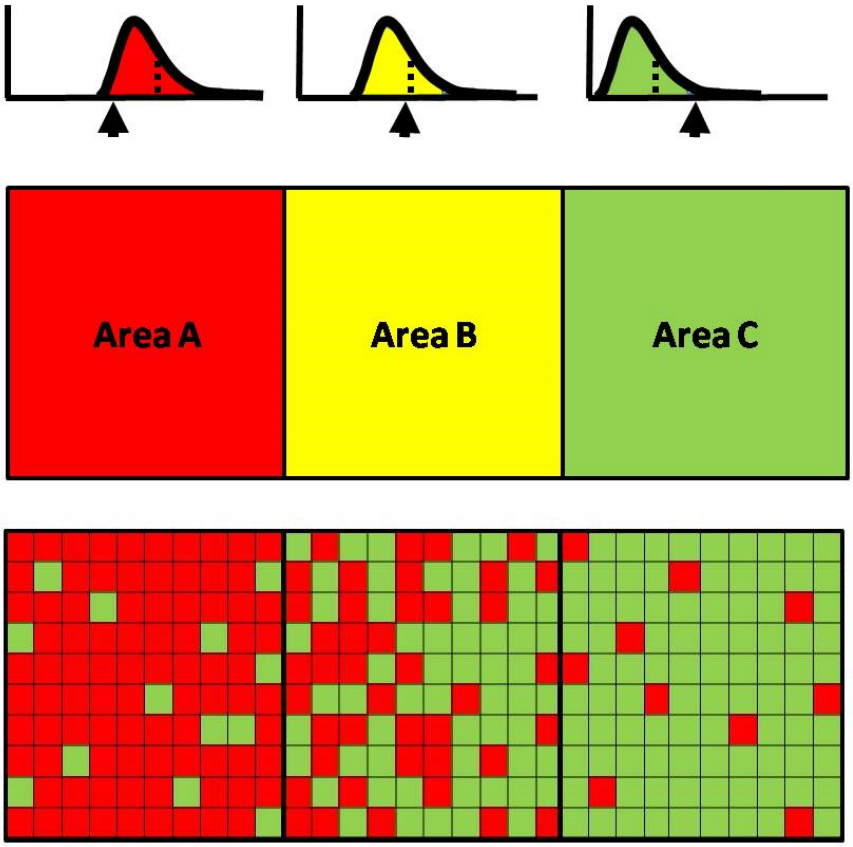


Figure 4-1. Influence of random, small-scale variability on comparison of discrete sample data to target screening level. Area A: Heavily contaminated, contaminant concentration in majority of discrete sample-size masses of soil fall above screening level (mean above screening level); Area B: Moderately contaminated, discrete samples fall both above and below screening level (mean above screening level); Area C: Low contamination, majority of discrete samples fall below screening level (mean below screening level). Discrete sample contaminant concentration frequency graphs noted above areas; dashed line represents mean; arrow represents soil screening level. Lower maps reflect random, small-scale distributional of contaminant in soil at the scale of a discrete sample relative to the target screening level (red = above, green = below). Note scattered "false negatives" in Area A, large number of "false negatives" in Area B and scattered "false positives" in Area C.

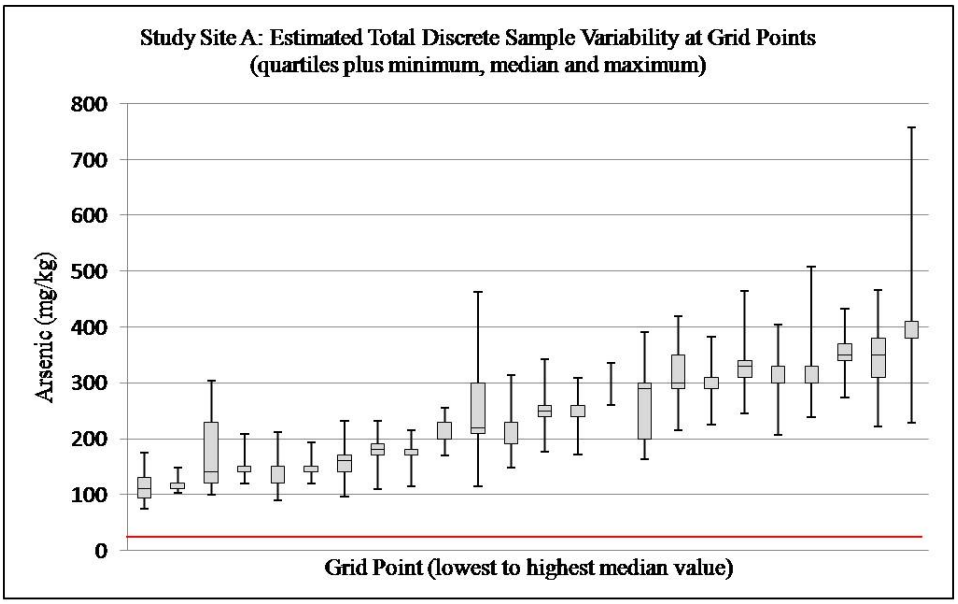


Figure 4-2. Box plots depicting estimated, total variability of total arsenic concentrations in discrete samples within 0.5m of grid points at Study Site A (lowest to highest median values for inter-sample data). Red line denotes HDOH total arsenic screening level of 24 mg/kg. Note relatively low, intra-sample variability of arsenic concentrations comparison to Study Sites B and C. Intra-sample data based on XRF analysis, not directly comparable to Method 6280 data for inter-sample data (XRF consistently higher).

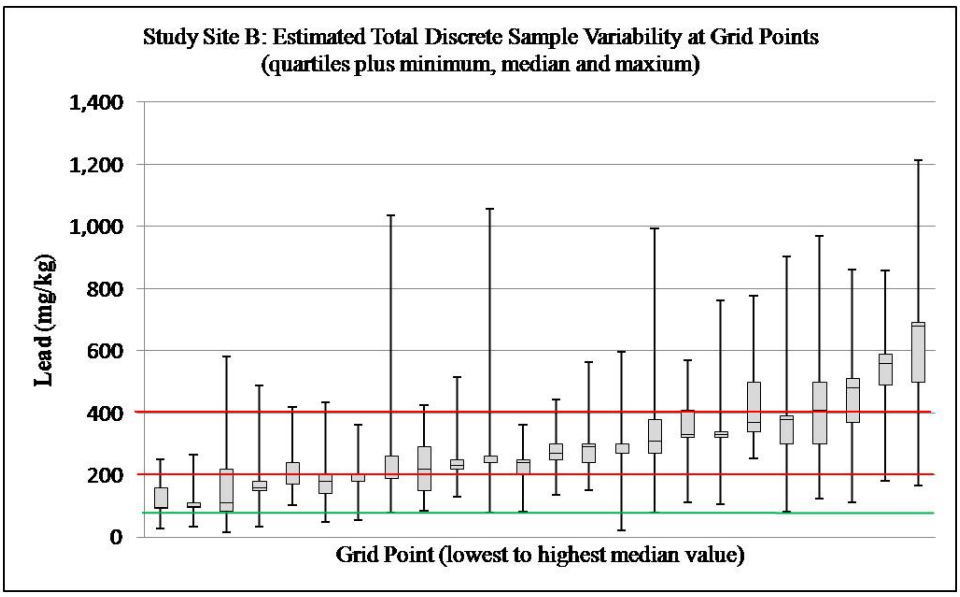


Figure 4-3. Box plots depicting estimated, total variability of lead concentrations in discrete samples within 0.5m of grid points at Study Site B (lowest to highest median for inter-sample data). Estimated range of lead concentrations falls both above and below HDOH residential soil action level of 200 mg/kg at twenty-three of twenty-four grid points and above USEPA residential screening level of 400 mg/kg at twenty of twenty-four points. HDOH default, upper background lead level of 75 mg/kg indicated for reference with full range of lead concentrations points reflecting the presumed mixture of native fill and lead-contaminate ash.

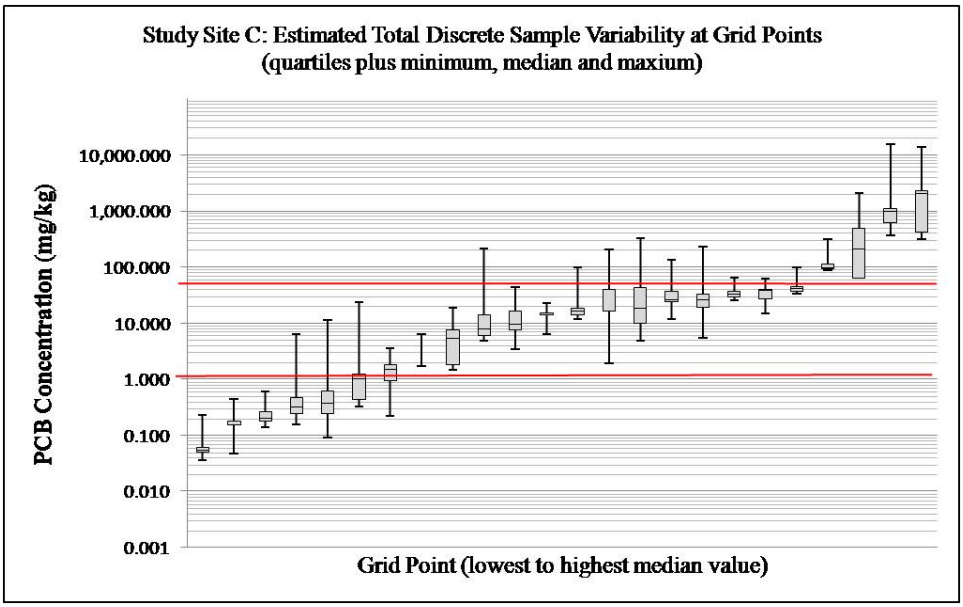


Figure 4-4. Box plots depicting estimated, total variability of total PCB concentrations in discrete samples within 0.5m of grid points at Study Site C (combined intra- and inter-variability; note use of log scale for vertical axis; lowest to highest median values for inter-sample data). HDOH residential PCB soil screening level of 1.1 mg/kg and TSCA level of 50 mg/kg noted for reference.

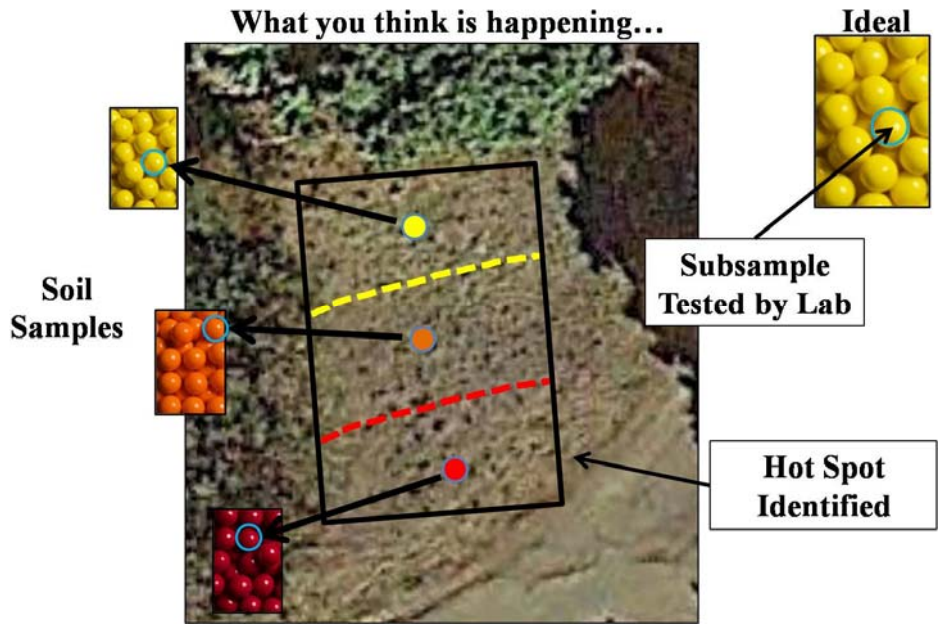


Figure 4-5. Hypothetical pattern of contaminated soil based on discrete sample data and assumption that subsample tested is representative of the sample collected (homogenous gumballs) as well as soil in the area surrounding the sample point.

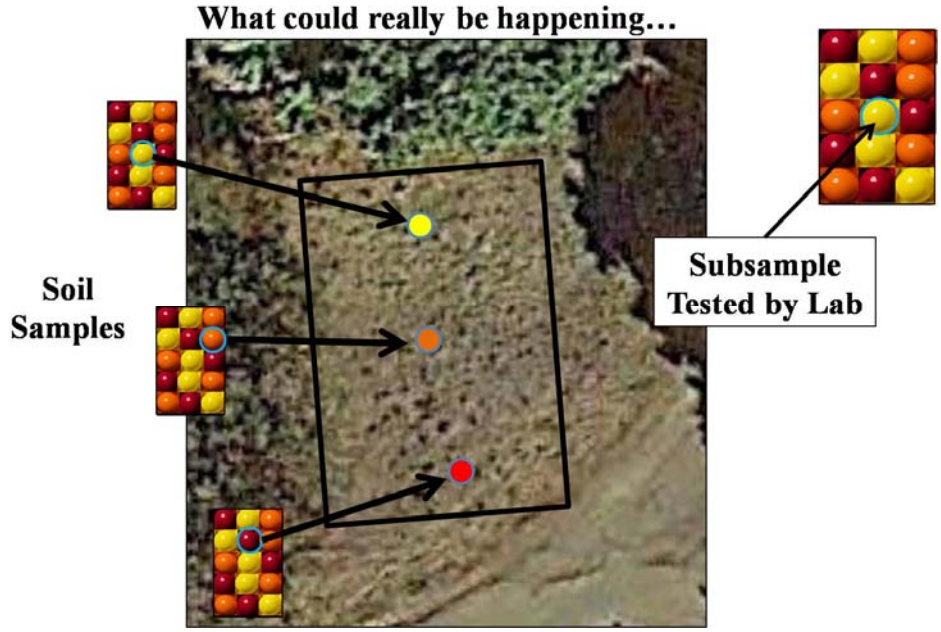


Figure 4-6. Effect of random, small-scale, distributional heterogeneity within discrete soil samples and resulting, erroneous laboratory data for mean concentration of contaminant in sample as a whole. The mean concentration of the contaminant within any given, discrete soil sample randomly collected from the area is in fact identical ("orange").

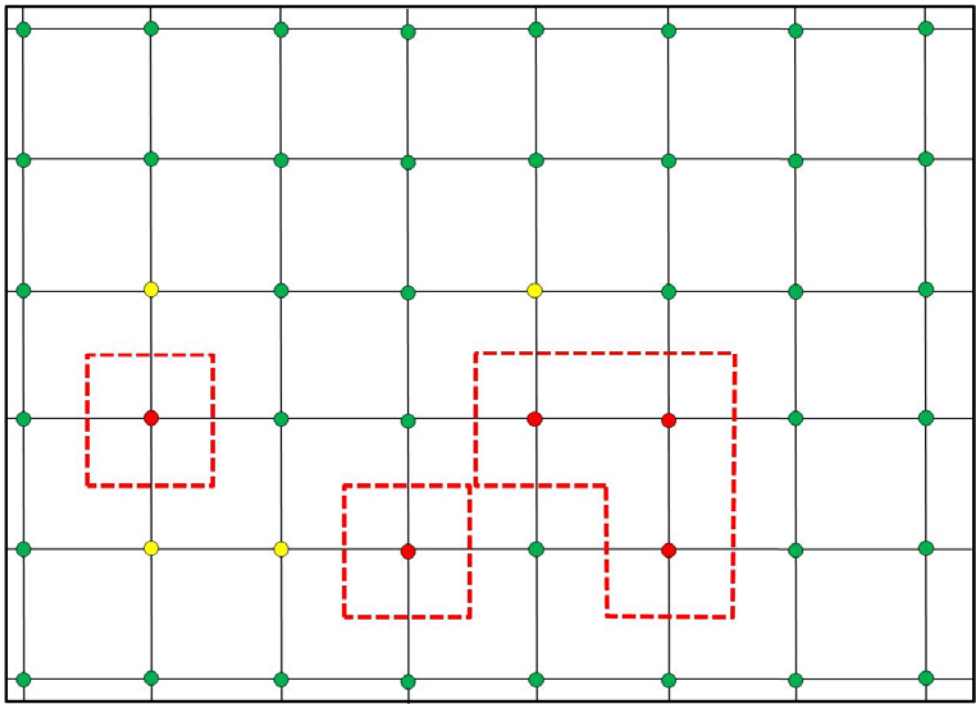


Figure 5-1. Estimated extent of soil contamination for hypothetical site based on closely-spaced discrete samples (red= above screening level; yellow = detected but below screening level; green = not detected). Dashed red lines indicate areas interpreted to require soil removal.

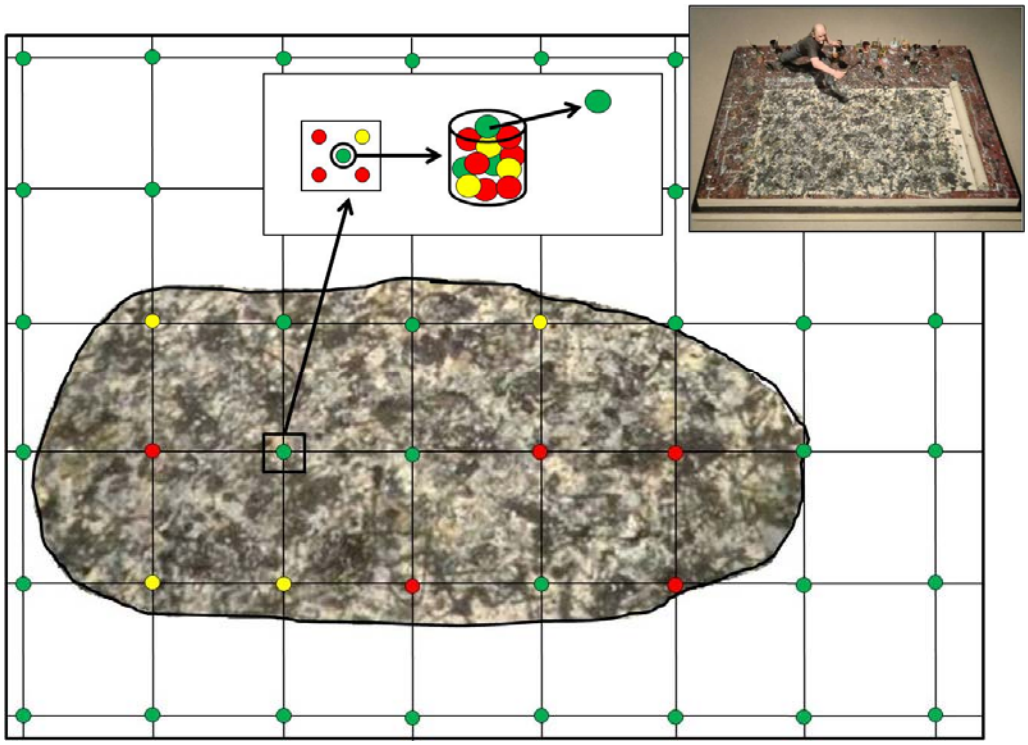


Figure 5-2. Actual extent of “contamination” in Figure 5-1; based on a cutout of the Jackson Pollock painting in Figure 5-8 of Part 1 of the study report.

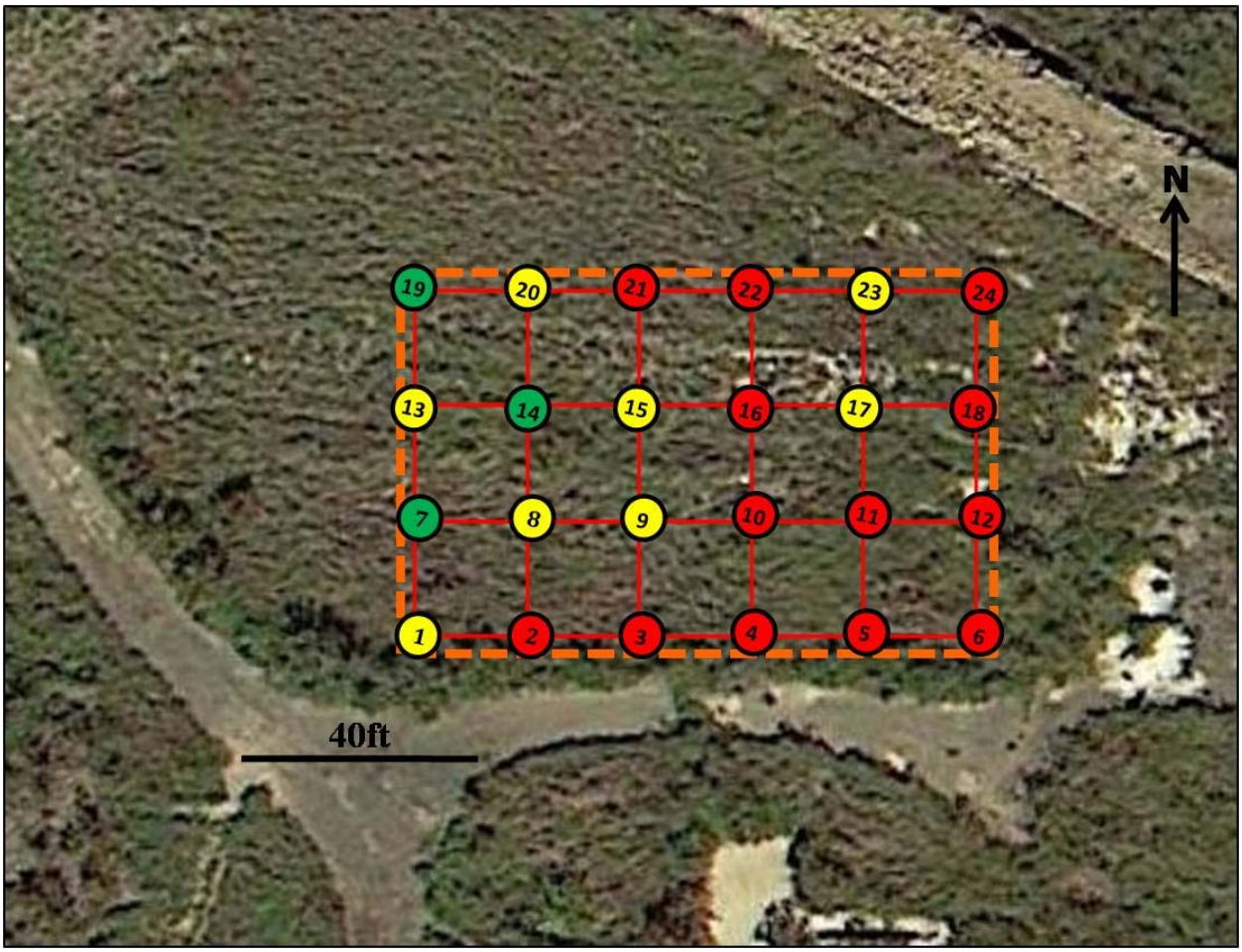


Figure 5.3 Estimated range of total PCB concentrations in discrete samples around individual grid points relative to the HDOH residential soil action level of 1.1 mg/kg (TSCA limit 1.0 mg/kg). Red: All samples likely to fall above action level; Yellow: Samples likely to fall both above and below action level; Green: All samples likely to fall below action level. Yellow areas especially prone to "false negatives" and failed confirmation samples when using discrete sample data for decision making purposes.

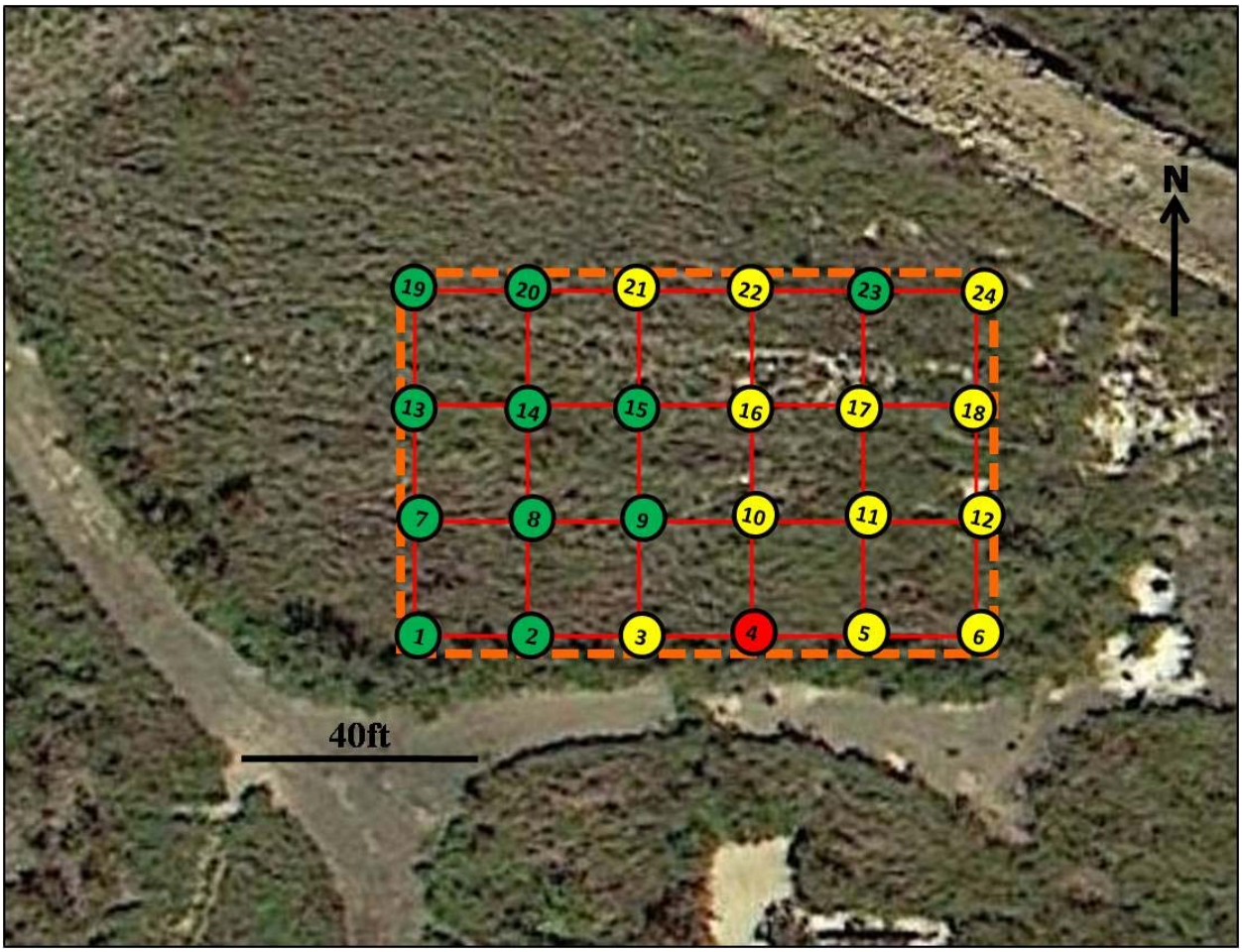


Figure 5.4 Estimated range of total PCB concentrations in discrete samples around individual grid points relative to the TSCA municipal landfill limit of 50 mg/kg. Red: All samples likely to fall above TSCA limit; Yellow: Samples likely to fall both above and below TSCA limit; Green: All samples likely to fall below TSCA limit.

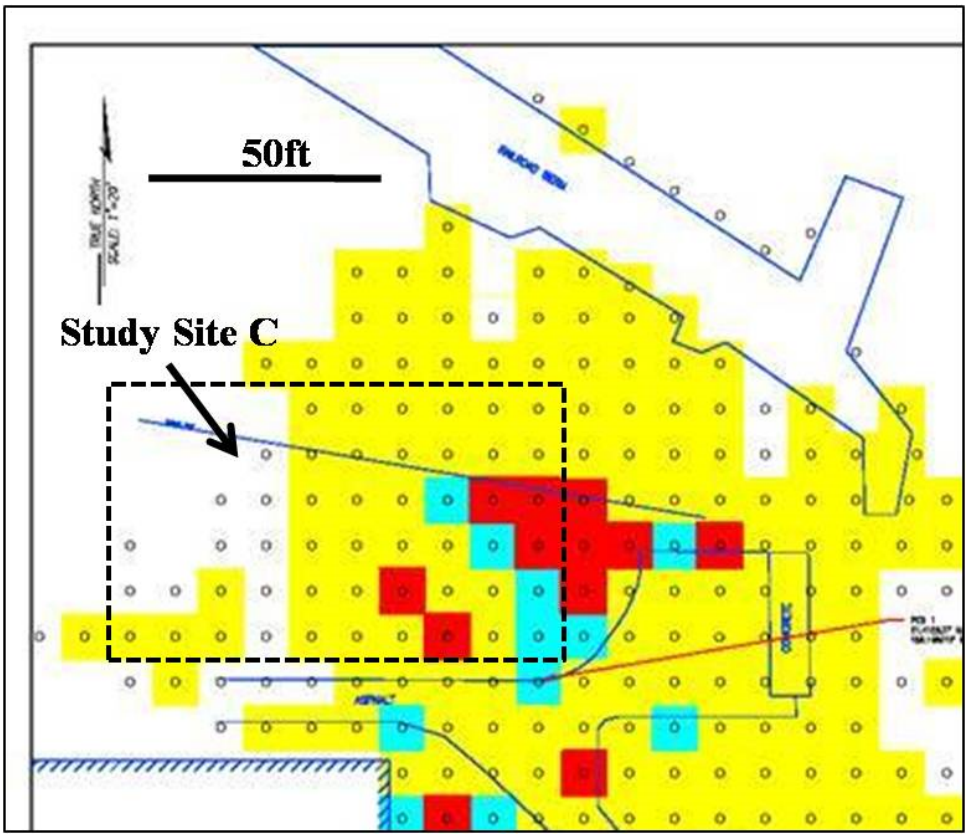


Figure 5-5. Discrete samples with reported concentration of total PCBs greater than 1 mg/kg collected from in the same vicinity as Study Site C in an earlier investigation with depth of impact noted (after USCG 2011); yellow = surface soil; blue = -2ft bgs; red = -4ft bgs. Border areas around marked grid points are highly prone to false negatives; outer areas prone to false positives. Compare to Figure 5-3; note location of potential false negatives.

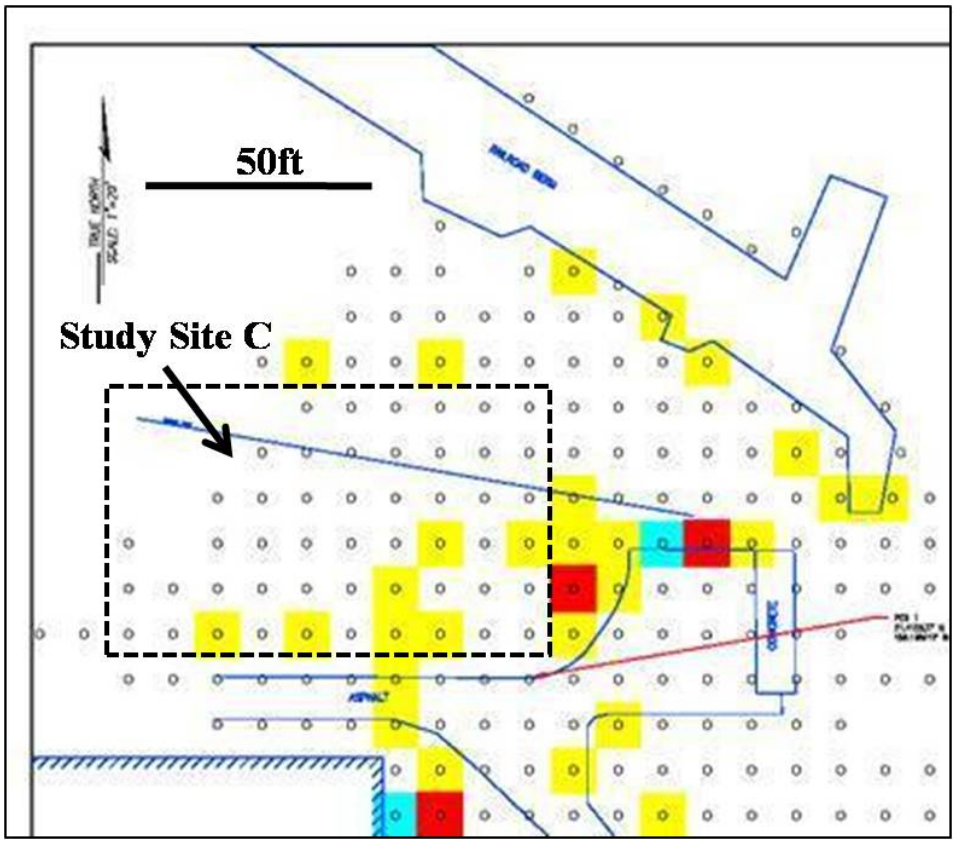


Figure 5-6. Discrete samples with reported concentration of total PCBs greater than 50 mg/kg collected from in the same vicinity as Study Site C in an earlier investigation (after USCG 2011); inset depicts HDOH sample points where PCBs >50 mg/kg reported. Isolated areas of higher PCB concentrations likely reflect random, small-scale, distributional heterogeneity of PCBs within the area as a whole rather than fortuitously identified "hot spots." Compare to Figure 5-4.

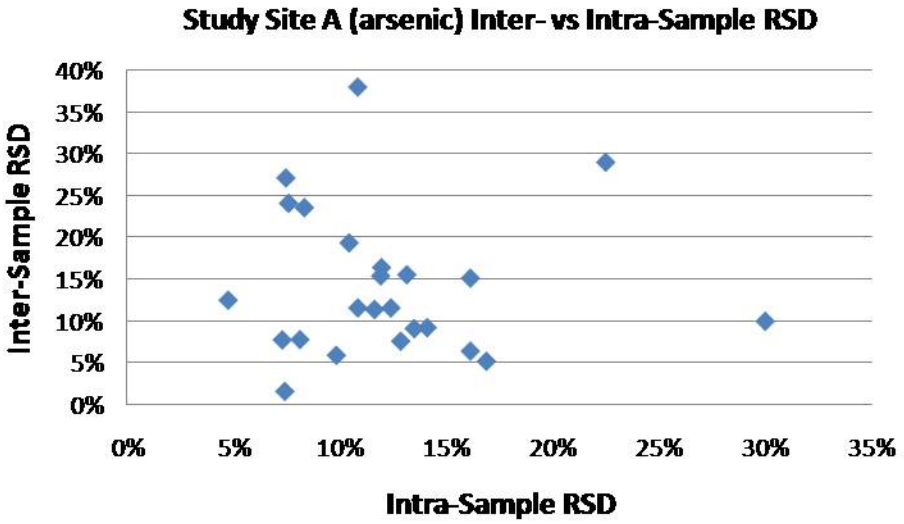


Figure 6-1. Comparison of intra-sample and inter-sample Relative Standard Deviations calculated for individual grid points at Study Site A (arsenic).

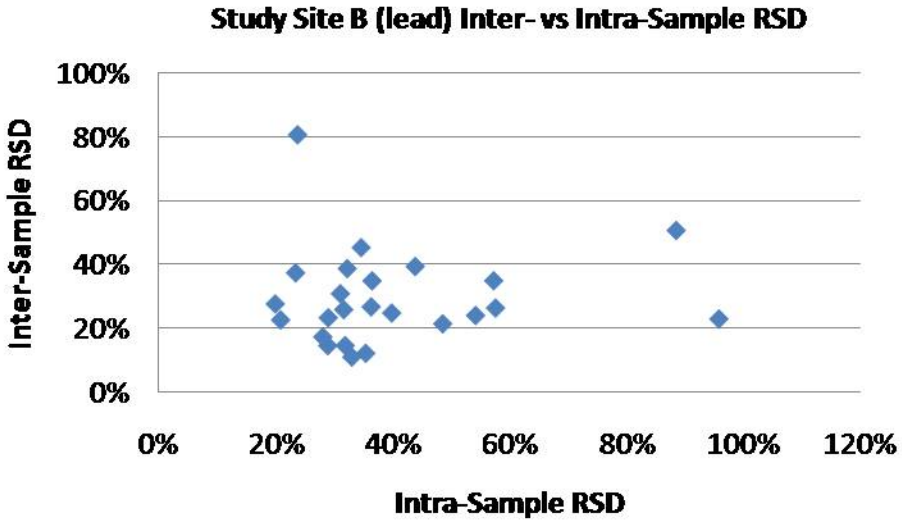


Figure 6-2. Comparison of intra-sample and inter-sample Relative Standard Deviations calculated for individual grid points at Study Site B (lead).

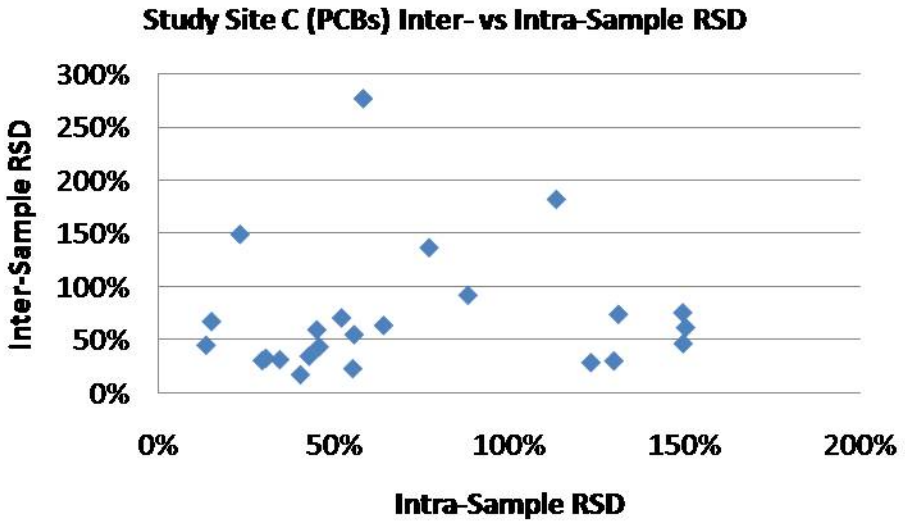


Figure 6-3. Comparison of intra-sample and inter-sample Relative Standard Deviations calculated for individual grid points at Study Site C (total PCBs).

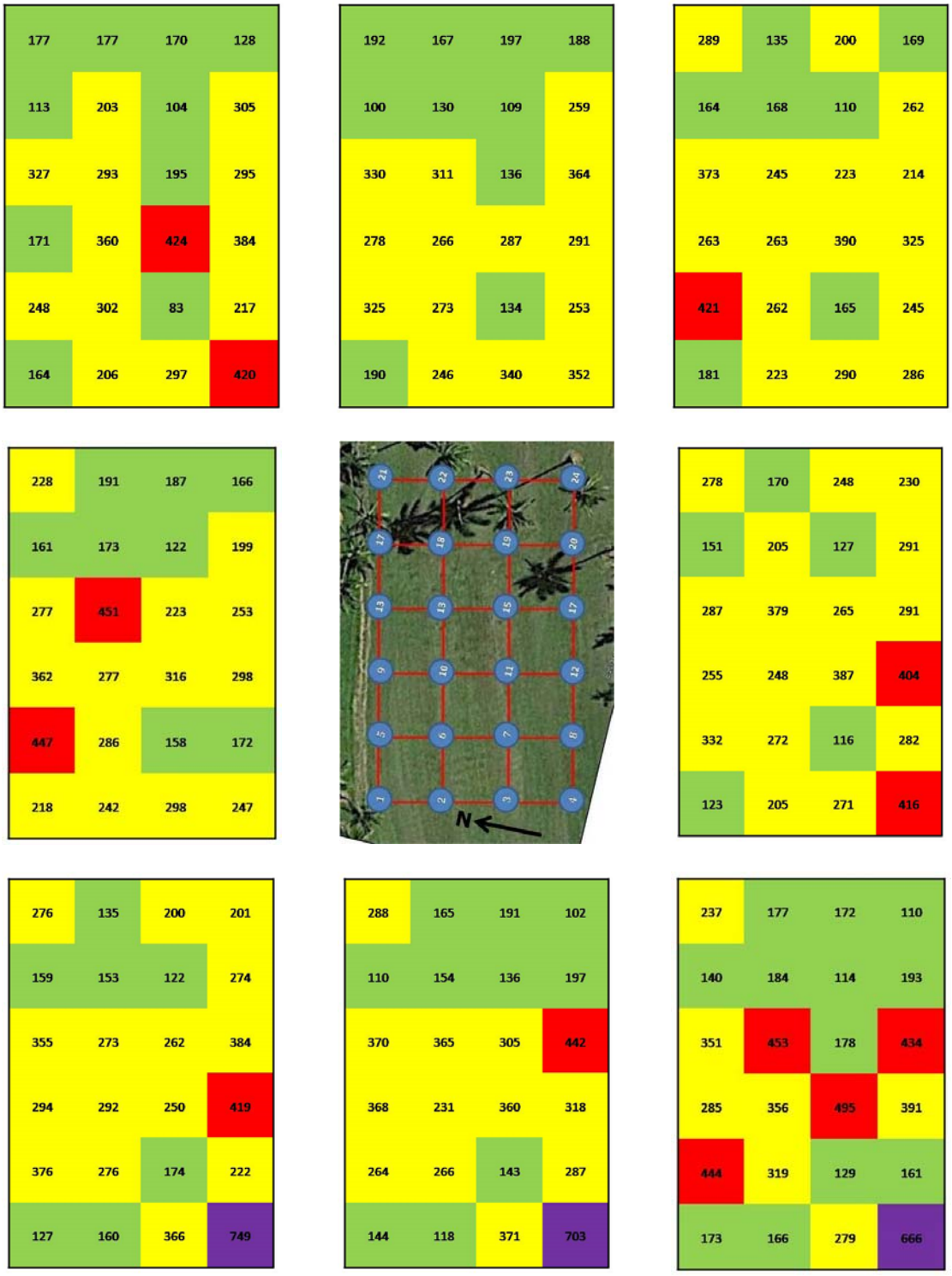


Figure 6-4. Eight artificial, small-scale, patterns of arsenic distribution at Study Site A based on random assignment of a concentration within the minimum and maximum range estimated for each grid point (study area pictured in center; grid area 13,500 ft^2). Patterns reflect hypothetical, independent resampling of the grid points and contrasts in resulting maps. Green <200 mg/kg ; Yellow ≥200 mg/kg; Red ≥ 400 mg/kg; Purple ≥ 600 mg/kg.

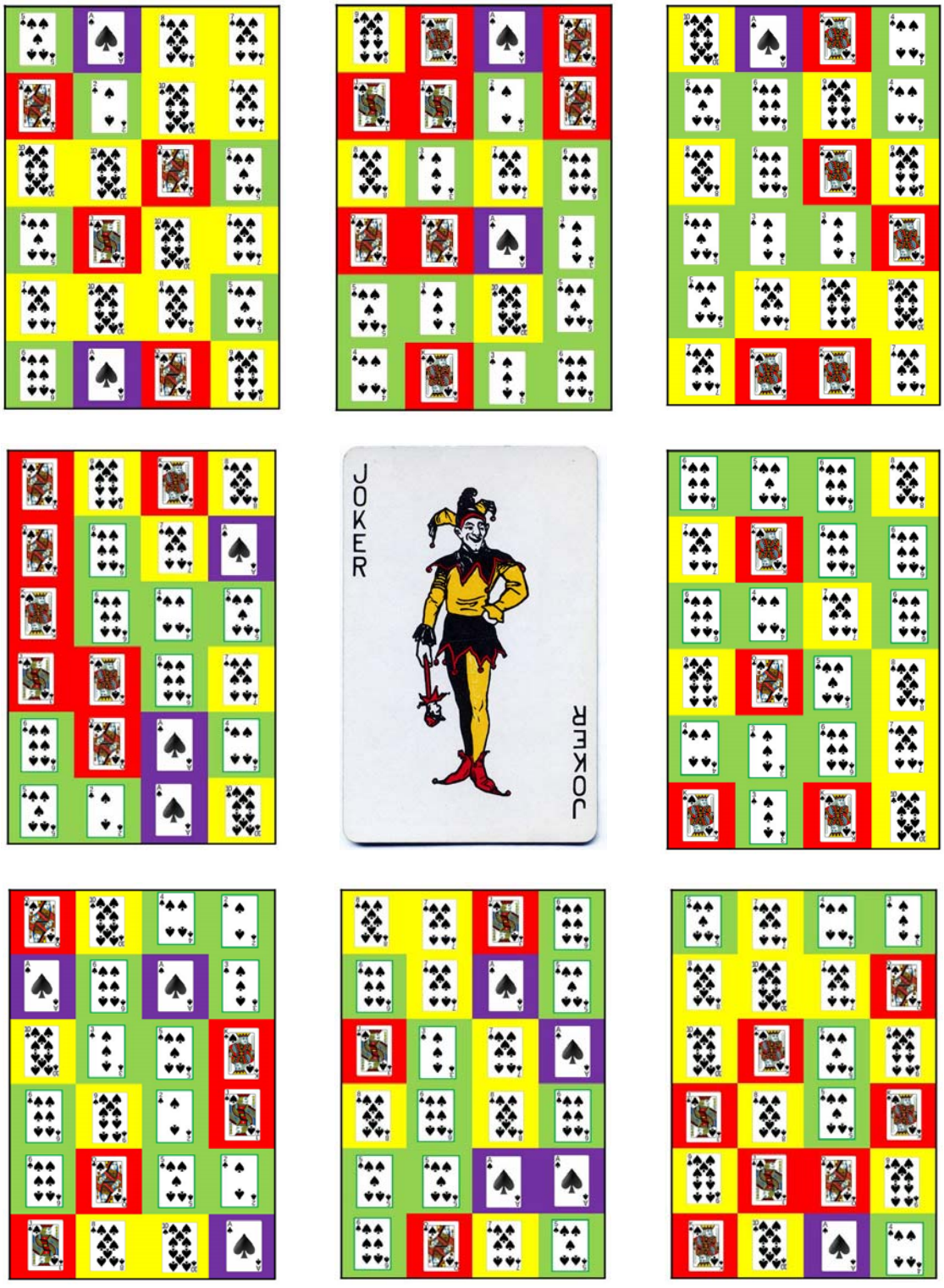


Figure 6-5. Artificial, small-scale patterns generated by random selection of the Ace through King of spades for each of 24 grid points (eight iterations). The true "mean" of each cell and the "study area" as a whole is "7" or "yellow." Green = 2 to 6 cards; Yellow = 7 to 10 cards; Red = face cards (11-13); Purple = Aces.

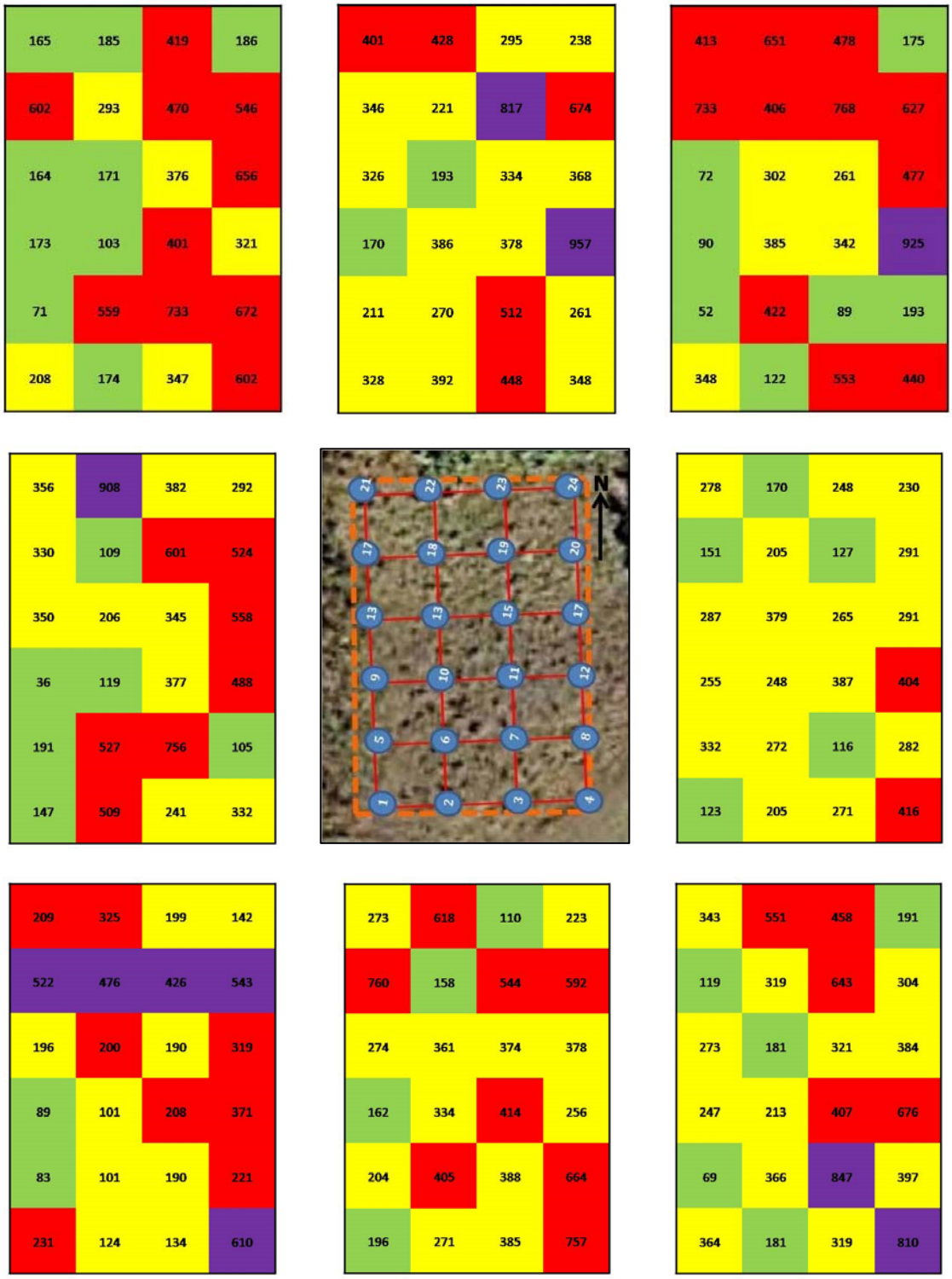


Figure 6-6. Eight artificial, small-scale, pattern of lead distribution at Study Site B based on random assignment of a concentration within the minimum and maximum range estimated for each grid point (study area pictured in center; grid area 1,500 ft^2). Patterns reflect hypothetical, independent resampling of the grid points and contrasts in resulting maps. Green <200 mg/kg; Yellow ≥200 mg/kg; Red ≥ 400 mg/kg; Purple ≥ 800 mg/kg.

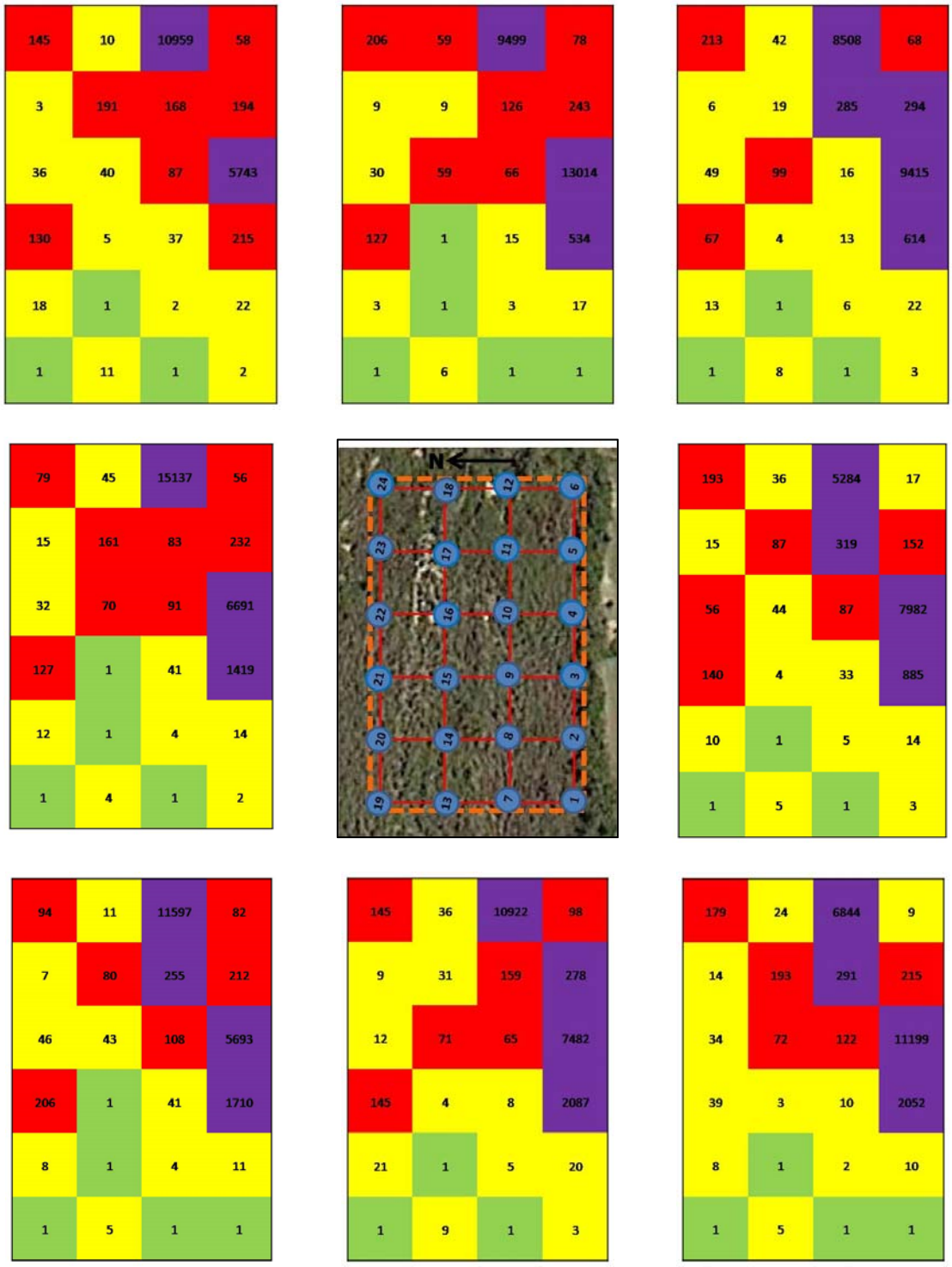


Figure 6-7. Eight artificial, small-scale, pattern of total PCB distribution at Study Site C based on random assignment of a concentration within the minimum and maximum range estimated for each grid point (study area depicted in center; grid area 7,200 ft^2). Patterns reflect hypothetical, independent resampling of the grid points and contrasts in resulting maps. Green <1.1 mg/kg; Yellow ≥1.1 mg/kg; Red ≥ 50 mg/kg; Purple ≥ 250 mg/kg.

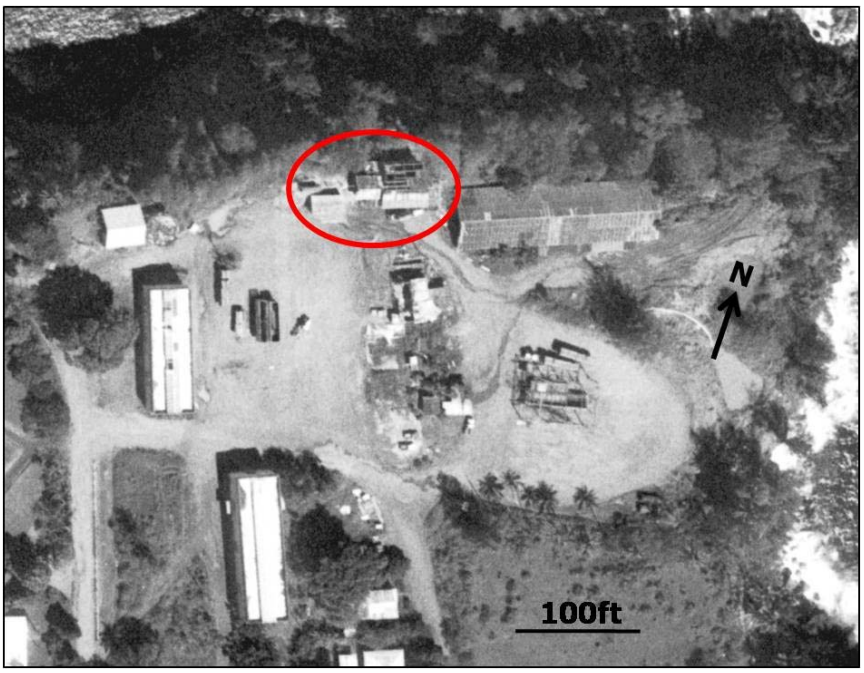


Figure 6-8. Former Hakalau pesticide mixing facility (circled) on the island of Hawai´i (1979 aerial photo).

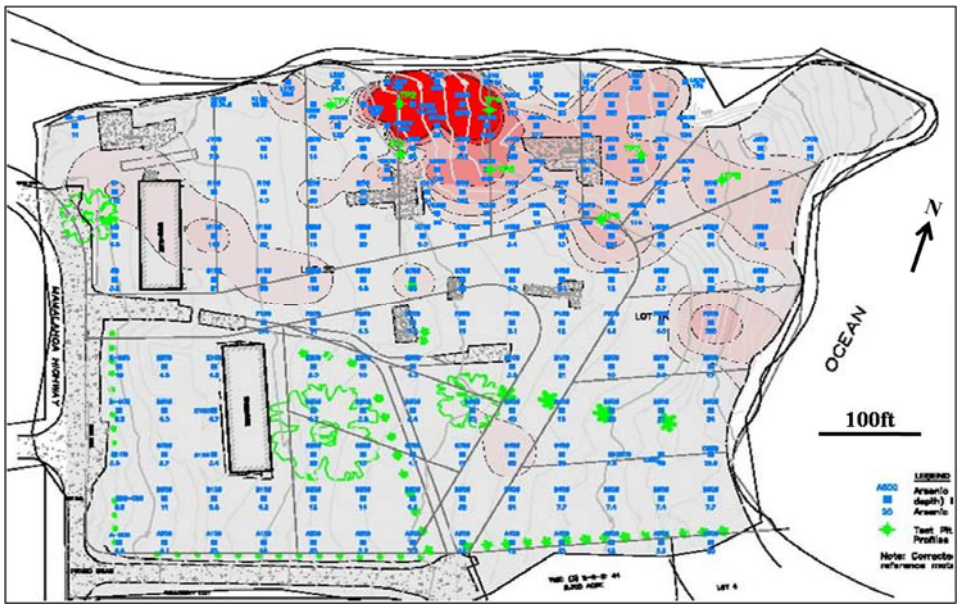


Figure 6-9. Isoconcentration map generated from discrete soil sample data collected at the arsenic-contaminated, Hakalau site on the island of Hawai´i (after ERM 2008; IDW distance decay parameter value = 5). Isoconcentration map generated from discrete soil sample data collected at the arsenic-contaminated, Hakalau site on the island of Hawai´i.

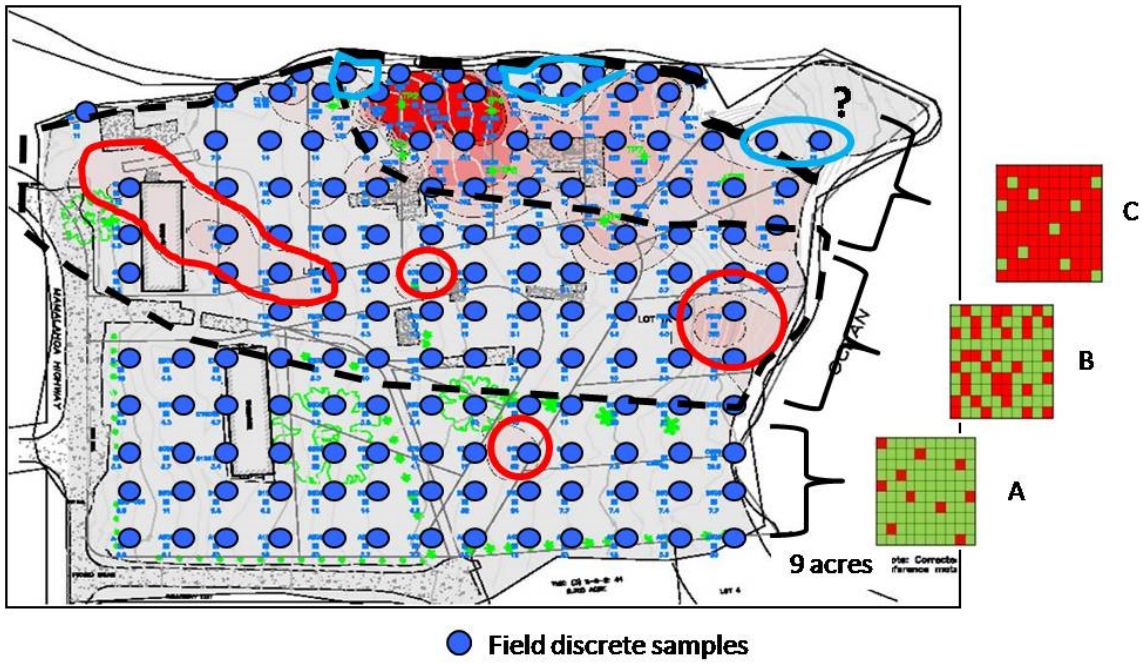


Figure 6-10. Random, small-scale variability expressed as isolated “hot spots” and “cold spots” within area (Zone B) separating areas of consistently low (Zone A) and high (Zone C) arsenic concentrations in soil (after ERM 2008).

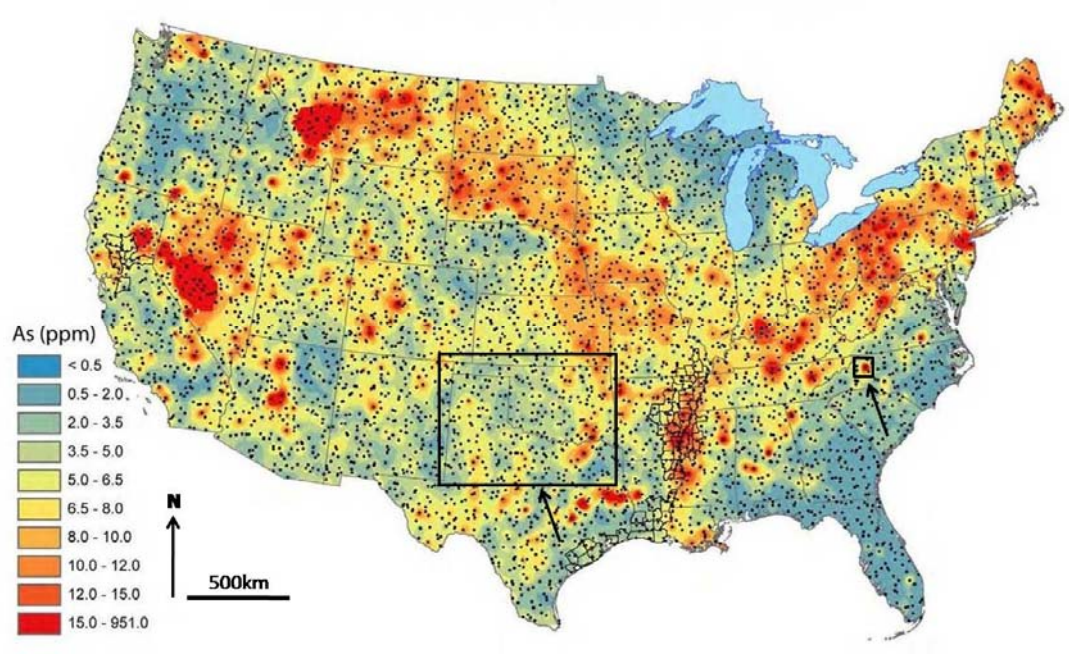


Figure 6-11. Isoconcentration map of naturally occurring arsenic in soils across the United States generated from data for composite soil samples collected over one-meter square areas (sample points depicted by black dots). Large-scale patterns primarily tied to geologic provinces of different rock and soil types (after USGS 2014). Small-scale "hot spots" are likely artifacts of random, small-scale heterogeneity within larger areas. Arrow points to areas depicted in Figure 6-17 and Figure 6-18.

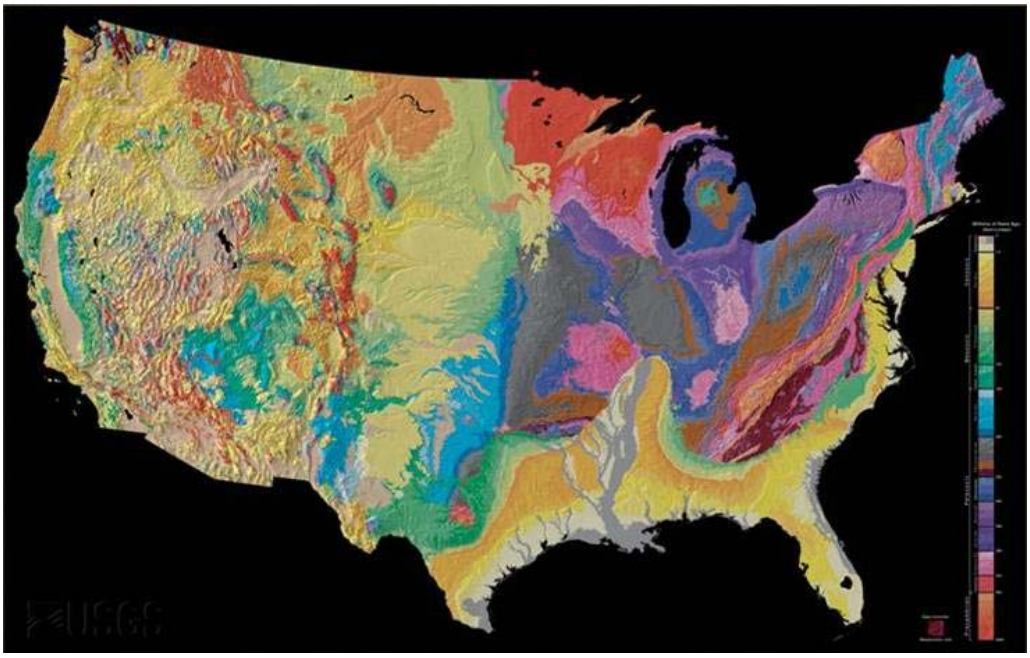


Figure 6-12. Geologic map of the United States (USGS 2004). Compare to patterns of arsenic distribution in surface soils depicted in Figure 6-11.

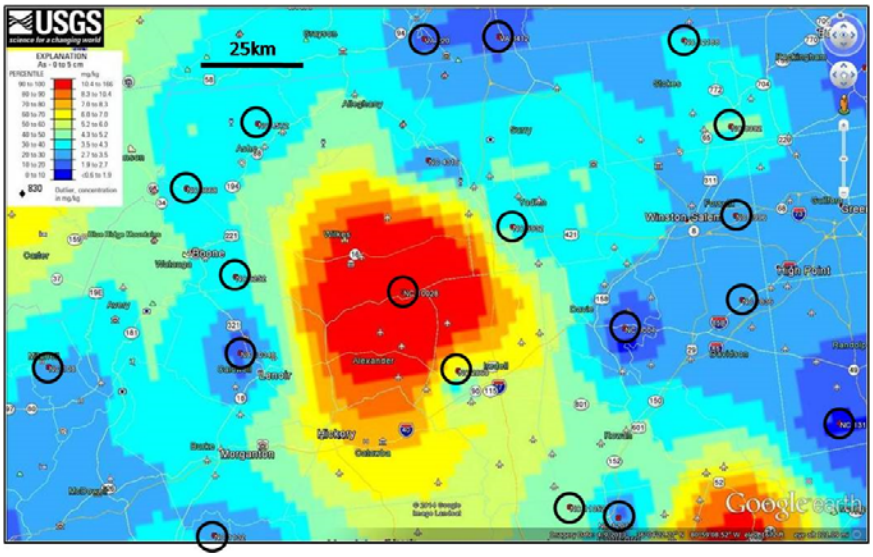


Figure 6-13. Artificial, 2,500km^2 arsenic "hot spot" in western North Carolina (see Figure 6-11) based on computer-generated extrapolation of two, one-square meter sample points separated by tens of kilometers (after USGS 2014; total 19 sample points within approximately 25,000km^2 area).

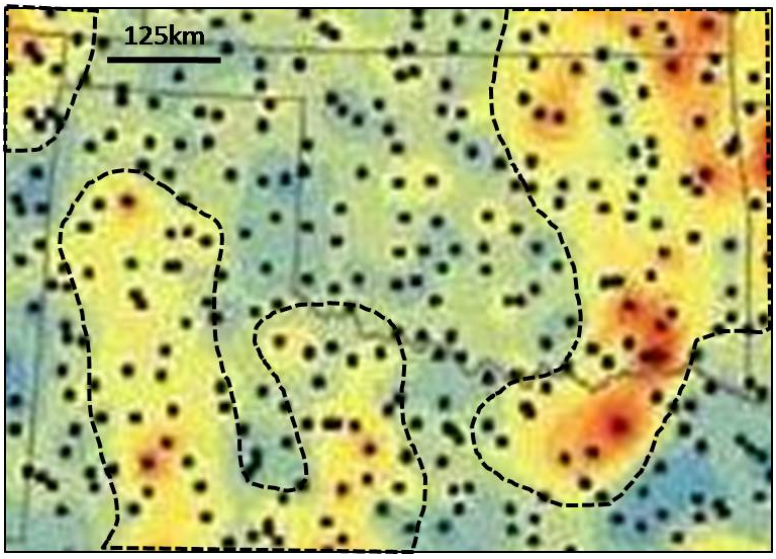


Figure 6-14. Larger-scale, likely reproducible and geologically-correlated patterns of soil arsenic variability in northern Texas and Oklahoma (see Figure 6-15; for example only, after USGS 2014). Each dot represents data for a one meter-square, composite soil sample. Smaller-scale patterns within larger areas most likely reflect random, small-scale heterogeneity and are not reliable indicators of arsenic concentrations at any given point within the map area.

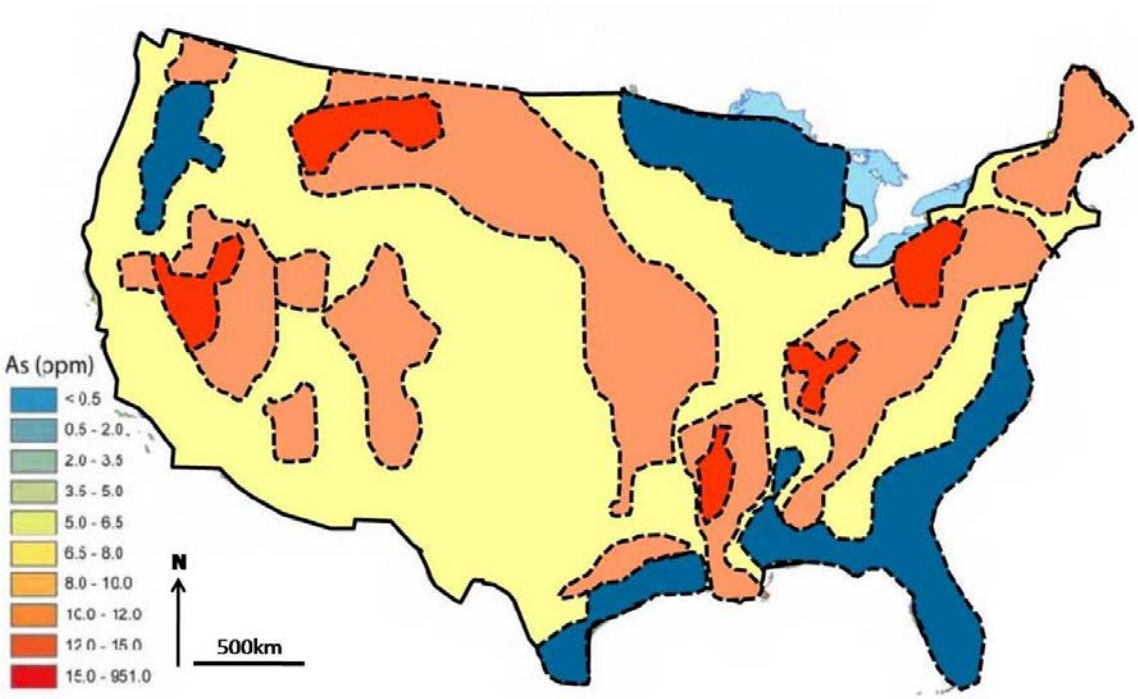


Figure 6-15. Hypothetical, large-scale variability of arsenic concentrations in soil across the US "filtered" to remove random, small-scale, random heterogeneity (compare to Figure 6-11; for example only). Elevated levels of arsenic in the upper, Mississippi flood plain could reflect deposition of fine sediment from more arsenic-rich regions of the upper watershed.

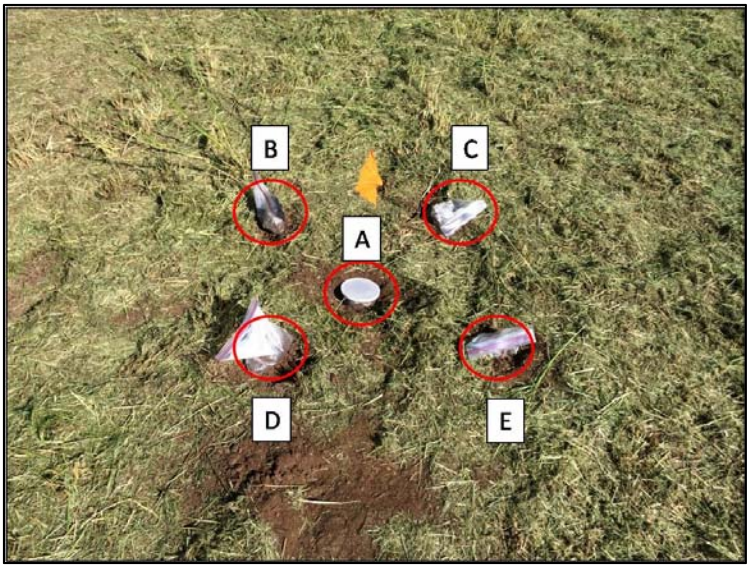


Figure 6-16. Pattern and labeling of discrete sample collection around grid points for evaluation of inter-sample variability (samples processed using MIS methods for analysis).

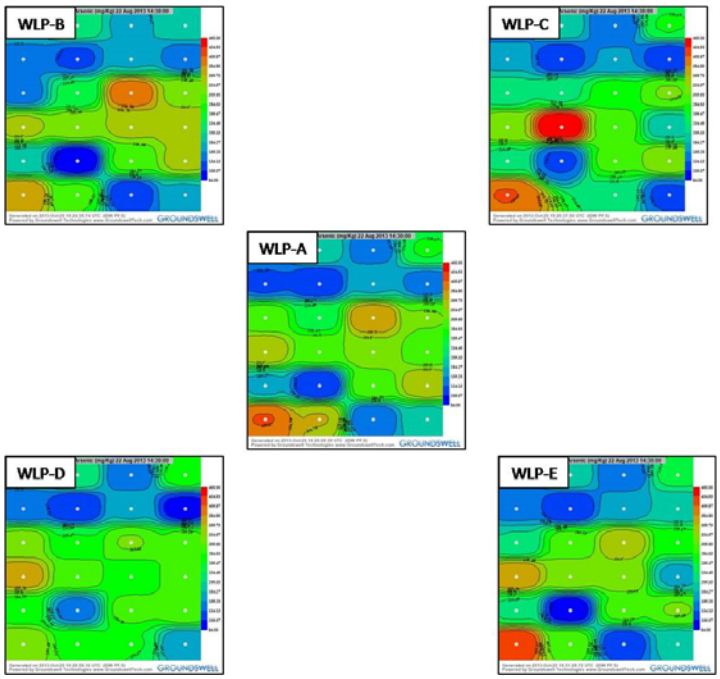


Figure 6-17. Changing locations of isolated “hot spots” and “cold spots” depending on use of arsenic data for “A,” “B,” “C,” “D,” or “E” processed sample sets for Study Site A (Groundswell Technologies; IDW Power Function = 5). Individual spots represent approximately 900 ft2 area (refer to Figure 2-4 in Part 1; Grid Point #1 in lower, left-hand corner). Changing patters reflect random, small-scale variability of arsenic concentrations around individual grid points and use of an unrealistically high, isoconcentration mapping power function.

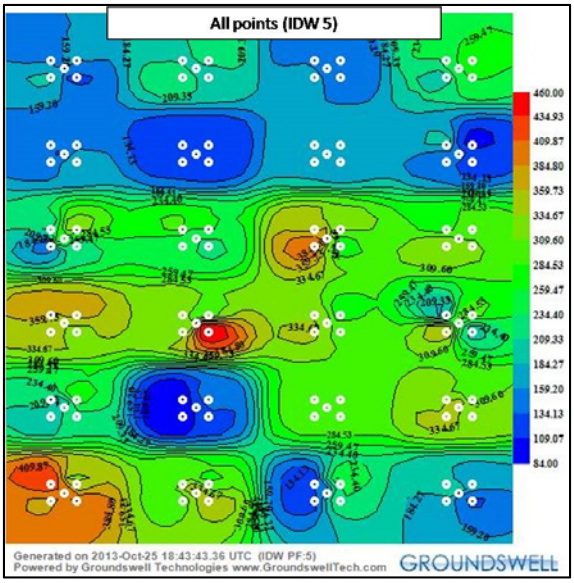


Figure 6-18. Similar artificial patterns of higher and lower arsenic concentrations in soil at Study Site A due to random, small-scale variability (13,500 ft^2 area; refer to Figure 2-4 in Part 1). Grid Point #1 in lower, left-hand corner. Map generated using all data for processed, discrete samples at grid points (5 per point) and typical mapping power function used for generation of contaminant isoconcentration maps (Groundswell Technologies; IDW Power Function = 5).

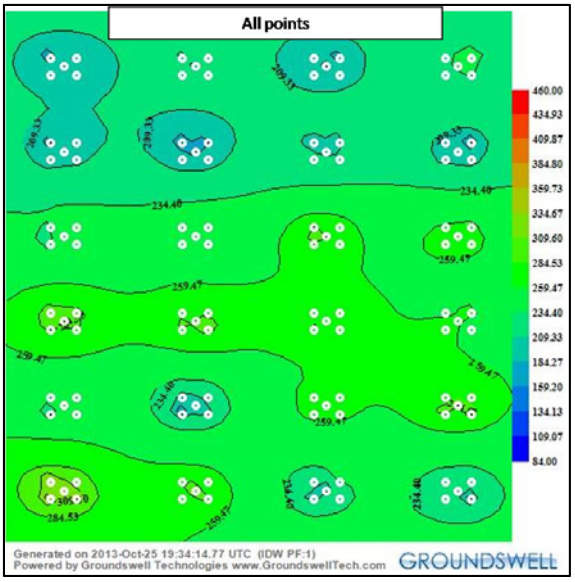


Figure 6-19. Reduced, artificial small-scale variability within Study Area A using all data for processed, discrete samples at grid points and minimizing interpretation of individual data points (Groundswell Technologies; IDW Power Function = 1). Apparent, isolated hot spots are artifacts of the interpolation algorithm but slightly higher concentrations of arsenic in upper third of study site area are presumably real (separated by 234 mg/kg isoconcentration contour).

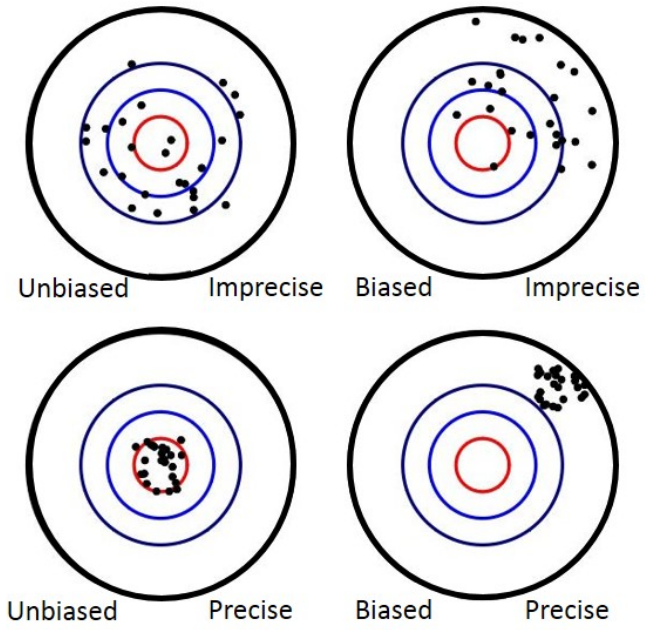


Figure 7-1. Four possible relationships between bias and precision (after ITRC 2012). The objective of an investigation is to obtain unbiased data with an acceptable level of precision.

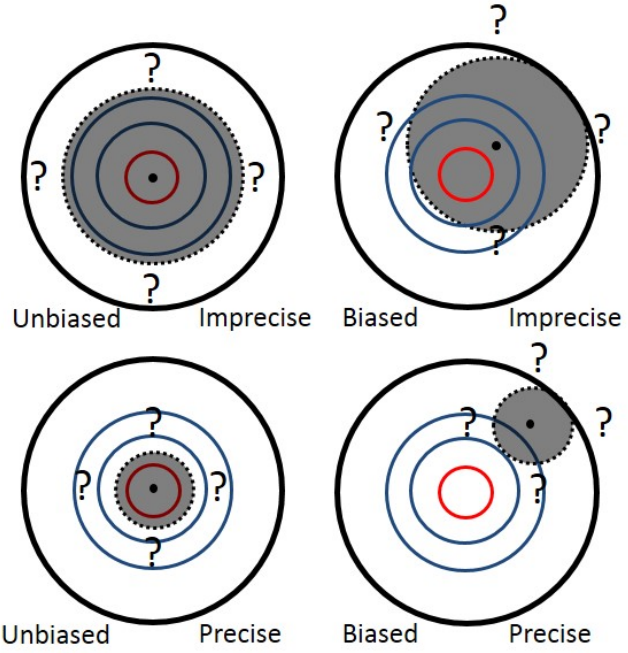


Figure 7-2. Relationships between bias and precision for a mean concentration estimated from a single, discrete sample data set.

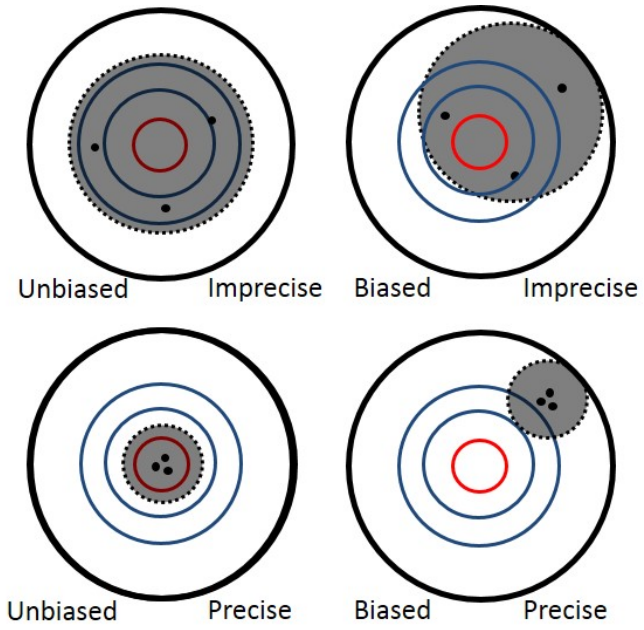


Figure 7-3. Relationships between bias and precision for a mean concentration estimated from a set of triplicate, incremental samples.

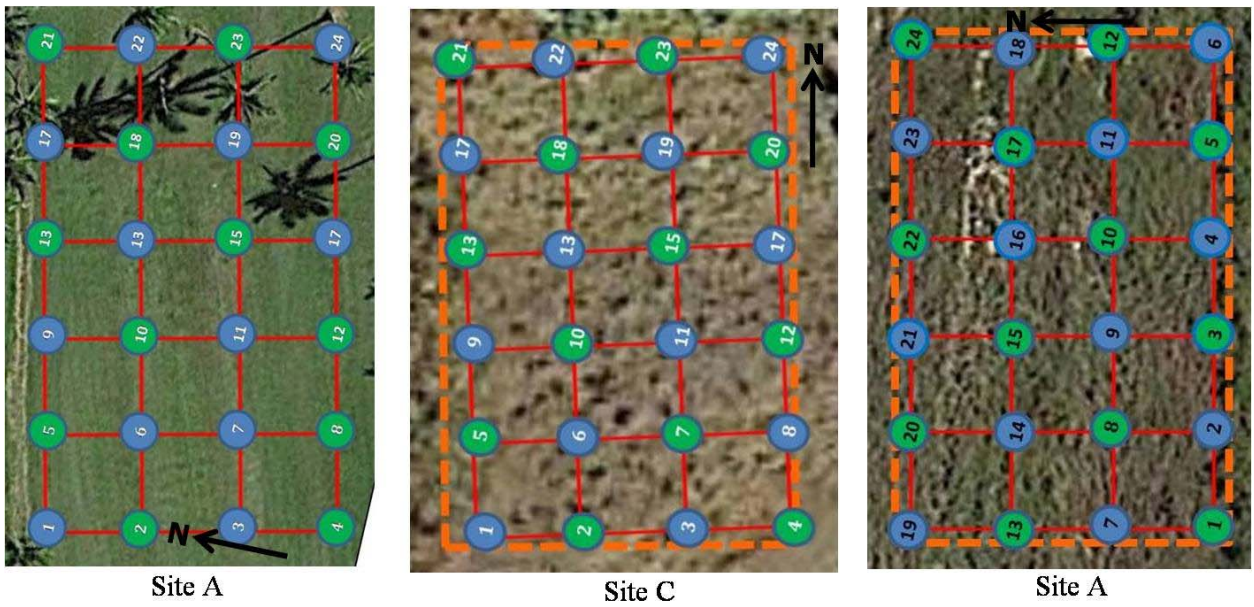


Figure 7-4. Twelve-point grid point sets for each study site used to evaluate field precision of random discrete data groupings.

Figure 8-1. Size of soil subsample masses typically tested for metals (one gram, left photo) and PCBs (ten grams, right photo) by commercial laboratories (minimum ten grams recommended for metals in HDOH guidance; HDOH 2008).

Figure 8-2. Sample from Study Site C mechanically "homogenized" by stirring, prior to the collection of a ten gram subsample mass from the top of the jar to be tested for PCBs.

Figure 8-3. Collection of subsample for analysis from processed soil sample following drying and sieving in accordance with incremental sampling procedures.

Figure 8-4. Irregular and disconnected spill patter due to flow of released milk along "preferential pathways" of low lying areas along the ground surface. Similar, vertical patterns of dispersion due to small differences in permeability might also characterize releases of liquids to the subsurface.

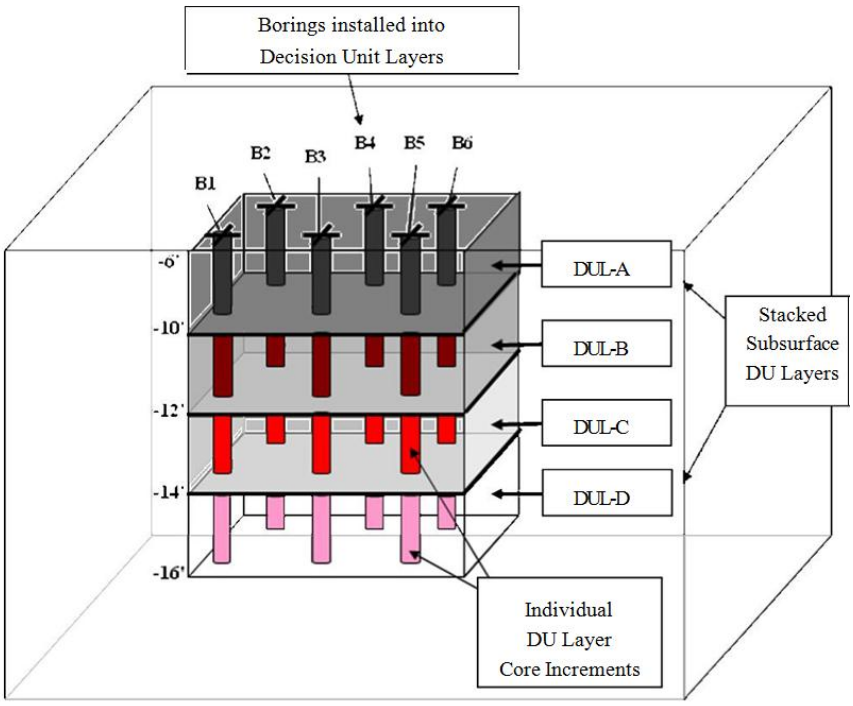


Figure 8-5. Decision Unit layers and associated core increment locations designated for the investigation of subsurface contamination.

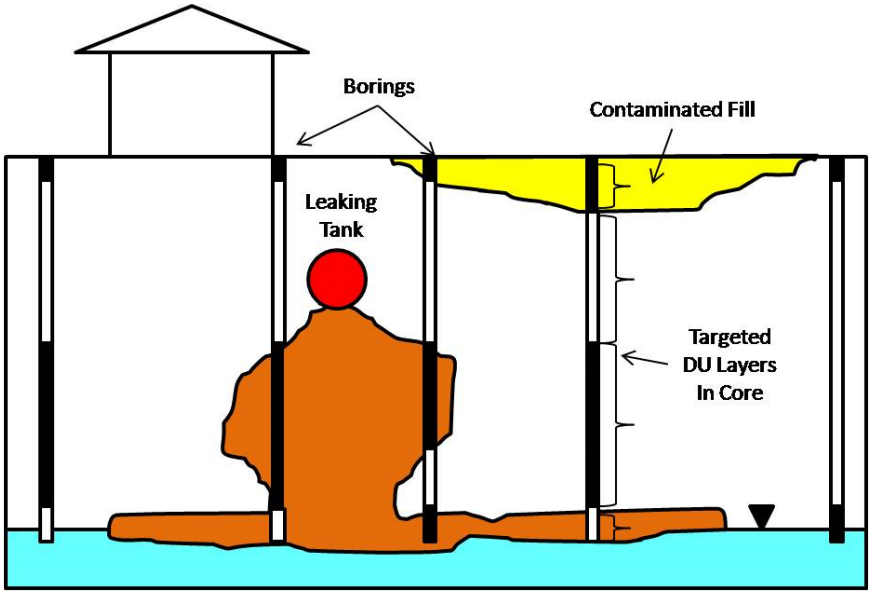


Figure 8-6. Use of a single or small number of “Borehole DUs" to estimate the vertical or lateral extent of contamination at a specific location within a site. Entire targeted core interval or representative subsample of interval submitted to lab for processing and testing.

Attachment 1: Concept of Large-Scale and Small-Scale “Hot Spots” in Early USEPA Guidance

1.0 Large-Scale Hot Spots

The identification and characterization of large-scale patterns of contamination such as the area of PCB contamination depicted in Figure 3-1 is the objective of most environmental investigations. Early USEPA guidance emphasizes the identification of large-scale "hot spots" as part of an environmental investigation (USEPA 1987; see also USEPA 1989a, 1991, 1992a):

> At sites or portions of sites where soil contamination is suspected but no definite sources have been identified, an objective of the remedial investigation might be to determine if soil contamination is present. Important decisions facing the site manager are how many samples must be taken to investigate the potentially contaminated area and where the samples will be located… The decision maker must determine… the acceptable probability of not finding an existing contaminated zone in the suspected area. For instance, it might be determined that a 20 percent chance of missing a 100ft-by-100ft (10,000ft^2) contaminated zone is acceptable but only a *5* percent chance of missing a 200ft-by-200ft (40,000ft^2)zone is acceptable.

The authors are clearly focusing on the identification of large-scale, i.e., "mappable," areas of elevated contamination. "Compositing" of samples collected from the grid area was discouraged due to potential "dilution" of large-scale areas of contamination and overlooking a significant "hot spot" (see USEPA 1987, 1989a, 1991, 1992a). In this sense, however, the guidance documents are describing the need to segregate and independently sample and characterize separate source areas to the extent known practical. Both HDOH and ITRC likewise make this requirement in their respective, incremental sampling guidance documents. The error in the early USEPA guidance documents was again the assumption that a single, discrete soil sample could be relied upon to identify and represent separate source areas, or to characterize the large-scale distribution and magnitude of contamination within a contaminated area.

Associated guidance written in the same time period states that "…there is no universal definition of what constitutes a hot spot" (USEPA 1989a). Discrete sample grids were specifically designated with the concept that the spacing between grid points represented the maximum allowable size hot spot that the sampling scheme should identify. As stated in the USEPA Data Quality Objectives guidance (USEPA 1987; emphasis added):

> The probability of not identifying a contaminated zone is related to the area or volume of the contaminated zone and the spatial location of the samples… To apply this method, the following assumptions are required… The shape and size of the contaminated zone must be known at least approximately. This known shape will be termed the target… *Any sample located within the contaminated zone will identify the contamination*. These assumptions are not severe and should be met in practice.

This is assumption is restated in the followup USEPA document *Methods for Evaluating the Attainment of Cleanup Standards* (USEPA 1989):

> When there is little distance between points it is expected that there will be little variability between points.

The target "contaminated zone" described in the guidance is referred to as a "Spill Area Decision Unit" in HDOH guidance (HDOH 2008; see also "Source Zone" DUs in ITRC 2012). Importantly, the USEPA guidance assumes that a single sample collected within the grid area will be adequate to represent the mean (or even maximum) contaminant concentration for that area. As discussed in Section 2 of the main text of this report, this assumption was based on experience in testing industrial waste streams and an assumption that contaminant concentrations in soil would be similarly "uniform."

The recommended grid and discrete sampling approaches were similarly incorporated into USEPA guidance documents for the investigation and cleanup of sites contaminated with polychlorinated biphenyls (PCBs; USEPA 1985, 1986, 1990). In this case requirements for the use of discrete sampling grids and prohibitions against "compositing" of samples were directly incorporated into formal regulations under the Toxic Substances Control Act that continue to be enforced in large part to this day (USEPA 2005a).

Several workers cautioned of potential problems at the time, but these concerns were overlooked as the investigation of tens or hundreds of thousands of sites across the US quickly began to be initiated, often as part of time-critical property transactions and redevelopment (e.g., Hadley and Sedman 1992; Pitard 1993; see also Ramsey and Hewitt 2005; Hadley et al. 2011; Hadley and Mueller 2012; Hadley and Petrisor 2013; Hadley and Bruce 2014).

As stated by Hadley and Sedman (1992):

> Every year across America, tens of thousands of soil samples are collected and analyzed for the presence of toxic contaminants. From among these sampling results, "hot spots" of soil contamination are identified. One or more hot spots on a property precipitates follow-up activities, typically at great expense. Given that costly action is undertaken as a result of this identification, it is surprising that there is no objective approach to identifying what is or is not a hot spot of soil contamination.

As further discussed by the same authors:

> Remediation of sites that pose clear threats to public health is acknowledged as the highest priority when expending the tens of billions of Superfund dollars projected for the national cleanup program… Given that… protection of human health appears to be a clear priority, a health-based measure and approach for evaluating the impact of soil contamination would appear to be an appropriate basis for determining whether a spot is "hot" or not… The term "hot spot" conveys a notion that a condition exists that merits consideration as a potential threat to the public health… The identification of a hot spot should not be site-specific or contaminant-specific, but, rather, risk-related … Only huge

> volumes of soil at a level of 1,000 ppm hold more gasoline than a person might be transporting in a spare 1-gal can in the trunk of their car. Clearly, identification of a hot spot should discriminate between minute and significant amounts of contamination.

To be more precise, approximately 3 metric tons (3,000 kg) of soil would be required to retain one gallon (3.6 liters) or approximately 3,000,000 mg of gasoline at an average concentration of 1,000 mg/kg. The risk posed by a handful of soil with an average gasoline concentration of 1,000 mg/kg would clearly be less than the risk posed by a football field area mass of soil with the same average concentration of gasoline.

This introduces the greater importance of the mean contaminant concentration for an "exposure area" over the concentration in an individual, discrete soil sample. Guidance on this subject was being developed and published by the USEPA and other entities in the same time period (see USEPA 1987, 1989a, 1991, 1992a, 1992b, 2005). The size of decision units designated for a site under investigation depends on the question being asked. Typical environmental concerns might include "Could leaching of contaminants from soil areas of the site where pesticides were mixed pose a risk to underlying groundwater?" or "Does contamination in soil in a yard pose a potential health risk to the residents?" The scale of the evaluation is important in both cases.

Of primary concern is continuous, long-term, "chronic" exposure to contaminants in soil over many years (refer to USEPA 1992b). This is clearly stated in more recent USEPA guidance documents (USEPA 2005; see also USEPA 2014a):

> The exposure unit generally is the geographic area within which a receptor comes in contact with a contaminated medium during the exposure duration… Exposure point concentration (EPC) is one of the key variables in estimating exposure in risk calculations. For purposes of this guidance, the EPC is not a point value but rather an average value for an exposure unit (EU)… The EPC is defined in EPA's *Risk Assessment Guidance for Superfund: Volume III - Part A* as "the average chemical concentration to which receptors are exposed within an exposure unit…" For "reasonable maximum exposure" (RME), the Risk Assessment Guidance for Superfund (RAGS) recommends using the average value with a specified level of confidence to represent "a reasonable estimate of the concentration likely to be contacted over time. This average value generally is based on the assumption that contact is spatially random.

The concept of "Decision Units (DUs)" is used in the HDOH and ITRC incremental sampling guidance documents to better define the scale at which an environmental investigation should be carried out (HDOH 2008, ITTC 2012). A Decision Unit is an area or more specifically the volume of soil which will be sampled and a decision made on the resulting data. Large-scale "hot spots" of contaminated soil, referred to as "Spill Area (HDOH 2008)" or "Source Area (ITRC 2012)" DUs, are areas of contamination associated with the specific release of a chemical and distinct from the surrounding areas. Areas of interest for investigation typically vary in size from several hundred to several thousand square feet but could be significantly larger. Examples

include areas of soil contaminated by disposal of waste solvents or petroleum at former industrial complexes, leaks of petroleum from tanks and pipelines, burning of wood coated with lead-based paint at former dump sites, spills of PCB containing oil at electrical facilities, etc. Characterization of large agricultural fields for residual pesticides could involve testing of tens or even thousands of acres as a single "spill area" if the objective is to determine the mean concentration of pesticides in the field as a whole.

In the absence of known or suspect spill areas, such areas are normally broken up into "exposure areas" DUs for independent testing. Exposure areas that exceed a target screening level for a contaminant could also be considered to be "hot spots," or more appropriately "hot areas" within an overall site. Residential exposure areas can be as small as a few hundred square feet of barren soil under and around a swing set or as large as several thousand square feet and include the entire yard. The size and shape of exposure areas at commercial and industrial properties varies with use but again tend to range in size from several hundred to several thousand square feet. Risk is assessed in terms of the *mean* concentration of the contaminant for the DU as a whole, with limitations on the maximum allowable size of DUs based on designated or default exposure areas (e.g., 1,000 ft^2 or 5,000$ft^2$to a depth of six inches).

Early USEPA guidance recognizes that small-scale heterogeneity within a spill area or exposure area DU can cause the reported concentration of a contaminant to range both above and below a target cleanup level at the scale of an individual, discrete sample (USEPA 1989a):

> When a sample is taken and the concentration of a chemical exceeds the cleanup standard for that chemical, it is concluded that the sampling position in the field was located within a hot spot… A site manager inevitably confronts the possibility of error in evaluating the attainment of the cleanup standard: is the site really contaminated because a few samples are above the standard? Conversely, is the site really "clean" because the sampling shows the majority of the samples to be within the cleanup standard?

This issue is unavoidable if discrete samples are used to characterize large-scale areas of soil contamination. As discussed in Part 1 of this study, at some point the variability of contaminant concentrations at the scale of a discrete sample will begin to fall both above and below the target screening level. Past USEPA guidance recognized this potential limitation in the use of a single discrete sample to represent a large area of soil. As discussed in the guidance document *Methods for Evaluating the Attainment of Cleanup Standards* (USEPA 1989a):

> This document assumes that… chemical concentrations *do not exhibit short-term variability* over the sampling period.

This caveat also applies to an assumed absence of random, small-scale, spatial variability of contaminant concentrations in soil. Potential problems with very small, discrete soil samples were further elaborated in the USEPA guidance document *A Rationale for the Assessment of Errors in the Sampling of Soils* in terms of "representative sampling" (USEPA 1990b):

> Soils are extremely complex and variable which necessitates a multitude of sampling methods… *A soil sample must satisfy the following*: 1) Provide an adequate amount of soil to meet analytical requirements and *be of sufficiently large volume as to keep short range variability reasonably small… The concentrations measured in an heterogeneous medium such as soil are related to the volume of soil sampled* and the orientation of the sample within the volume of earth that is being studied. The term 'support" is used to describe this concept.

The same document warned that errors in the collection and representativeness of soil samples were likely to far outweigh errors in analysis of the samples at the laboratory (USEPA 1990b):

> During the measurement process, *random errors* will be induced from: sampling; handling, transportation and preparation of the samples for shipment to the laboratory; taking a subsample from the field sample and preparing the subsample for analysis at the laboratory, and analysis of the sample at the laboratory (including data handling errors)… *Typically, errors in the taking of field samples are much greater than preparation, handling, analytical, and data analysis errors*; yet, most of the resources in sampling studies have been devoted to assessing and mitigating laboratory errors.

Addressing random errors in the laboratory was and has continued to be "low hanging fruit" that received the greatest focus of attenuation over the past 20 to 30 years (USEPA 1990b):

> It may be that those errors have traditionally been the easiest to identify, assess and control. This document adopts the approaches used in the laboratory, e.g. the use of duplicate, split, spiked, evaluation and calibration samples, to identify, assess and control the errors in the sampling of soils.

The implications of these important ideas in the field were, unfortunately, never fully discussed in guidance documents. Ultimately, confusion over the need to determine the "maximum" contaminant concentration within a targeted area and search for sample-size "hot spots" continued (and still continues) to plague the industry, and reliance on often scant discrete soil sample data for decision making quickly became routine.

2.0 Small-Scale Hot Spots

Perhaps the greatest source of confusion in environmental investigations is the need (and ability) to identify and characterize "hot spots" of elevated contaminant concentrations at the scale of an individual, discrete sample. Refer again to Figure 3-2 and Figure 3-3. From a field perspective, the mass of soil traditionally collected as a discrete sample has no basis in science or sampling theory. The mass of a soil sample, typically a few hundred grams, is instead determined by the mass of soil needed by the laboratory to carry out the requested analyses and related quality assurance and control measures for the analyses. Since they must store and ultimately dispose of any soil received, it is in the laboratories interest to request only the smallest mass necessary in

order to optimize storage and testing space and minimize disposal costs. In this sense, the mass of traditional discrete soil samples is driven almost completely by laboratory needs, rather than consideration of representativeness in terms of the area from which the sample was collected (refer to ITRC 2012).

Sampling theory and the need to ensure that a soil sample is in fact representative of the area from which it was collected is touched upon in the USEPA guidance document *Preparation of Soil Sampling Protocols* (USEPA 1992a; see also Pitard 1993, 2005, 2009; Minnitt et al 2007):

> Gy's theory makes use of the concept of sample correctness which is a primary structural property… A sample is correct when all particles in a randomly chosen sampling unit have the same probability of being selected for inclusion in the sample…

The authors use the term "sampling unit" in the same sense of a "decision unit" as described above and in HDOH guidance (HDOH 2008). The authors go on to describe in detail the types of error that can be associated with sample representativeness in accordance with Gy's sampling theory. They focus in particular on the variability of contaminant distribution at the scale of individual particles (e.g., fundamental error) and the need to collect a sufficient mass of soil to ensure that very small, "micro-scale" distributional heterogeneity is adequately captured in the sample collected. The authors caution against the over-interpretation of traditional discrete samples collected without an adequate understanding of basic sampling theory (USEPA 1992a):

> "Grab samples" or judgmental samples lack the component of correctness; therefore, they are biased. The so-called grab sample is not really a sample but a specimen of the material that may or may not be representative of the sampling unit. Great care must be exercised when interpreting the meaning of these samples.

The document points out the important distinction between what they refer to as "short-range" (i.e., "small-scale") and "long-range" (i.e., "large-scale") variability, following the terminology used by the mining industry (USEPA 1992a):

> Long-Range Heterogeneity (is)… created by local trends and is essentially a nonrandom, continuous function. This heterogeneity is the underlying basis for much of geostatistics and kriging… The short-range heterogeneity… is essentially a random, discontinuous function… This error is the error occurring within the sampling support.

The concept of "sample support" refers to the representativeness of the sample(s) collected. Although not explicitly stated, the document goes on to imply that a soil sample or set of samples must be adequate to overcome and capture random, short-range heterogeneity in order to reliably represent the mean contaminant concentration for any given area (and volume) of soil as well as for decision making regarding non-random, large-scale trends of interest. The later point is important and is discussed in more detail in Section 7 of this report, which explores the reliability of isoconcentration maps based on traditional, discrete sample data.

The authors of the 1992 USEPA guidance were well ahead of their time in terms of environmental investigations. Sampling theory was later invoked as a basis for processing and testing of soil samples received by a laboratory (USEPA 2003; refer to Section 8.1 in main text of report), but is only now being applied to the representativeness of the samples actually collected in the field. This is in part due to the continued confusion by the authors over the need to understand contaminant concentration at the scale of a still largely arbitrary, discrete sample mass, as described in the same guidance document (USEPA 1992a):

> Pitard (1989) recommends developing a sample by taking a large number of small increments and combining them into a single sample submitted to the laboratory… One of the problems with compositing samples is the loss of information and the loss of sensitivity because of dilution of the samples.

This could perhaps be considered a second, critical juncture in the use of discrete rather than incremental sampling methodologies to characterize sites with contaminated soil, with the first being a failure to collect large sets of replicate samples during initial testing of grid schemes for PCBs in 1986 (refer to Section 8.3 in main part of text). After presenting a strong review of sampling theory and error associated inherent, random, small-scale variability of contaminant concentrations in soil, the authors fall into the same "hot spot" trap and the need for decision making on a sample-by-sample basis reflected in similar guidance being written in the same time period (USEPA 1992a):

> …the effects of contaminant dilution can be reduced by specifying the minimum detection limit (MDL) for the analytical procedure and… the action level (AL)… for the site. Using this information, the maximum number of samples or increments that can be composited (n) is given by: n = AL/MDL… Test statistics (are used) for determining if any sample within the group of samples combined into the composite were above the AL. Those groups that fail the test are then analyzed as individuals to determine which support fails the AL criterion.

The authors are mistakenly assuming that risk-based action levels starting at that time to be published for direct-exposure concerns, including USEPA Preliminary Remediation Goals had to be met by any given, discrete sample mass of soil within a site or exposure area (now referred to as Regional Screening Levels; USEPA 2014b). Such screening levels in fact apply to chronic health risk posed by long-term exposure to contaminants in soil within a designated exposure area. In such cases, the screening levels are intended to apply to the mean contaminant concentration for the exposure area, not to individual samples collected within the exposure area. Use of the mean acknowledges that concentrations at the scale of a discrete sample can be expected to fall both above and below these screening levels.

This repeats a mistake made seven years earlier in guidance for testing of PCB-contaminated soils, which states that no more than nine (ten in other documents) samples should be composited as a single sample, in order to ensure that no single sample might have exceeded an risk-based

screening level if PCBs are not identified above the laboratory method detection level, even though these screening levels again apply to long-term, chronic exposure (USEPA 1985).

> Once the samples have been collected at a site, the goal of the analysis effort is to determine whether at least one sample has a PCB concentration above the allowable limit. This sampling plan assumes the entire spill area will be recleaned if a single sample contaminated above the limit is found. Thus, it is not important to determine precisely which samples are contaminated or even exactly how many. This means that the cost of analysis can be substantially reduced by employing compositing strategies, in which groups of samples are thoroughly mixed and evaluated in a single analysis. If the PCB level in the composite is sufficiently high, one can conclude that a contaminated sample is present; if the level is low enough, all individual samples are clean.

Guidance published the following year goes so far as to specify the number of discrete samples that can be combined for a single analysis, most likely based on a target action level of 1 mg/kg and a then method detection limit of approximately 100 µg/kg (see USEPA 1986). As somewhat ironically stated in the same document that discusses the important of Gy's sampling theory (USEPA 1992a):

> Do not form a composite with more than 10 samples, since in some situations compositing a greater number of samples may lead to such low PCB levels in the composite that the recommended analytical method approaches its limit of detection and becomes less reliable.

Subsequent guidance, even noting that the exceedence of a risk-based screening level in a single discrete sample may not necessarily indicate a risk to human health and the environment, calls for a halt of sampling and move to remediate the entire site if such a scenario is encountered (USEPA 1989a; annotation added):

> Because of this requirement (i.e., cleanup required if any single sample exceeds a screening level) it may be advisable, after identifying the presence of a single hot spot, to continue less formal searching followed by treatment throughout the entire sample area.

While perhaps easy to suggest from a regulatory perspective, such a misunderstanding of risk no doubt led to unnecessary cleanup at a large numbers of sites, with significant expense and legal burdens imposed on the property owner. As discussed above and in Section 4 of the main text, risk-based soil screening levels under development at the time applied to the mean concentration of a contaminant in soil within large-scale, exposure areas, not to individual points within those areas.

The above approach was even more unfortunately codified in Toxic Substances Control Act regulations regarding testing of soils for PCBs, with the maximum number of discrete soil samples that could be composited being reduce to nine (USEPA 1998a; refer to Subpar O). This

is to large extent still enforced to this day, even though data are compared to screening levels specifically developed to address long-term exposure to PCBs in soil, which concurrent USEPA guidance states should be carried out by comparison to the mean. This unfortunate, but somewhat understandable turn in the 1992 USEPA document given the infancy of the industry at the time, helped to secure the continued use and misuse of discrete soil sample data for the next two-plus decades.

Science-based decisions in environmental investigations are rarely if ever made at the scale of an individual sample (refer to Section 4 in the main text). As stated in the USEPA Superfund Environmental Assessment Manual (USEPA 1988; see also USEPA 1989b,c):

> In most situations, assuming long-term contact with the maximum concentration is not reasonable.

In spite of this reasonable observation, subsequent USEPA guidance repeatedly discusses the need to collect discrete soil samples in order to verify the presence of absence of sample-size "hot spots" with data to be compared to unspecific "acute toxicity" or "not-to-exceed" screening levels (e.g., USEPA 1989a, 1992a). This concept was made prominent in the USEPA document *Guidance on Surface Soil Cleanup at Hazardous Waste Sites: Implementing Cleanup Levels*, with such criteria referred to as "Remedial Action Levels" (USEPA 2005; annotation added; note that this document is a Peer Review Draft and to our knowledge has not been finalized):

> Because soils with contaminant concentrations exceeding the cleanup level will be left onsite, it is important to ensure that those concentrations are not so high that they pose acute or subchronic health risks if exposure to them occurs. Therefore, if this approach is used, the RPM should conduct a separate assessment of potential acute effects to determine the contaminant concentration at which acute effects are likely to occur. The RAL should be below that concentration to ensure protection against acute effects. If acute toxicity data are insufficient to either determine whether the Remedial Action Level (i.e., level intended to be protective of short-term health effects) is protective for acute effects or to establish an alternative protective level, then the area average approach should not be used.

To those unfamiliar with risk assessment or sampling theory this may at first seem reasonable. The document continues by further discussing the difference between soil screening levels intended to be protective of chronic versus acute health risks (USEPA 2005):

> The Remediation Action Level in most cases is the maximum concentration that may be left in place within an exposure unit such that the average concentration (or 95% UCL of the average) within the EU is at or below the cleanup level… A vital concept in this document is the difference between the implementation of a cleanup level as a not-to-exceed level or as an area average. The not-to-exceed option typically entails treating or removing all soil with contaminant concentrations exceeding the cleanup level. The area

> average option typically involves treating or removing soils with the highest contaminant concentrations such that the average (usually the upper confidence limit of the average) concentration remaining onsite after remediation is at or below the cleanup level... The method used in implementing the cleanup level should be compatible with the method used in establishing the cleanup level.

The concept of theoretical, "acute hot spots" is then specifically introduced (USEPA 2005):

> Contaminants present at hazardous waste sites may pose human health risks from short-term exposures, as well as from long-term exposures. Therefore contaminants need to be evaluated for their acute and chronic toxicity, and the toxicity generally should be matched to the exposure duration and frequency… At most sites, it is reasonable to assume that random exposure occurs over the long-term. Short-term exposures, however, may be non-random. For example, a resident may move randomly across his/her property spending equal amounts of time in all areas over the long-term period of residence, but intense short-term exposure may occur as a result of a construction project, such as building a shed… To help risk managers decide whether to implement cleanup levels as not-to-exceed levels or as area averages, this part of the guidance discusses these options with respect to their advantages, disadvantages, and appropriate use.

The document then states that "all soil" must meet acute and not-to-exceed screening levels, again without stating the mass of soil at which this should be assessed or discussing how this would be implemented in the field (USEPA 2005):

> Implementing the cleanup level as a not-to-exceed value normally means that soil removal or treatment will continue until the analysis of soil samples indicates that all *soil with contaminant concentrations exceeding the cleanup level has been removed or treated*… Remediating or removing all soil with contaminant concentrations above the Remedial Action Level should enable risk managers to ensure that the estimated post-remediation EPC achieves the cleanup level…

In spite of the alleged importance of this issue the document provides no guidance on the calculation of either "acute" or "not-to-exceed" screening levels, nor does it provide guidance on sampling methods to establish with any degree of reliability the presence or absence of contaminated soil that could pose such concerns. Acute or not-to-exceed soil screening levels (e.g., health effects within minutes or a few days) have never, to the authors' knowledge, been published by the USEPA. This is in fact acknowledged in the same document (USEPA 2005):

> At present EPA does not have acute toxicity criteria, therefore consultation with a toxicologist may be necessary to determine if the RAL is sufficiently protective for acute effects.

The manner in which a toxicologist is to assess acute health risk, given the absence of this knowledge by even the authors of the 2005 guidance document, is again not discussed. The document states, however, that if acute health risks from exposure to very small masses of soil with hypothetical, very high concentrations of contaminants cannot be ruled out, then the entire site must be remediated under the assumption that such "spots" could indeed be present (USEPA 2005):

> If site characterization or sampling data are insufficient to provide confidence in the use of the area average method, then the cleanup level should be implemented as a not-to-exceed level because it generally provides more certainty about the protectiveness of the cleanup. The area average approach is specifically intended for situations where adequate site characterization data are available. Applications of area average methods to sites with limited or incomplete data are inappropriate. However, if the quality of site characterization data is the only factor limiting the use of the area average approach, it may be more cost-effective to spend more on sampling to improve the quality of the data before deciding to implement the cleanup level as a not-to-exceed level where the area average approach could save on remediation costs.

Realistically, this would be the case at any site since it is economically impractical if not technically impossible to determine with a reasonable degree of confidence that no single, discrete sample-size mass of soil among tens or hundreds of thousands (or more) of potential sample-size masses within an exposure area does not exceeds a hypothetical, maximum-allowable level. Acute toxicity would in practice need to be tied to ten-gram or smaller masses, the default assumed to be ingested by a pica child (USEPA 2011; default non-pica child soil ingestion rate 200 mg/day). Each ten-gram mass of soil at a site then becomes an individual "Decision Unit." To put this in perspective, a $100m^2$ ($1{,}000ft^2$) area to a depth of 15cm (six inches) includes approximately 15,000 kg of soil, or 150,000 hypothetical, ten-gram, "acute toxicity" DUs. The level of effort to prove beyond a reasonable doubt that no single, ten-gram mass of soil posed acute toxicity risks for even relatively small areas would be enormous and not ultimately feasible from either a technical or financial standpoint.

The hypothetical importance of identifying and removing small, isolated "hot spots" is carried to an extreme later in the guidance, through a remediation approach referred to as "Iterative Truncation" (USEPA 2005):

> (The iterative truncation method) is based on the identification and removal of soils with high contaminant concentrations to lower estimated post-remediation Exposure Point Concentrations (EPCs) to levels at or below the cleanup levels. Iterative truncation is used for non-spatial data, it assumes that each sample is an uncorrelated, unbiased representation of a remediation area within the site or Exposure Unit (EU). As indicated, iterative truncation involves removing (truncating) high values in the sample

> concentration measurements and calculating a hypothetical post-remediation EPC. For this reason, it is inappropriate to use composite samples.

In essence and as implemented in the field, this method involves excavation of soil around individual sample points where the reported concentration of a contaminant exceeded a screening level. One objective of the approach is to reduce the mean contaminant concentration within an exposure area to at or below a target screening level (or to meet a target risk). The document rightly cautions, however, that reducing the mean to address chronic, long-term exposure concerns may not be adequate to ensure that no single "spot" exceeds hypothetical acute or otherwise not-to-exceed soil screening levels.

The authors acknowledge that this approach is only defensible if the sample data accurately reflect conditions in the field on a point-by-point basis (USEPA 2005):

> To use this method with confidence, it is important to have good site characterization based on extensive, unbiased, and representative sampling, and the resulting data should adequately represent random, long-term exposure to receptors… Simple random sampling may fail to represent a patchy distribution of contaminants… If the highest sample concentrations are not representative of the highest concentrations in the (Exposure Unit) and there are actually areas with higher concentrations, then the resulting (maximum concentration left in place) may not be protective.

As demonstrated in Part 1 of this report, contaminant concentrations at the scale of a discrete soil sample are always likely to reflect a random "patchy distribution," referred to in sampling theory as distributional heterogeneity. The latter will of course always be the case, since both the mass of soil designated to assess the "maximum concentration" of a contaminant is never defined and in practice sampling will never be adequate to accomplish such an objective if it were possible. The maximum concentration of a contaminant in soil at the scale of a discrete sample, which represents the average concentration of the contaminant in the mass of soil actually analyzed, will never be known, nor does this need to be known for decision making purposes.

Small-scale, random variability of contaminant concentrations in soil negates the implementability of "Iterative Truncation" methods to remediate areas of contaminated soil. As discussed throughout the main text of this report, removal of soil within the immediate vicinity of a sample point where the initial concentration of a contaminant was reported above a screening level cannot be assumed to have significantly reduced the mean contaminant concentration for the area as a whole. Doing so is equivalent to removal of a single, randomly plucked red marble from a bucket of mixed marbles, with each marble representing a discrete soil sample, and assuming that this has significantly reduced the average redness for the bucket of marbles as a whole. In practice this would impossible to know without knowledge of every single marble in the bucket.

Recalculation of a mean, contaminant concentration based removal of the "hot spot" data point using data for remaining sample points is invalid, since the sample set as a whole has now been biased. This is true even if "confirmation" samples were collected around the excavated sample point. The "hot spot" removed in all likelihood is one of many and simply reflects the chance of identifying a "hot spot" given the total number of samples collected. For example, the identification of a small-scale hot spot at 2 out of 20 discrete sample points infers that such hot spots collective comprise 10% of the overall area and volume of soil at the site.

Re-estimation of a mean contaminant concentration for the area as a whole would require recollection of a new, independent set of discrete samples from separate sample and randomly selected points. Even this would not be fully adequate, since the representativeness of any single set of samples is unknown. As discussed in Section 7 of the main text, estimation of the precision of the resulting data set can still only be reliably accomplished by the collection of completely independent, replicate sets of discrete samples. Precision is evaluated by comparison of mean contaminant values for each replicate set of data to the original set of data, in the same manner as done for incremental soil sample replicates.

The underlying basis of the "iterative truncation" concept is again understandable. Attempts to investigation and remediate a site to the resolution of a single, discrete soil sample are destined to failure, however. This is due to lack of both toxicological information to develop acute-based soil screening levels and more importantly the impracticality of demonstrating with any degree of statistical certainty through representative sampling that no such "spots" are indeed present. The cost to private property owners and businesses to comply with such a requirement would be enormous. Fortunately, serious implementation of such an approach has not been implemented on a widespread basis in the authors' knowledge, although misguided and costly attempts to do so have certainly been carried out.

In reality investigating a site with the intent of demonstrating that none of the soil (i.e., any testable mass) does not exceed a risk-based, acute or not-to-exceed screening level, if such levels could in fact be developed, is both unnecessary and impractical. To the authors' knowledge, acute health effects have rarely if ever been reported due to exposure to contaminants in soil, aside from exposure to mine tailings and other industrial waste. Acute toxicity factors are also only available for a very small number of chemicals and applied primarily in industrial settings, where exposure to pure product might occur. As discussed above, health risk is instead most efficient evaluated in terms of long-term, chronic exposure to very low levels of contaminants in soil within a large-scale exposure area averaged over many years, rather than short-term exposure to very high levels of contaminants in very small masses of soil over a matter of seconds or even a few days.

More importantly, the basic principles of risk assessment and sampling theory would still apply even if such toxicity factors and screening levels were available. The sampling scheme would have to be designed in such a manner that the probably of making an error, for any given,

discrete sample-size mass of soil that was not tested as part of the investigation was acceptably small. The fact that the 2005 USEPA document leaves the definition of "any soil" undefined highlights the fact that this recommendation had not been well thought out in advance. Acute exposure would presumably be evaluated based on accidental ingestion of a ten-gram mass of soil, the default mass assumed to be ingested in any given event by a pica child (USEPA 2011). For comparison a relatively small, 1,000ft^2 (100m^2) area of soil to a depth of six inches (15cm) contains approximately 18,000 kilograms of soil, or 1,800,000 potential ten gram masses of soil. The level of sampling required to convincingly demonstrate that no single, ten-gram mass of soil exceeds an acute toxicity-based soil screening level with any given degree of confidence would be enormous, and impractical from both a cost and sampling perspective.

This is acknowledged but the significance unrecognized in early USEPA guidance (USEPA 1989a):

> The more (discrete) samples collected, the more likely that one sample will exceed a cleanup standard. That is, it is more likely to measure a rare high value with a larger sample (number).

As discussed in Section 8, detection of a "high value" of contamination in a small number of samples from a large data set can cause significant problems with statistical evaluation of the database. The same guidance document introduces the misused concept of "outliers" as a means to inappropriately ignore these data in a risk assessment and decision making (USEPA 1989a):

> Because of the chance of outliers, it may be that the (not-to-exceed) rule that allows one or more exceedances… in order to still have the site judged in attainment of the cleanup standard.

The inclusion of "outliers" in the data is in contrast an important part of sample representativeness and accurate estimation of risk. This issue is explored in more detail in Section 8 in the main text of this report. As discussed in that section, it is somewhat ironic that early USEPA sampling guidance emphasizes the need to identify sample-size hot spots while subsequent risk assessment guidance attempts to justify why such "outliers" can be ignored since they disrupt geostatistical models for calculation of mean contaminant concentrations from sets of discrete samples.

3.0 Current and Future USEPA Guidance

Problems with earlier USEPA and related guidance for testing soils are progressively being realized and more up-to-date guidance published. Many of the issue above are discussed in the document *Hot Spots: Incremental Sampling Methodology (ISM) FAQs* published by the USEPA Superfund office (USEPA 2014a). At the writing of this report, a number of individual offices within the USEPA are implementing incremental sampling approaches into their projects. It is

hoped that the field study of discrete sample error presented in this report will aid in these discussions.

References

Hadley, P.W. and R.M. Sedman, 1992, How· Hot Is That Spot?: Journal of Soil Contamination, (3), pp 217-225.

Hadley, P.W., Crapps, E. and A.D. Hewitt, 2011, Time for a Change of Scene: Environmental Forensics, 12, pp 312-318.

Hadley, P.W. and S.D. Mueller, 2012, Evaluating "Hot Spots" of Soil Contamination: Soil and Sediment Contamination, 21. Pp 335-350.

Hadley. P.W. and I.G. Petrisor, 2013, Incremental Sampling: Challenges and Opportunities for Environmental Forensics: Environmental Forensics, 14. Pp 109–120.

Hadley, P.S. and Bruce, M.L., 2014, On Representativeness: Environmental Forensics, 15:1, pp1-3.

HDOH, 2009, *Technical Guidance Manual* (2009 and updates): Hawai‘i Department of Health, Office of Hazard Evaluation and Emergency Response.

HIDOH, 2011, *Screening for Environmental Concerns at Sites with Contaminated Soil and Groundwater* (Fall 2011 and updates): Hawai’i Department of Health, Office of Hazard Evaluation and Emergency Response/

ITRC, 2012, Incremental Sampling Methodology: Interstate Technology Regulatory Council, February 2012.

Minnitt, R.C.A., Rice, P.M. and C. Spangenberg, 2007, Part 1: Understanding the components of the fundamental sampling error: a key to good sampling practice: The Journal of the Southern African Institute of Mining and Metallurgy, August 2007, Vol. 107.

Pitard, F., F., 1993, Pierre Gy’s Sampling Theory and Sampling Practice: CRC Press, New York, NY.

Pitard, F.F., 2005, Sampling Correctness - A Comprehensive Guideline: Sampling and Blending Conference Sunshine Coast, Queensland, Australia, May 9-12, 2005.

Pitard, F.F., 2009, Theoretical, practical and economic difficulties in sampling for trace constituents: Fourth World Conference on Sampling & Blending, The Southern African Institute of Mining and Metallurgy, 2009.

Ramsey, C. A. and A.D. Hewitt, 2005, A Methodology for Assessing Sample Representativeness: Environmental Forensics, 6:71–75, 2005.

USEPA, 1985, *Verification of PCB Spill Cleanup by Sampling and Analysis*: U.S. Environmental Protection Agency, Office of Toxic Substances, EPA-560/5-85·026, August 1985, Washington DC.

USEPA, 1986, *Field Manual for Grid Sampling of PCB Spill Sites to Verify Cleanups*: U.S. Environmental Protection Agency, Office of Toxic Substances, EPA-560/5-86·017, May 1986, Washington DC.

USEPA, 1987, *Data Quality Objectives for Remedial Response Activities*: U.S. Environmental Protection Agency, Office of Emergency and Remedial Response, EPA/540/G-87/003, March 1987, Washington DC.

USEPA, 1988, *Superfund Exposure Assessment Manual*: U.S. Environmental Protection Agency, Office of Remedial Response, EPA/540/1-881001, April 1988, Washington DC.

USEPA, 1989a, *Methods for Evaluating the Attainment of Cleanup Standards, Volume 1: Soils and Solid Media*: U.S. Environmental Protection Agency, Office of Policy, Planning, and Evaluation, EPA 230, U2-89-042, February 1989, Washington DC.

USEPA, 1989b, *Risk Assessment Guidance for Superfund, Volume I, Human Health Evaluation Manual (Part A):* U.S. Environmental Protection Agency, Office of Policy, Planning, and Evaluation, EPA/540/1-89/002, December 1989, Washington DC.

USEPA, 1989c, *Risk Assessment Guidance for Superfund, Volume II, Environmental Evaluation Manual:* U.S. Environmental Protection Agency, Office of Policy, Planning, and Evaluation, EPA/540/1-89/001, March 1989, Washington DC.

USEPA, 1990, *Guidance on Remedial Actions for Superfund Sites with PCB Contamination:* U.S. Environmental Protection Agency, Office of Emergency and Remedial Response, EPA/540/G-90/007, August 1990, Washington DC.

USEPA, 1991, *Guidance for Data Usability in Risk Assessment*: Environmental Protection Agency, Office of Research and Development, EPA/540/R-92/003, December 1991, Washington DC.

USEPA 1992a, *Preparation of Soil Sampling Protocols: Sampling Techniques and Strategies*: Environmental Protection Agency, Office of Research and Development, EPA/600/R-92/128, July 1992, Washington DC.

USEPA, 1992b, *A Supplemental Guidance to RAGS: Calculating the Concentration Term*: U.S. Environmental Protection Agency, Office of Solid Waste and Emergency Response, EPA 9285.7-081, May 1992, Washington DC.

USEPA, 1998, 40 CFR Part 761.61, PCB Remediation Waste: U.S. Environmental Protection Agency, Code of Federal Regulations.

USEPA, 2003, Guidance for Obtaining Representative Laboratory Analytical Subsamples from Particulate Laboratory Samples: U.S. Environmental Protection Agency, Office of Research and Development, EPA/600/R-03/027, November 2003, Washington DC.

USEPA, 2005, *Guidance on Surface Soil Cleanup at Hazardous Waste Sites: Implementing Cleanup Levels (Peer Review Draft)*: U.S. Environmental Protection Agency, Office of Emergency and Remedial Response, EPA 9355.0-91, April 2005.

USEPA, 2014a, Hot Spots: Incremental Sampling Methodology (ISM) FAQs: U.S. Environmental Protection Agency, Superfund, March 27, 2014.

USEPA, 2014b, *Screening Levels for Chemical Contaminants*: U.S. Environmental Protection Agency, (May 2014), prepared by Oak Ridge National Laboratories.