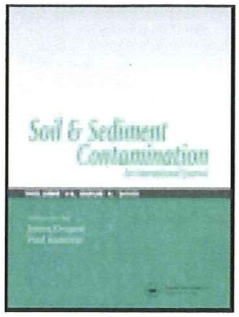

Taylor & Francis
Taylor & Francis Group

Soil and Sediment Contamination: An International Journal

ISSN: 1532-0383 (Print) 1549-7887 (Online) Journal homepage: http://www.tandfonline.com/loi/bssc20

A Critical Review of Discrete Soil Sample Data Reliability: Part 2—Implications

Roger Brewer, John Peard & Marvin Heskett

To cite this article: Roger Brewer, John Peard & Marvin Heskett (2016): A Critical Review of Discrete Soil Sample Data Reliability: Part 2—Implications, Soil and Sediment Contamination: An International Journal, DOI: 10.1080/15320383.2017.1244172

To link to this article: http://dx.doi.org/10.1080/15320383.2017.1244172

SOIL AND SEDIMENT CONTAMINATION
2016, VOL. 0, NO. 0, 1–22
http://dx.doi.org/10.1080/15320383.2017.1244172

Taylor & Francis
Taylor & Francis Group

A Critical Review of Discrete Soil Sample Data Reliability: Part 2—Implications

Roger Brewer[a], John Peard[a], and Marvin Heskett[b]

[a]Hawaii Department of Health, Honolulu, HI, USA; [b]Element Environmental, Aiea, HI, USA

ABSTRACT

Part 2 of this study investigates the implications of random, small-scale contaminant concentration variability in soil for reliance on discrete soil sample data to guide environmental investigations. Random variability around an individual point limits direct comparison of discrete sample data to risk-based screening levels. "False negatives" can lead to premature termination of an investigation or remedial action. Small-scale distributional heterogeneity of contaminants in soil is expressed as artificial, seemingly isolated "hot spots" and "cold spots" in isoconcentration maps. Surgical removal of hot spots can lead to erroneous conclusions regarding the magnitude of remaining contamination. The field precision of an individual discrete sample data set for estimation of means for a contaminant in a risk assessment is not directly testable. Omission of "outlier" data in order to force data to fit a geostatistical model distorts estimates of mean concentrations and introduces error into a risk assessment. The potential for such errors was pointed out in early USEPA guidance but largely ignored or misunderstood. Decision Unit and *Multi Increment* sample investigation methods, long known to the agricultural and mining industries, were specifically developed to overcome these inherent shortcomings of discrete sampling methods and provide more reliable and defensible data for environmental investigations.

KEYWORDS

Soil sample; sampling theory; *Multi Increment* sample; incremental sampling methodology; environmental site investigation

Introduction

The field study presented in Part 1 of this paper (Brewer *et al.*, 2016) was designed to address a basic question: What is the variability of contaminant concentrations in soil around a fixed point at the scale of a typical, discrete soil sample? A significant variability in data for "co-located" or "split" samples as well as data for replicate analyses by the laboratory samples is often simplistically blamed on "laboratory error." The results of this study document that the error more likely lies in a misunderstanding of the heterogeneous nature of contaminants in particulate media such as soil (HDOH, 2015a,b).

CONTACT Roger Brewer roger.brewer@doh.hawaii.gov Hawaii Department of Health, 919 Ala Moana Blvd., Room 206, Honolulu, HI 96814, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/bssc.

Supplemental data for this article can be accessed on the publisher's website.

SOIL AND SEDIMENT CONTAMINATION
2016, VOL. 0, NO. 0, 1–22
http://dx.doi.org/10.1080/15320383.2017.1244172

Taylor & Francis
Taylor & Francis Group

A Critical Review of Discrete Soil Sample Data Reliability: Part 2—Implications

Roger Brewer[a], John Peard[a], and Marvin Heskett[b]

[a]Hawaii Department of Health, Honolulu, HI, USA; [b]Element Environmental, Aiea, HI, USA

ABSTRACT

Part 2 of this study investigates the implications of random, small-scale contaminant concentration variability in soil for reliance on discrete soil sample data to guide environmental investigations. Random variability around an individual point limits direct comparison of discrete sample data to risk-based screening levels. "False negatives" can lead to premature termination of an investigation or remedial action. Small-scale distributional heterogeneity of contaminants in soil is expressed as artificial, seemingly isolated "hot spots" and "cold spots" in isoconcentration maps. Surgical removal of hot spots can lead to erroneous conclusions regarding the magnitude of remaining contamination. The field precision of an individual discrete sample data set for estimation of means for a contaminant in a risk assessment is not directly testable. Omission of "outlier" data in order to force data to fit a geostatistical model distorts estimates of mean concentrations and introduces error into a risk assessment. The potential for such errors was pointed out in early USEPA guidance but largely ignored or misunderstood. Decision Unit and *Multi Increment* sample investigation methods, long known to the agricultural and mining industries, were specifically developed to overcome these inherent shortcomings of discrete sampling methods and provide more reliable and defensible data for environmental investigations.

KEYWORDS

Soil sample; sampling theory; *Multi Increment* sample; incremental sampling methodology; environmental site investigation

Introduction

The field study presented in Part 1 of this paper (Brewer *et al.*, 2016) was designed to address a basic question: What is the variability of contaminant concentrations in soil around a fixed point at the scale of a typical, discrete soil sample? A significant variability in data for "co-located" or "split" samples as well as data for replicate analyses by the laboratory samples is often simplistically blamed on "laboratory error." The results of this study document that the error more likely lies in a misunderstanding of the heterogeneous nature of contaminants in particulate media such as soil (HDOH, 2015a,b).

CONTACT Roger Brewer roger.brewer@doh.hawaii.gov Hawaii Department of Health, 919 Ala Moana Blvd., Room 206, Honolulu, HI 96814, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/bssc.

Supplemental data for this article can be accessed on the publisher's website.

The concentration of a chemical in a heterogeneous particulate media such as soil is directly tied to the mass of soil represented by the field sample and subsample tested by the laboratory. As documented in Part 1, discrete sample data provided by the laboratory cannot reliably be assumed to represent the sample submitted, and the sample submitted cannot reliably be assumed to represent the immediate area where the sample was collected. Data reported by a laboratory for a discrete soil sample can only be assumed to be representative of the subsample mass actually extracted and tested. The data for a given sample therefore may or may not have any relevance to the overall objectives of a site investigation.

The area, volume, and total mass of soil for which an average concentration of a chemical is desired are referred to in Sampling Theory as the "Decision Unit" (DU; HDOH, 2016; ITRC, 2012; see also Minnitt *et al.*, 2007; Pitard, 1993, 2009; Ramsey and Hewitt, 2005; USEPA, 1999). Note that all data for particulate media such as soil represent an "average" of the particles tested, regardless of how the sample was collected and analyzed. The concept of DUs is well understood in the agriculture, mineral exploration, and food industries, where it is referred to as a "batch" or "lot" (see also AAFCO, 2015). The level of available nutrients for a particular field or area of a field might be desired to assist in fertilizer application. Estimation of the mean concentration and ultimately the total mass of gold in a preliminarily identified ore deposit is needed to determine if the deposit is economically viable. In each case, the objective is to determine the concentration of a targeted chemical for the DU volume of media as a whole. Ideally, the entire DU would be submitted to a laboratory for analysis of a single unit. This is not practical in most cases, so a sample of the DU media must be collected.

A well-designed, step-by-step approach to representative sampling based on sampling theory has only recently been utilized for the investigation of potentially contaminated soil in the environmental industry (e.g., HDOH, 2016; ITRC, 2012). Testing of soil in many cases and many areas is still reliant on discrete sample methods described in Part 1 of this paper. Multiple factors are responsible for the continued use of discrete sampling and subsampling methods in the field and in the laboratory. The number of environmental investigations rose exponentially in the 1980s following passage of federal legislation such as the Resource Conservation and Recovery Act and the publishing of associated regulations and guidance. With little experience to go by, and only marginally aware of the field of Sampling Theory developing in other industries, the authors of early environmental guidance quickly adopted sample collection and testing methods already in place for evaluation of liquid wastes (e.g., USEPA, 1985, 1986, 1989a). Testing of small subsamples from a relatively small number of samples is common practice for these types of media, where the concentration of a contaminant in any given volume/mass of a waste stream can be assumed to be reasonably uniform. Test methods were designed to evaluate temporal rather than spatial variability in contaminant concentrations. The size of the sample collected was largely driven by laboratory needs with respect to the analytical method to be employed and any additional mass required for quality assurance and control measures. Collecting relatively few numbers of discrete samples intended to represent a specific contaminated area also limited costs associated with sample acquisition.

Authors of early soil sampling guidance assumed that a similar approach would be adequate to identify "hot spots" of contaminated soil that could pose a potential risk to human health (USEPA 1987; USEPA 1989a, 1991, 1992a; refer to supplement). This is depicted in Figure 1, taken from the USEPA Methods for Evaluating the Attainment of Cleanup

Standards guidance (USEPA 1989a). The figure is used to illustrate how an excessively large grid spacing might inadvertently miss large "hot spots" that would otherwise be detected with a single discrete soil sample.

Such methods are not, however, effective for heterogeneous particulate media such as soil, where spatial rather than temporal variability of contaminant concentrations is of primary importance. Unlike a liquid, and as demonstrated in Part 1 of this paper, the concentration of a contaminant in small masses of soil can vary dramatically and randomly both within an individual discrete sample and between closely spaced co-located samples. A laboratory will, of course, report a concentration for the mass of soil tested, but the relevance of the resulting data to the objectives of the investigation will be uncertain. Representative testing of contaminants in heterogeneous particulate matter requires much greater attention to the desired resolution of concentration data in terms of the site investigation objectives. This is carried out in the DU designation stage of an investigation (HDOH, 2016; ITRC, 2012).

Concern regarding potential error associated with reliance on traditional discrete soil sampling methods has been growing for some time (e.g., Hadley and Sedman, 1992; Pitard, 1993; Ramsey and Hewitt, 2005). As stated by Hadley and Petrisor (2013):

> 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. (p. 109)

In this paper, we briefly review the causes of random small-scale variability of contaminant concentrations in soil. We then use data from the field study described in Part 1 as well as other examples to explore the specific types of sampling error likely to be associated with the use and interpretation of discrete sample data in environmental investigations. In the Supporting Information provided with this paper, we also trace the roots of the entrenchment of discrete soil sampling methods in the environmental industry, highlight multiple calls for caution, and emphasize the eventual need for development of robust and reliable investigation methods.

Heterogeneity and hot spots

Distributional heterogeneity

The variability of contaminant concentrations observed between and within discrete soil samples collected during the study is controlled by three factors, each of which is a function of what Sampling Theory (Pitard, 1993) refers to as "distributional heterogeneity": 1) large-scale differences in the amount of the contaminant released in different parts of the study sites; 2) random, small-scale heterogeneity of contaminant distribution in soil at the scale of the sample collected (e.g., five grams to a few hundred grams); and 3) random, small-scale heterogeneity of contaminant distribution within an individual sample at the scale of the mass of soil analyzed by the laboratory (e.g., 0.5 to 30+ g). Large-scale variability is related to the release of greater amounts of a contaminant in one area, typically several hundred to several thousand square feet and a volume of several hundred to several thousand cubic

yards of soil (HDOH, 2016; ITRC, 2012). The identification of such areas and assessment of the potential risk to human health and the environment is the primary objective of most site investigations. The term "small-scale" variability is used in the context of this report to collectively describe random *intra-sample* and *inter-sample* variability of contaminant concentrations in discrete samples collected around an individual grid point, typically at the scale of a few grams to a few tens or hundreds of grams (refer to Part 1). While it is important to consider and capture small-scale variability when designing a sampling plan, attempt to characterize a site at the scale or resolution of a discrete sample is neither practical nor necessary in terms of evaluating potential risk to human health and the environment (see HDOH, 2016).

The magnitude of random variability increases as the sample mass decreases. Variability in contaminant concentrations within an individual 200-g discrete sample at the scale of a laboratory subsample (e.g., 0.5–30 g) might span one or more orders of magnitude. At some scale, perhaps the scale of an individual particle or even the coating on a particle, the minimum and maximum concentrations of a contaminant in soil will necessarily be 0% and 100%. Attempt to identify the "maximum" concentration of a contaminant in soil at the arbitrary scale of a discrete soil sample or laboratory subsample is both impractical and again irrelevant in terms of evaluating potential risk to human health and the environment. The maximum concentration of a contaminant identified in a small set of discrete samples collected within an area cannot be assumed to represent the maximum concentration of the contaminant present for the tested mass of soil. A relatively small 100 m^2 area to a depth of 5 cm will contain, for example, approximately 5,000 kg of soil. Identification of the maximum concentration of a contaminant in any given 0.5 g, 30 g, or even 200 g mass of soil within this area would require an enormous amount of sample collection and provide no added benefit to the objectives of the site investigation.

The cause of random variability of contaminant concentrations in discrete soil samples is straightforward: the mass of the sample and the area over which the sample is collected are too small to overcome random distributional heterogeneity of the contaminant within the soil. This dilemma, well known in the agriculture and mining industry, is identified as "Fundamental Error" in the Gy Sampling Theory (Pitard, 1993, 2009; see also Minnitt *et al.*, 2007; Ramsey and Hewitt, 2005; USEPA, 1999). 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 a representative, unbiased manner (e.g., collection of adequate sample mass from an adequate number of locations with an appropriate collection tool in both the field and the laboratory). Error associated with random distributional variability of a contaminant within a sample, referred to as "*intra-sample* variability" in Part 1, can in theory be largely eliminated by the use of proper field collection, processing, and laboratory subsampling techniques (Minnitt *et al.*, 2007; Pitard, 1993, 2005, 2009). Error associated with random distributional heterogeneity between closely spaced discrete soil samples, referred to as "*inter-sample* variability" in Part 1, cannot be eliminated, since this is an inherent property of the soil under investigation (Minnitt *et al.*, 2007; Pitard, 1993, 2005, 2009). This error can, however, be minimized through the use of Decision Unit and *Multi Increment* sample (DU-MI) investigation approaches for soil (HDOH, 2016), also referred to as Incremental Sampling Methodology or "ISM" (ITRC, 2012; The term "Multi Increment®" is trademarked by Charles Ramsey and EnviroStat, Inc.; see Ramsey and Hewitt, 2005.)

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* discussed the need for "representative sampling" (USEPA, 1990):

> 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. (p. 5)

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, 1990):

> 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. (p. 3)

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

> 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. (p. 3)

Random small-scale variability of contaminant concentrations in small masses of soil is predicted by sampling theory, but outside of munitions-related sites, had not been widely studied in the field (e.g., see USACE, 2009). 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.

Implications

Comparison of discrete sample data to screening levels

Direct comparison of discrete sample data points to screening levels can lead to significant errors in environmental investigations. Risk-based soil screening levels, including the Regional Screening Levels (RSLs) published by the USEPA (USEPA, 2015), are intended for comparison to the concentration of a contaminant within a targeted area of concern or "Decision Unit" as a whole (i.e., the "average") rather than discrete points within this area. This was made clear in early USEPA soil sampling guidance (USEPA, 1989a, emphasis added; see also USEPA, 2014 and Supporting Information):

> 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. (p. 6–19)

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., 100–200 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. The USEPA document *Guidance on Surface Soil Cleanup at Hazardous Waste Sites* (USEPA, 2005) notes:

> 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. (p. 5)

Grids of discrete data can sometimes be useful for gross approximation of contaminated versus clean areas (HDOH, 2016). The reliability of the data for final decision-making depends in part on the magnitude of small-scale variability of contaminant concentrations in soil with respect to the screening level being used.

Consider, for example, the range of lead concentrations estimated for discrete samples around grid points at Study Site B (refer also to Part 1 supplement). Box plots of total estimated variability depicted in Figure 2 fall both above and below the HDOH residential soil action level for lead of 200 mg/kg (HDOH, 2011) at 23 of the 24 grid points. Discrete sample concentrations at 20 of the 24 grid points similarly fall both above and below the USEPA residential soil screening level of 400 mg/kg (USEPA, 2015). The wide range of estimated concentrations matches well with the assumed incomplete mixture of lead-contaminated ash

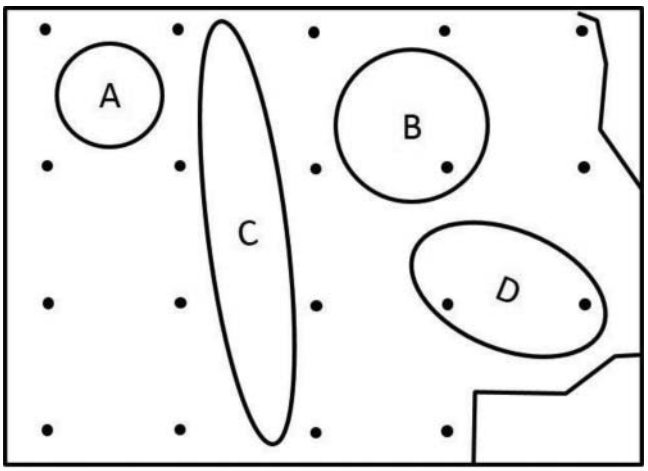

Figure 1. Discrete sampling grid designated for a site under investigation overlaid with hypothetical "hot spots" superimposed (USEPA, 1989a). Under this approach, an individual discrete soil sample was assumed to be adequate to identify large areas of contamination above potential levels of concern.

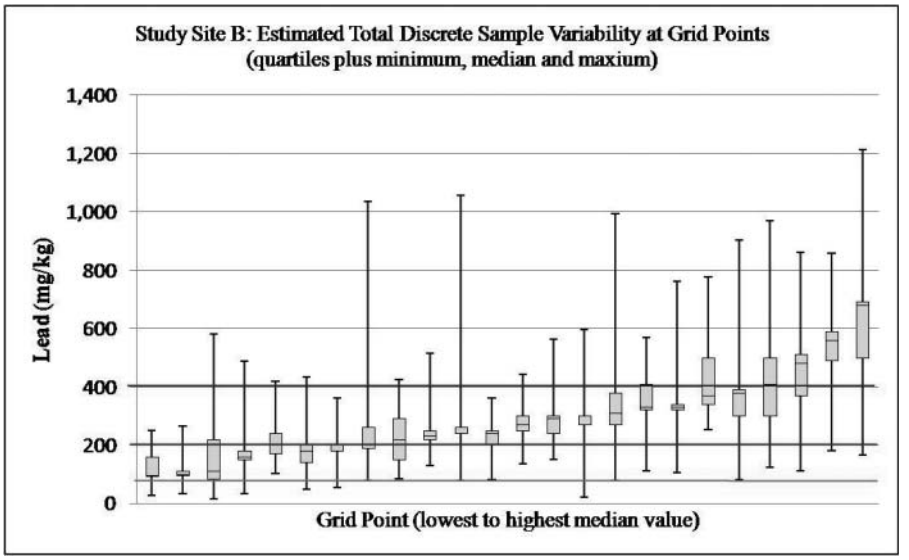

Figure 2. Box plots depicting estimated total variability of lead concentrations in discrete samples within 0.5 m of grid points at Study Site B (lowest to highest median for *inter-sample* data). Estimated range of lead concentrations falls both above and below HDOH residential soil action level of 200 mg/kg at 23 of 24 grid points and above USEPA residential screening level of 400 mg/kg at 20 of 24 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-contaminated ash.

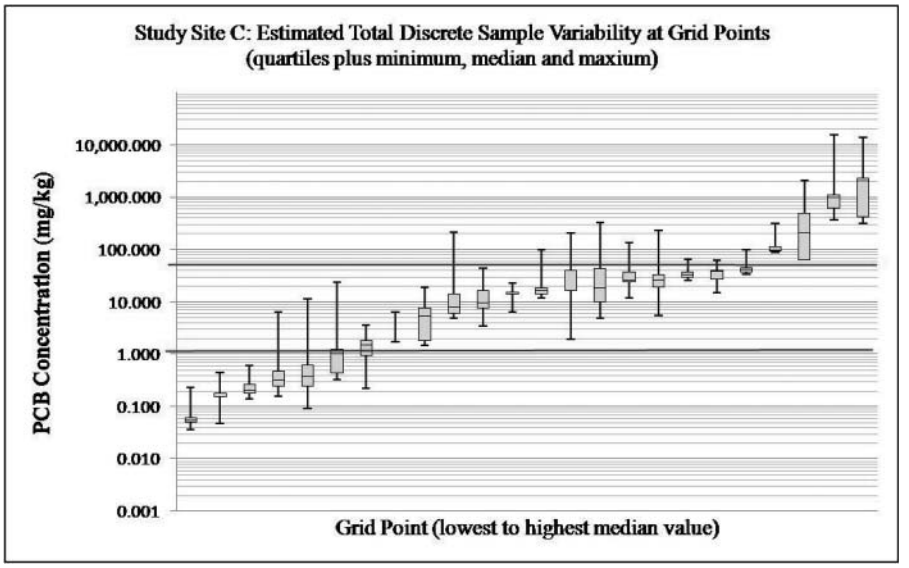

Figure 3. Box plots depicting estimated total variability of total PCB concentrations in discrete samples within 0.5 m of grid points at Study Site C (combined *intra-* and *inter-sample* variability; note use of log scale for vertical axis; lowest to highest median values for *inter-sample* data). Hawaii Department of Health Residential PCB soil screening level of 1.1 mg/kg and USEPA TSCA level of 50 mg/kg noted for reference.

and fill soil at the site. Reported concentrations of lead in soil below 100 mg/kg imply that the subsample tested consisted primarily of native fill material, with concentrations approaching natural background (upper threshold limit 73 mg/kg; HDOH, 2012). Higher reported concentrations of lead imply a more significant proportion of incinerator ash in the tested soil (typically 1,000–4,000 mg/kg; Shulgin and Duhaas, 2008).

Box plots for data from Study Site C depict the extreme variability of total polychlorinated biphenyls (PCBs) concentrations both in subsamples of individual discrete samples as well as in estimated total variability around individual grid points when data for processed samples are considered (Figure 3). Lines denoting screening levels of 1.1 mg/kg (HDOH residential screening level) and 50 mg/kg are included in the graph. Note the random variability of PCB concentrations both above and below these levels at multiple grid points across the study area. 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 highly prone to false negatives and early termination of the investigation. As discussed in the "Environmental risk assessment" discussion below, such high variability can also confound estimation of mean PCB concentrations for targeted exposure areas.

The implications are significant. Data for discrete soil samples cannot be reliably assumed to represent either the soil immediately surrounding a sample collection point or the sample submitted to the laboratory for analysis. Direct comparison of data for individual grid points could in theory declare the site to be either completely "clean" (i.e., all discrete samples $\leq$200 mg/kg lead) or completely "contaminated" (i.e., all discrete samples $>$200 mg/kg lead) depending on the mass of soil randomly collected for testing around a particular grid point.

Estimation of extent of contamination

The use of discrete sample data to define large-scale contamination patterns of potential interest is highly prone to error. Drawing a line between "contaminated" and "clean" areas of a site for characterization and risk assessment is fundamental to environmental investigations and necessary to design appropriate remedial actions. As is obvious from the previous discussion, random small-scale variability of contaminant concentrations can significantly affect the accuracy of discrete soil sample data to estimate the lateral and vertical extent of larger-scale patterns of interest. Consider the following text from the USEPA document *Data Quality Objectives for Remedial Response Activities* (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. (p. C-4)

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 the precision of estimated mean contaminant concentrations within targeted areas (e.g., USEPA, 1987, 1990, 1991). The cost of replicate sample collection and the premature conclusion in USEPA guidance that the variability of contaminant concentrations within an individual sample would

Figure 4. Irregular and disconnected spill patterns on soil made by a release of milk. This might also mimic vertical patterns for subsurface releases of liquids.

be minimal negated serious efforts to evaluate this critical issue in more detail (USEPA, 1989a):

> When there is little distance between points it is expected that there will be little variability between points. (p. 10-2)

Consider again the example from Study Site B above. The occurrence of "false negatives" and premature termination of an investigation using a progressive, step-out discrete sample collection approach is unavoidable. At some random point, the reported concentration of lead in an individual sample will fall below the target screening level although the concentration of lead for the area as a whole is still well above the screening level. The potential for this type of field error was recognized in early USEPA guidance documents (USEPA, 1991):

> 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. (p. 40)

Larger scale heterogeneity in the manner in which a contamination was released to the soil can also be expected to confound attempts to determine the extent of contamination based on discrete sample data. Refer, for example, to the pattern of "contamination" in soil caused by an overturned milk truck in Figure 4. Assume that the milk was present but not visible to field investigators, as is the case for most contaminants. 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.

The same potential for error exists in the use of discrete sample data to assess the vertical extent of contamination. Random small-scale variability of contaminant concentrations in soil limits the reliability of interpolation between individual discrete sample points. This is due to the fact that the sample is collected over an area too small to capture smaller scale

Table 1. Range and sum of *intra-sample* and *inter-sample* relative standard deviation (RSD) of discrete sample variability around individual grid points (refer to Part 1 Supporting Information).

	Intra-sample data	*Inter-sample* data	Combined data
Study site	Range RSD (%)	Range RSD (%)	Range RSD (%)
Site A (arsenic)	4.8–30	1.5–38	9–52
Site B (lead)	20–96	11–81	44–139
Site C (PCBs)	17–277	14–151	58–336

random heterogeneity within the overall spill area. Such confounding factors are the primary cause of "failed" confirmation samples and of the need for repeated sample collection and over-excavation of contaminated soil with no clear end point in sight.

Interpretation of isoconcentration maps

The problems discussed previously become readily apparent in computer- or hand-generated isoconcentration maps of contaminant distribution. Use of geostatistical methods to interpolate contaminant concentrations between discrete sample 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, and 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 are 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. As demonstrated in the field study presented in Part 1 of this paper, this requirement is unlikely to be met on a point-by-point basis for typical discrete sample data.

Table 1 summarizes the Relative Standard Deviation (RSD) measured and estimated for discrete samples around individual grid points at each of the study sites (refer to Part 1 Supporting Information). The RSDs estimated for total variability around grid points vary widely between individual grid points both within and between the sites, ranging from 9% to 52% at Study Site A (arsenic), 44% to 139% at Study Site B (lead), and 58% to 336% at Study Site C (total PCBs). Additional sample collection and testing would likely be resulting in a higher RSD. Although larger scale contaminant distribution patterns might indeed be real, small-scale patterns generated by a single point or even a small cluster of points could be random artifacts of small-scale heterogeneity and not reproducible.

This has significant implications for attempts to remove apparent isolated "hot spots" in order to reduce the overall mean concentration of a contaminant within a targeted area, referred to as "Iterative Truncation" in some USEPA guidance documents (USEPA, 2005; refer to Supporting Information). Consider the removal of a few randomly selected higher concentration sample data represented in the bar graphs for Study Site B depicted in Figure 3. Removal of such spots cannot be considered to have significantly reduced the mean concentration of the contaminant in soil for the targeted area as a whole. In addition to the likelihood of "failed" confirmation samples following removal, the remaining sample points could no longer be representative of the

variability of contaminant concentrations at the scale of a discrete sample. If a new, independent set of samples were to be collected, then a similar random and artificial pattern of isolated "hot spots" and "cold spots" as originally identified would again be generated, but in different locations. The same holds true for selective removal of soil around discrete sample data points at Study Site C, where a concentration of >50 mg/kg PCBs was reported for a discrete sample randomly collected around the point (see Figure 3). While they shed some light on the range of contaminant concentrations within the targeted area as a whole, data for any individual point cannot be considered to be representative of the area around that point and, in the absence of other information (e.g., obvious staining or other direct signs of contamination in the field), cannot defensibly be relied upon for design of remedial actions.

The presence of artificial small-scale patterns of contaminant distribution is exemplified in a series of isoconcentration maps prepared for Study Site A (Figure 5). The maps depict patterns generated based on separate groupings of data for different groups of processed,

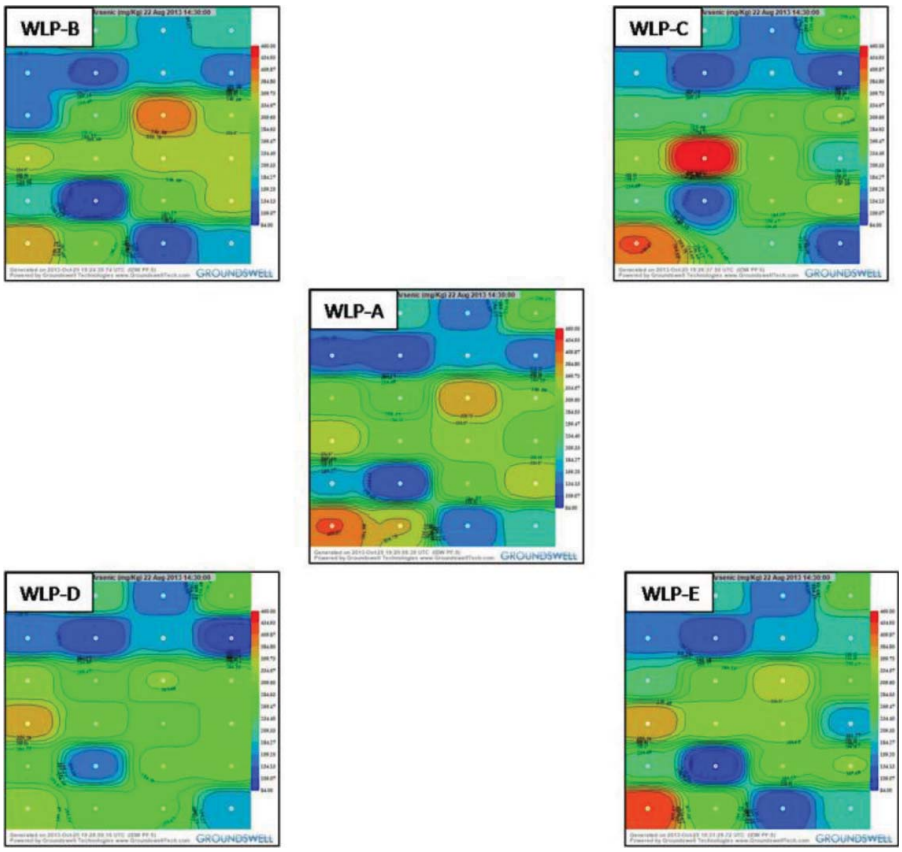

Figure 5. Changing locations of isolated "hot spots" and "cold spots" depending on use of arsenic data for "A," "B," "C," "D," or "E" processed sample sets for Study Site A (Groundswell Technologies, Inverse Distance Weighted Power Function = 5). Red tones represent higher concentrations. Blue tones represent lower concentrations. Individual spots represent approximately 900 ft^2 area (refer to Part 1; Grid Point #1 in lower left hand corner). Changing patterns reflect random small-scale variability of arsenic concentrations around individual grid points and use of an unrealistically high isoconcentration mapping power function.

discrete samples collected around each of the 24 grid points (samples sets "A," "B," "C," "D," and "E"; see Figure 2 in Part 1). 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 the figure. 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, 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. 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-g discrete soil samples is real, but the map patterns generated from the data are not. This is because the concentration of the contaminant in any given 200 g mass of soil from the grid cell area is likely to be random with respect to concentrations in immediately adjacent soil. Collection and testing of an independent set of co-located samples might yield a similar degree of variability, but the apparent distribution of this variability within the grid soil might be very different.

Over-interpretation of individual discrete sample data points can be addressed in part by selection of a mapping option that de-emphasizes data for individual points and instead focuses on apparent larger scale patterns (HDOH, 2015b). This is normally accomplished by selection of a lower value "distance decay parameter" value in mapping program. However, most mapping software is still unlikely to be fully able to overcome random small-scale

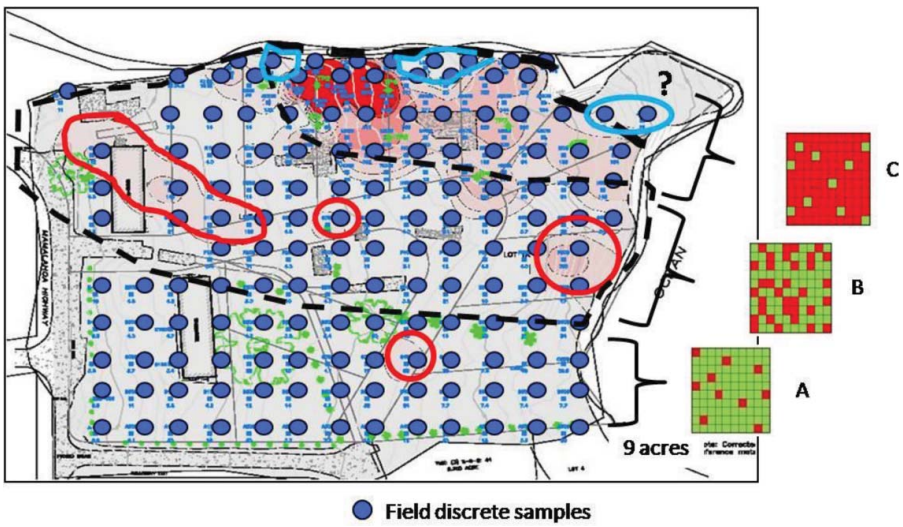

Figure 6. Isoconcentration map generated from discrete soil sample data collected at a known arsenic-contaminated site on the Island of Hawaii (after ERM, 2008). Random small-scale variability of arsenic concentrations in soil at the scale of a discrete sample are expressed in the map as isolated "hot spots" and "cold spots," particularly within the transitional area (Zone B) that separates areas of consistently low (Zone A) and high (Zone C) arsenic concentrations. Red and green cells in inserts to the right of the map illustrate hypothetical distribution of discrete sample points above (red) and below (green) the target arsenic screening level within each zone.

variability of contaminant concentrations around and between individual grid points, and care must be taken in the use of such maps for final decision making purposes.

Small-scale contaminant patterns reflective of random heterogeneity are endemic in computer-generated isoconcentration maps of discrete sample data. Consider, for example, the nine-acre site on the Island of Hawaii depicted in Figure 6. The site was formerly used to mix and store arsenic-based herbicides and was being considered for residential redevelopment. A tight grid of discrete samples was collected across the site in order to help identify large scale contamination patterns (ERM, 2008). The figure depicts an isoconcentration map generated from the discrete data grid points. A background threshold value of 24 mg/kg total arsenic was used to screen the site, with red shades in excess of this level (HDOH, 2012).

A large area of heavy contamination at the northern edge of the site is clearly apparent from the discrete sample data. Three large-scale zones of arsenic concentrations in soil are apparent on the map. 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. 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 random small-scale variability of arsenic concentrations in soil rather than actual large-scale areas of higher or lower concentrations.

Such artificial "Zone B" type patterns are readily apparent in isoconcentration maps used for presentation of contaminant distributions in soil. Examples of large-scale patterns of background metal concentrations in soil reflective of the underlying geology as well as presumed artificial small-scale patterns reflective of random small-scale variability are evident in nationwide isoconcentration maps recently published by the U.S. Geological Survey (USGS, 2014). A review of these maps is included in Part 2 of the original report for HDOH field study (HDOH, 2015b). In one case, data for a single composite sample collected from a 1-m square area are extrapolated to imply the presence of a 2,400 km^2 "hot spot" of arsenic-contaminated soil in a geologic terrane known to be highly heterogeneous. This is almost certainly an artifact of random small-scale heterogeneity that would not be reproducible in the field, and the document appropriately cautions users not to over-interpret data on a point-by-point basis. A more detailed discussion of the use of "inverse distance weighting" methods and "power functions" to help reduce, but not fully eliminate, such inherent errors in isoconcentration maps is also provided (Lu and Wong, 2008).

Environmental risk assessment

Estimation of the mean contaminant concentration for a targeted area and volume of soil is a key element of environmental risk assessment (USEPA, 1987, 1988, 1989a,b,c,d, 1991, 1992a, 2011b; see also USEPA, 2014). The accuracy of the estimate in terms of bias and precision is

a function of multiple factors, including (see Minnitt *et al.*, 2007; Pitard, 1993, 2009): 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. The effect of each of these factors on the estimated mean is in theory evaluated during the "data validation" stage of a project (USEPA, 2002a). In practice, data validation procedures primarily focus on data quality objectives associated with analytical measurements. The precision and reproducibility of the data in terms of field representativeness are rarely if ever directly and adequately assessed.

The USEPA *Supplemental Guidance to RAGS: Calculating the Concentration Term* (USEPA, 1992a) document suggests that a minimum of 20–30 discrete soil samples is required to reliably estimate the mean concentration of a contaminant in soil for a targeted area (USEPA, 1992a):

> Sampling data from Superfund sites have shown that… data sets with 20–30 samples provide fairly consistent estimates of the mean (i.e., there is a small difference between the sample mean and the 95 percent UCL). (p. 3)

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 in the 1980s (USEPA, 1991; refer to Exhibit 23 in document). The manner in which the evaluation was carried out is not discussed in the documents, however. Recommendations for the use of independent replicate sets of discrete sample data to assess field representativeness were recognized to likely be inadequate early on (USEPA, 1990, annotations and emphasis added):

> 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 (replicate) samples required to detect random bias will depend on the distribution of the biasing errors, and this distribution will generally be unknown… *The importance of pilot studies to the overall monitoring effort cannot be stressed enough.* (p. 9)

The "random bias" that the guidance document warns against is the random small-scale variability highlighted in Part 1 of this paper. The document also stressed the need for "pilot studies" to assess the reliability of grid and discrete sample characterization methodologies being proposed and published at that time. To the authors' knowledge, such studies were never carried out to the same level of detail presented in Part 1 of this paper.

The USEPA ProUCL guidance for the statistical evaluation of discrete sample data sets states the size of a discrete sample data set should be based on "appropriate DQOs processes," with a minimum of ten data points recommended (USEPA, 2013). The document is largely referring to data quality objectives (DQOs) for the desired statistical precision of the test used, rather than a more holistic sense of DQOs in terms of both statistical and field precision. Methods are provided to estimate the number of discrete samples required to achieve an acceptable level of precision in the estimate of a mean based on the variance measured for an initial set of samples, but the numbers generated are often well beyond the financial resources available for the project. The authors acknowledge this complication in the document (USEPA, 2013, annotations added):

> Due to resource constraints, it may not be possible to collect as many samples as determined by using a DQOs based sample size formula… It is suggested to collect at least 10 (discrete samples) before using statistical methods. (p. 24)

The authors are clearly uncomfortable with the use of statistical tests to draw conclusions from what they would consider an inadequately representative set of data (USEPA, 2013; annotations added):

> Statistics (derived from dataset that does not meet DQO goals) may not be considered representative and reliable enough to make important cleanup and remediation decisions. It is recommended not to use those statistics to draw cleanup and remediation decisions potentially impacting human health and the environment. (p. 24)

In practice, however, investigators are most often left with little other recourse once funds for sample collection and laboratory analysis have been expended, and decisions must be made on how to proceed forward.

Potential problems with the representativeness of discrete sample sets are further recognized, but not fully explored, in more recent USEPA guidance (USEPA, 2005, emphasis added):

> 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*… Uncertainty associated with sampling error can be very large, particularly at sites where there is significant spatial heterogeneity in contaminant concentrations. (p. 24)

The effect of field error in estimation of the mean contaminant concentration for a targeted area is highlighted by estimations of means and 95% Upper Confidence Limit (UCL) values for random non-stratified groupings of ten discrete sample data points at Study Site C (HDOH, 2015b). The USEPA ProUCL software was then used to generate a 95% UCL of the arithmetic mean for data set (USEPA, 2013). Twenty iterations of random data groupings were carried out for each set of study site data. The variance between estimated UCLs for each study site was then used to assess the precision of random sets of discrete soil samples to characterize a site.

The range in estimates of the mean arsenic concentration in soil for the study area again reflects the high combined small- and large-scale variability of total PCB concentrations in the soil identified for the study area in Part 1 (see Part 1 Supporting Information). 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 RSDs for the data point groupings are similarly high, with a range of 124–315%. Individual RSDs suggest a very poor precision in the estimate of mean concentration of PCBs estimated for any given ten-point set of discrete samples and the statistical tests used.

Results for the arsenic and lead study sites were less dramatic but still significant. Calculated 95% UCL arsenic concentrations for random ten-point groupings of discrete sample data range from 403 to 776 mg/kg. RSDs for the groupings range from 34% to 67%. Calculated 95% UCL lead concentrations for the random groupings of discrete sample data at

Study Site B ranged from 201 to 439 mg/kg, with a median of 345 mg/kg and a mean of 343 mg/kg, with a corresponding RSD range of 20–86%.

Compare these results to DU-MI triplicate sample data collected for each site (refer to Part 1). A 95% UCL of 259 mg/kg was calculated for arsenic at Study Site A with an RSD of only 6.5%, inferring very good total precision. A 95% UCL of 383 mg/kg was calculated for lead at Study Site B, with a slightly higher but still strong RSD of 20%. A 95% UCL of 346 mg/kg was calculated for PCBs at Study Site C, with the concentration of PCBs in individual MI samples reported at 19, 24, and 270 mg/kg. The replicate data RSD of 138% immediately flagged the data as unreliable due to low field precision, however. Such a test of field precision cannot be carried out on an individual set of discrete sample data, and biases due to field error could go unnoticed. These evaluations are for illustration purposes only. The range of RSDs and estimated means indicates a similarly poor lack of field precision for any given data set. Larger discrete sample databases for each site could be expected to identify even greater variance between random sample sets of data and lower total precision of any given set of data.

Several USEPA guidance documents also discuss the use of discrete soil samples to assess the presence or absence of very small "hot spots" that could pose "acute" toxicity risks (USEPA, 1989a, 1992b, 2005; see also Supporting Information). Data were to be compared to risk-based screening levels for acute toxicity or "not-to-exceed" concentrations. While understandable in concept, this approach suffers from two significant flaws. Acute toxicity factors, reflecting health effects within minutes to a few days following exposure (USEPA, 2011a), are not available for direct ingestion for the majority of contaminants assessed as part of a typical environmental investigation. Similarly, "acute" or "not-to-exceed" soil screening levels have never been published by the USEPA, to the authors' knowledge.

None of the documents noted provide either guidance on the calculation of such hypothetical soil screening levels or guidance on sampling methods to establish with any degree of reliability the presence or absence of contaminated soil that could pose such concerns. Assessment of acute toxicity would by necessity need to be tied to a target mass of incidentally ingested soil, such as a default of 10 g assumed to be ingested by a pica child (USEPA, 2011b). Each 10-g mass of soil at a site then becomes an individual "Decision Unit." For comparison, a relatively small 10 m $\times$ 10 m exposure area to a depth of just one centimeter contains approximately 1,000 kg of soil, or 100,000 potential 10 g DUs. Were acute toxicity factors and screening levels in fact available, the level of effort to demonstrate with an acceptable level of confidence, if such a level exists for theoretical acute toxicity, would be enormous and not feasible from either a technical or financial standpoint.

Decision-making is instead made based on remediation to meet potential long-term chronic health risk from exposure to much lower concentrations of the contaminant, with a perhaps unspoken assumption that this will also address hypothetical acute risks. The authors are unaware of any cases where such an approach has been demonstrated to be inadequate. If acute health risks are indeed a concern at a site 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). This might include, for example, concerns regarding the incidental ingestion by children of lead-based paint chips or lead shot randomly scattered in soil. Soil samples could also be ground to help assess the potential presence of large nuggets of targeted contaminants (HDOH, 2016; ITRC, 2012).

Similar debate and confusion exist in the interpretation and use of apparent "outlier" discrete sample data as part of an environmental investigation. 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 are therefore carefully designed to capture and represent "outliers" in order to make sound decisions (Pitard, 1993). Over-representation can lead to overestimates of the mass of the targeted mineral present and subsequent economic failure of the venture. Under-representation can lead to missed opportunities and shortages of minerals critical for economic development.

The same concepts apply to the investigation of contaminants in soil. The importance of capturing the full distributional heterogeneity of a contaminant in a targeted area and volume of soil is recognized in the USEPA guidance document *Methods for Evaluating the Attainment of Cleanup Standards* (USEPA, 1989a):

> 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. (p. 2–16)

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

> 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. 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). (p. vi)

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 erroneous for testing of particulate matter such as soil. The objective of any soil investigation is to estimate the mean concentration a contaminant within a designated DU area and volume of soil, regardless of how large or how small the desired DU is. 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, such as removing chips of lead-based paint, would yield erroneous data and conclusions. A mineral exploration company certainly would never omit "outlier" concentrations of gold in veins running through an ore body.

This conflict was recognized in earlier USEPA guidance on the estimation of mean contaminant concentrations in exposure areas (USEPA, 2002b):

> There are a variety of statistical tests for determining whether one or more observations are outliers. These tests should be used judiciously, however. It is common that the distribution of concentration data at a site is strongly skewed so that it contains a few very high values corresponding to local hot spots of contamination. The receptor could be exposed to these hot

> spots, and to estimate the EPC correctly it is important to take account of these values. Therefore, one should be careful not to exclude values merely because they are large relative to the rest of the data set. (p. 3)

The problem with "outliers" lies not with the statistical test used but with the inappropriately small scale of observation that the sample data represent. The concentration of a chemical in soil at the scale of an individual discrete soil sample or subsample tested by a laboratory has no direct relevance to the assessment of health risk. 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 (sic) operator. (p. 34)

Pitard repeatedly emphasizes the need for sampling methods that accurately represent all parts of the investigation area (Pitard, 2005, annotations and emphasis added):

> *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). (p. 56)

This includes the need to retain "outlier" data (Pitard, 2009, emphasis added):

> 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*. (p. 5)

Pitard notes that exclusion of "outlier" data can lead to significant error in decision-making (Pitard, 1993):

> …the above sampling protocol (e.g., discrete samples and 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. (p. 173)

The recommendation in the ProUCL guidance to similarly ignore "non-detect (ND)" results in the statistical evaluation of a data set is likewise inappropriate for soil data (USEPA, 2013; see also USEPA, 2002b). 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 used to estimate a mean from both a field and statistical standpoint, rather than an error in the data provided.

Summary and discussion

The results of this field study highlight the need to transition from traditional discrete soil sample investigation methods to more science-based and reproducible Decision Unit and *Multi Increment* sampling methodologies. The data are irrefutable that the concentration of a contaminant reported by a laboratory for a discrete soil sample cannot reliably be assumed to be representative of the sample provided. A single sample (or even small group

of samples) likewise cannot reliably be assumed to be representative of the area from which it was collected or reliably indicative of large-scale trends of potential interest. This can lead to significant but largely hidden errors when discrete sample data are used as a basis for decision-making in environmental investigations. However, false negatives, false positives, confusion over seemingly isolated, and potentially artificial "hot spots" and "cold spots," inappropriate omission of "outlier" data in environmental risk assessments, and the lack of sufficient replicate data to verify the field precision of data sets are unavoidable. Although useful in some cases for rough approximation of large-scale contaminant patterns, discrete soil sampling methodologies are at best highly inefficient and wasteful of resources, and at worst, highly misleading.

The fact that it has taken over 30 years to begin to address this problem is attributable to multiple factors, including: 1) the lack of training of environmental regulators and consultants in more up-to-date concepts of Sampling Theory for particulate media, 2) the lack of field studies to investigate and quantify potential error in discrete sample data, 3) the mistaken assumption that disparities between replicate data, when collected, were due primarily to laboratory error, and 4) the lack of a final test of data quality comparable to those of the agriculture and mining industry to assess data reproducibility and the corresponding lack of market forces to push controlling regulatory agencies to make the process more efficient and effective. It is hoped that the field study presented in this paper will in part help to address these deficiencies.

The much-needed transition to Decision Unit and *Multi Increment* sampling methodologies will necessarily be disruptive to regulatory agencies and consultants entrenched in discrete sample investigation methods. The State of Hawaii instituted the change over a period of several years, beginning in 2004 and publishing the first formal guidance in 2008. An estimated 15,000+ MI samples have since been collected. Previously completed investigation and remediation actions were only revisited in a small number of cases, and then usually only as part of a new property redevelopment or transaction. Field data typically indicate that the core of contamination was indeed removed, although the process was highly inefficient in terms of time and cost, and outer areas of moderate contamination were sometimes overlooked. The use of DU-MI methodologies was encouraged but not necessarily required for projects already underway, particularly where work plans for site characterization had already been completed. The collection of DU-MI data to confirm the results of site investigations and remedial actions was especially recommended. Intensive training for regulators and consultants in Sampling Theory and the implementation of DU-MI in the field was carried out during the same period and continues on a regular basis. Experience gained was progressively incorporated into the state's *Technical Guidance Manual*, which continues to be updated as more efficient and effective investigation methods are developed (HDOH, 2016).

The results of this study also have clear implications for reliance on discrete sample data for investigation of contaminated sediment, which requires similar but largely untested assumptions of uniformity at the scale of the samples collected and corresponding predictable trends between individual sample points. A detailed overview of this issue is beyond the scope of this paper, however, and detailed field research is again lacking. The push for change elsewhere will be driven in part by a better understanding of the science of Sampling Theory. The need for change will perhaps be

driven even more by responsible parties required to pay for the investigations and the environmental experts who put their credibility at stake each time an investigation is carried out. Additional pushes for change will come from insurers forced to pay for unanticipated cleanup costs, attorneys representing parties involved in property transactions, and the financiers behind these transactions, each of whom could have the most to lose due to faulty data. Not the least, however, is the increased confidence gained in the protection of human health and the environment.

Acknowledgments

The field study carried out by the Hawaii Department of Health was conducted under a grant from the United States Environmental Protection Agency. The authors wish to acknowledge the numerous environmental consultants and regulators who provided input and often lively debate during preparation of this paper. The conclusions expressed in the paper are, however, our own and may not necessarily reflect those of the reviewers.

Funding

Guidance published by the Hawaii Department of Health and referenced in this paper was funded partly through the use of U.S. EPA State Response Program Grant funds. Its contents do not necessarily reflect the policies, actions, or positions of the U.S. Environmental Protection Agency. The Hawaii Department of Health does not speak for or represent the U.S. Environmental Protection Agency.

Conflict of interest

The authors declare no conflict of interest.

References

Association of American Feed Control Officials (AAFCO). 2015. *GOODSamples: Guidance on Obtaining Defensible Samples*. Champaign, IL.

Brewer, R., Peard, J., and Heskett, M. 2016. A critical review of discrete soil sample reliability: Part 1—Field study results. *Soil Sediment Cont*. Available at: http://dx.doi.org/10.1080/15320383.2017.1244171

Environmental Resources Management. 2008. *Sampling and Analysis Plan Amendment Former Pepe'ekeo Sugar Company property, Hakalau, Hawaii*. Prepared for Aloha Green Inc. Hilo, HI,.

Groundswell Technologies, Inc. 2013. *Groundswell Technologies Contouring Web Application and API*. Santa Barbara, CA.

Hadley, P. W. and Petrisor, I. G. 2013. Incremental sampling, challenges and opportunities for environmental forensics. *Environ. Forensics* **14**, 109–120.

Hadley, P. W. and Sedman, R. M. 1992. How hot is that spot? *J. Soil Cont.* **3**, 217–225.

Hawaii Department of Health (HDOH), Office of Hazard Evaluation and Emergency Response. 2011. *Screening for Environmental Concerns at Sites with Contaminated Soil and Groundwater*. Honolulu, HI.

Hawaii Department of Health (HDOH), Hazard Evaluation and Emergency Response. 2012. *Hawaiian Islands Soil Metal Background Evaluation*. Honolulu, HI.

Hawaii Department of Health (HDOH), Hazard Evaluation and Emergency Response. 2015a. *Small-Scale Variability of Discrete Soil Sample Data, Part 1: Field Investigation Of Discrete Sample Variability*. Honolulu, HI.

Hawaii Department of Health (HDOH), Hazard Evaluation and Emergency Response. 2015b. *Small-Scale Variability of Discrete Soil Sample Data, Part 2: Causes and Implications for Use in Environmental Investigations*. Honolulu, HI.

Hawaii Department of Health (HDOH), Office of Hazard Evaluation and Emergency Response. 2016. *Technical Guidance Manual*. Honolulu, HI.

Interstate Technology Regulatory Council (ITRC). 2012. *Incremental Sampling Methodology*. Washington, DC.

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

Minnitt, R. C. A., Rice, P. M. and Spangenberg, C. 2007. Part 1: Understanding the components of the fundamental sampling error: A key to good sampling practice. *J. South. Afr. Inst. Min. Metall.* **107**, 505–511.

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

Pitard, F. F. 2005. *Sampling correctness—A comprehensive guideline*. Proceedings of 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*. Proceedings of the 4th World Conference on Sampling and Blending, The Southern African Institute of Mining and Metallurgy, Cape Town, October 19–23, 2009.

Ramsey, C. A. and Hewitt, A. D. 2005. A methodology for assessing sample representativeness. *Environ. Forensics* **6**, 71–75.

Shulgin, A. I. and Duhaas, L. 2008. *Evaluation of HMA treatment to allow for reuse of waste ash as landfill daily cover*. Prepared for H-Power, Inc. Honolulu, HI, October 30, 2008.

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

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

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

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

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

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

U.S. Environmental Protection Agency (USEPA), Environmental Monitoring Systems Laboratory. 1989b. *Soil Sampling Quality Assurance User's Guide*. Washington, DC (EPA/600/8-69/046).

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

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

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

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

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

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

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

U.S. Environmental Protection (USEPA), Office of Environmental Information. 2002a. *Guidance on Environmental Data Verification and Data Validation*. Washington, DC (EPA/240/R-02/004).

U.S. Environmental Protection (USEPA), Office of Emergency and Remedial Response. 2002b. *Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites*. Washington, DC (OSWER 9285.6-10).

U.S. Environmental Protection Agency (USEPA), Office of Emergency and Remedial Response. 2005. *Guidance on Surface Soil Cleanup at Hazardous Waste Sites* (peer review draft). Washington, DC (EPA 9355.0-91).

U.S. Environmental Protection Agency (USEPA), National Center for Environmental Assessment, Integrated Risk Information System. 2011a. *IRIS Glossary*. Washington, DC.

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

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

U.S. Environmental Protection Agency (USEPA), Superfund. 2014. *Hot Spots: Incremental Sampling Methodology (ISM) FAQs*. Washington, DC.

U.S. Environmental Protection Agency (USEPA), Superfund. 2015. *Screening Levels for Chemical Contaminants*. Washington, DC.

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

A Critical Review of Discrete Soil Sample Reliability: Supporting Information

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

[1] Hawaii Department of Health, 919 Ala Moana Blvd., Room 206, Honolulu, HI 96814, USA; E-mail: roger.brewer@doh.hawaii.gov, randall.peard@doh.hawaii.gov
[2] Element Environmental, 98-030 Hekaha St #9, Aiea, HI 96701, USA; E-mail: mheskett@e2hi.com

Reference:
Roger Brewer, John Peard & Marvin Heskett (2016): A Critical Review of Discrete Soil Sample Data Reliability: Part 2—Implications, Soil and Sediment Contamination: An International Journal, DOI: 10.1080/15320383.2017.1244172. Available from: http://dx.doi.org/10.1080/15320383.2017.1244172

1 Early Concepts of Hot Spots

Early guidance published in the 1980s centered on the need to identify "hot spots" of heavily contaminated soil that could pose a significant risk to human health and the environment. Exactly what constituted a "hot spot" in terms of size and risk was at that time still being debated, and guidance that still serves as the basis of environmental risk assessment was still under development.

As stated in the USEPA document *Methods for Evaluating the Attainment of Cleanup Standards* (USEPA 1989a, notation added):

> 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. (p. 9-1)

Two distinct scales of "spots" are discussed (see also USEPA 1987, 1989b, 1991, 1992a; see also USEPA 2014): 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 sample-sized spots within a spill area or exposure area that could pose hypothetical acute toxicity concerns or hypothetical "not-to-exceed" screening levels (i.e., health effects within 24–96 hours; USEPA 2011a). The first types of "hot spots" are referred to as "Spill Area" Decision Units (DUs) in the HDOH *Technical Guidance Manual* (HDOH 2016). The ITRC guidance on Incremental Sampling Methodology refers to these as "Source Area" DUs, but the intent is identical. The second type of "hot spot" was understandable in concept, but not practical or necessary in practice.

Although the terms are not specifically used, correlation of the concept of "hot spots" to what are now referred to as "Spill Area" (HDOH 2016) or "Source Area" (ITRC 2012) 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. (p. 11)

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%. (p. 74)

Guidance for more up-to-date "DU-MI" investigation methods specifically calls for the identification and designation of suspect Spill Areas or Source Areas for separate characterization (ITRC 2012, HDOH 2016). This is carried out in much the same manner as used to determine locations for discrete samples, except that the area for which the sample is intended to represent is specifically stated upfront.

The identification and characterization of large-scale patterns of contamination such as the area of PCB contamination depicted in Figure 6 of the main paper 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. (p. A-8)

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 under-representation of a significant "hot spot" (see USEPA 1987, 1989a, 1991, 1992a). The guidance documents are, however, describing the need to segregate and independently sample and characterize suspect spill/source areas from anticipated clean areas to the extent practical, based on existing knowledge of the site. Both HDOH and ITRC likewise make this requirement in their respective guidance documents. An error in the early USEPA guidance documents was the assumption that an individual discrete soil sample or, in essence, a "single-increment" 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 suspected, contaminated area.

2 "Hot Spot" Characterization

The next task in the early guidance 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 operating 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 and emphasis 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*. (p. 45)

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. (p. A-8)

The first assumption reflects the more current concept of DU. The second assumption is, however, is at odds with Sampling Theory and is restated in the follow-up USEPA document *Methods for Evaluating the Attainment of Cleanup Standards* (USEPA 1989a):

> When there is little distance between points it is expected that there will be little variability between points. (p. 10-2)

The recommended grid and discrete sampling approaches were incorporated into USEPA guidance documents for the investigation and cleanup of sites contaminated with polychlorinated biphenyls (PCBs; USEPA 1985, 1986, 1990). As discussed in the USEPA document *Verification of PCB Spill Cleanup by Sampling and Analysis* (USEPA 1985, notation 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. (p. 11)

Although erroneous, this assumption 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, notation added):

> The grid spacings are obtained so that the consumer's [i.e., of soil] risk is held to an acceptable level. (p. 131)

The cited authors' use of grid spacing is identical to the use of "Exposure Area" DUs in *Multi Increment* sampling guidance documents to establish a minimum resolution for a soil investigation (ITRC 2012, HDOH 2016). For example, a maximum DU area of 5,000 ft^2 is recommended in HDOH guidance for investigation of a site that is intended to be used for residential or other sensitive land uses (HDOH 2016). Designation of appropriate DU areas is necessarily site-specific. This is recognized in the USEPA *Data Quality Objectives* guidance, restating the above quote (USEPA 1987; see also USEPA 1989a, 1991, 1992a):

> 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. (p. A-8)

In this case, the cited authors are assuming that an individual discrete sample will be adequate to represent contaminant levels within any given "contaminated zone," or DU. This is illustrated in Figure 1 of the main paper, taken from the USEPA *Methods for Evaluating the Attainment of Cleanup Standards* guidance (USEPA 1989a). The various "hot spots" within the 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 to illustrate how an excessively large grid spacing might inadvertently miss exposure-area-sized "hot spots" within the hypothetical site. As discussed in the following discussion of small-scale distributional heterogeneity, tight grids of discrete samples can nonetheless be useful in the identification of large-scale patterns of soil contamination of interest, but are highly prone to false negatives and confusion over artificial, seemingly isolated "hot spots" based on an individual sample or even a small cluster of samples.

3 Composite Samples Versus Composite DUs

"Compositing" of samples collected from the grid area was discouraged due to potential dilution of large-scale "hot spot" areas with outlying 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. (p. C-5)

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 continues to be enforced in large part to this day (USEPA 2005a).

The guidance documents are again describing the need to segregate and independently sample and characterize separate spill or source areas to the extent practical, based on existing information, just as discrete samples intended to represent suspect "hot spots" should not be mixed or "composited" with samples anticipated to represent clean areas. Composting of individual *Multi Increment* samples intended to represent separate spill areas, exposure areas, and other intentionally designated DUs is likewise not recommended (HDOH 2016).

The Interstate Technology and Resource Council (ITRC) ISM guidance document refers to an ISM sample as a "structured composite" (ITRC 2012). This was so named to help field workers visualize collection of a *Multi Increment* sample in the field. In terms of Sampling Theory, and particularly in a regulatory connotation, an MI or ISM sample does not, however, represent a composite sample (Pitard 1993, HDOH 2016). It is simply a sample, as might be collected from any heterogeneous mixture of particulate matter. The term "composite" is only strictly applicable when soil from two or more initially designated DUs is intentionally combined and tested as a single sample. As stated above, compositing of MI (or ISM) samples is typically not recommended and generally defeats the purpose of DU-MI investigation methods (HDOH 2016).

4 Chronic Risk and Large-Scale "Hot Spots"

Early soil investigation guidance warns of potentially significant flaws in the use of simplistic grid and discrete sample investigation methods, should assumptions regarding the "uniformity" of contaminant concentrations within spill areas prove erroneous (Gilbert 1987, notations 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]." (p. 119)

At that time, the absence of three factors—1) a "clear and unambiguous," concept of DUs for risk-based definitions of "hot spots", 2) rigorous field-based studies of small-scale variability of contaminant concentrations in soil, and 3) meaningful evaluation of the reproducibility of discrete sample data—all set the stage for an unavoidable failure in the reliability and efficiency of many environmental investigations that followed.

Several investigators cautioned about the potential problems at the time, but these concerns went generally unheeded as the investigation of thousands of sites was 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. (p. 217)

In contrast to the agriculture and mining industry where crop yields of mineral production are routinely measured, true confirmation of data quality is not tested in the environmental industry under standard, discrete-sample characterization methods, and the precision of estimates of health risk and even *in situ* contaminant mass, including both lab and field error, is largely unknowable.

As further discussed by Hadley and Sedman (1992):

> 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. (p. 219)

To be more precise, approximately 3 metric tons (3,000 kg) of soil would be required to retain 1 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 2015):

> 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 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… 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", the Risk Assessment Guidance for Superfund 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. (p. 3)

The concept of "Decision Units" (DUs) is used in the HDOH and ITRC guidance documents to better define the scale at which an environmental investigation should be carried out (HDOH 2016, ITRC 2012). A Decision Unit is an area, or more specifically, the volume of soil that will be sampled and a decision made on the resulting data. Large-scale "hot spots" of contaminated soil, referred to as "Spill Area (HDOH 2016)" or "Source Area (ITRC 2012)" DUs, are areas of contamination associated with the specific release of a chemical and are 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, sites are normally broken up into "exposure area" DUs for independent testing. This is commonly done as part of due diligence for property transactions. 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 including the entire yard. The size and shape of exposure areas at commercial and industrial properties vary 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., 1,000ft^2 or 5,000ft^2 to a depth of 6 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? (p. 9-2)

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 an individual 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, emphasis added):

> This document assumes that… chemical concentrations *do not exhibit short-term variability* over the sampling period. (p. 2-17)

This caveat also applies to an assumed absence of random small-scale, spatial variability of contaminant concentrations in soil. The potential problem with very small discrete soil samples was 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, emphasis added):

> 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. (p. 5)

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, emphasis added):

> 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… *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. (p. 3)

Addressing random errors in the laboratory analytical process 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. (p. 3)

The implications of these important ideas in the field were, unfortunately, never fully discussed in guidance documents. Even the critical importance of correct laboratory subsampling methods for bulk field soil samples (either discrete or MI), which is the largest potential source

of laboratory error by far, has not yet been incorporated systematically in USEPA SW-846 methods. As a result, many commercial environmental analysis laboratories still do not utilize representative subsampling methods. Ultimately, confusion over the need to determine the "maximum" contaminant concentration within a targeted area and search for sample-sized "hot spots" continued (and still continues) to plague the environmental industry, and reliance on often scant discrete soil sample data for decision making quickly became routine.

5 Misinterpretation of Sample-Size 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. The mass of soil traditionally collected as a discrete sample, typically ten to a few hundred grams, has no basis in science or sampling theory. The mass of a soil sample is instead most often 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 laboratory's interest to request only the smallest mass necessary in order to optimize storage and testing space and minimize disposal costs. Consequently, 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. (p. 2-4)

The authors use the term "sampling unit" in the same sense as a "decision unit," as described above and in HDOH guidance (HDOH 2016). 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 data for discrete samples intentionally collected from a suspect spill/source area 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. (p. 2-4)

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. (p. 2-6)

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 latter point is important and is discussed in more detail in the main text of this paper, 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), but it is only now being applied to the representativeness of the samples actually collected in the field. In addition, the correct laboratory subsampling methods recommended by USEPA in 2003 were not subsequently incorporated systematically in the USEPA SW-846 laboratory methods typically utilized by environmental laboratories, and consequently were not (and are generally still not) widely adopted and practiced. This is in part due to the continued confusion over the need to understand contaminant concentration at the scale of a discrete sample, 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. (p. 3-5)

This could perhaps be considered a second critical juncture in the use of discrete rather than *Multi Increment* sampling methodologies to characterize sites with contaminated soil, with the first being a failure to collect multiple sets of stratified, random replicate samples during initial testing of grid schemes for PCBs in 1986. After presenting a strong review of sampling theory and error associated with 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. This is further reflected in the guidance for the interpretation of data for composite samples (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. (p. 3-6)

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 (now referred to as Regional Screening Levels; USEPA 2015), had to be met by any given discrete sample mass of soil within a site or exposure area. This is incorrect. The models used to develop the screening levels are based on long-term (i.e., 6–30 years) chronic exposure to contaminants in soil within a designated exposure area. In such cases, the screening levels are intended to apply to the concentration of the contaminant for the targeted volume of soil as a single unit—for example, the upper 5 cm of a 100 m^2 exposure area (approximately 5,000 kg of soil). Under ideal circumstances, the entire mass of soil that comprises the Exposure Area Decision Unit would be tested. Since this is not practical, a representative sample must be collected from the DU. There is only one answer to this question. The objective is *not* to identify the "mean" or "average" concentration of the contaminant in soil, but simply the concentration for the unit of soil as a whole.

The 1992 USEPA document repeats a mistake made seven years earlier in guidance for testing of PCB-contaminated soils. It erroneously states that no more than nine (or ten in other documents) samples should be composited as a single sample, in order to ensure that no individual sample might have exceeded a 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. (p. 23)

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 previous method detection limit of approximately 100 µg/kg (see USEPA 1986). As stated (ironically) in the same document (USEPA 1986):

> 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. (p. A-2)

Subsequent guidance, even noting that exceeding a risk-based screening level in an individual discrete sample may not necessarily indicate a risk to human health and the environment, calls for a halt to sampling and a move to remediate the entire site if such a scenario is encountered (USEPA 1989a, notation 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. (p. 9-3)

While it might be easy to suggest, from a regulatory perspective, perhaps such a misunderstanding of risk 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 the main text of the paper, 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 approach above was 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 2005; refer to Subpar O). This is in 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 decision 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. As stated in the USEPA Superfund Environmental Assessment Manual (USEPA 1989c; see also USEPA 1988, 1989b,d):

> In most situations, assuming long-term contact with the maximum concentration is not reasonable. (p. 6-19)

In spite of this observation, subsequent USEPA guidance repeatedly discusses the need to collect discrete soil samples in order to verify the presence of absence of sample-sized “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, notation added; note that this document is a peer-review draft, but, to our knowledge, was not 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. (p. 10)

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. (p. 4)

"Not-to-exceed" levels are intended to address hypothetical, acute toxicity concerns (USEPA 2005):

> 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. (p. 1)

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. (p. 7)

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, emphasis added):

> 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… (p. 9)

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 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. (p. 11)

The manner in which a toxicologist assesses acute health risk is not discussed. The document states, however, that if acute health risks from exposure to very small masses of soil with (hypothetically) 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. (p. 13)

Under this rationale, a decision to initiate a remedial action would be the typical case at any site since it is economically impractical, if not technically impossible, to determine with a reasonable degree of confidence that no individual discrete sample-sized mass of soil among tens or hundreds of thousands (or more) of potential sample-sized masses within an exposure area does not exceed a hypothetical maximum-allowable level. Acute toxicity would in practice need to be tied to 10 g masses or smaller, the default assumed to be ingested by a pica child (USEPA 2011b; default non-pica child soil ingestion rate 200 mg/day). Each 10 g 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 15 cm (6 inches) includes approximately 15,000 kg of soil, or 150,000 hypothetical 10 g "acute toxicity" DUs. The level of effort to prove beyond a reasonable doubt that no individual 10 g mass of soil poses acute toxicity risks for even small areas would be enormous and generally not feasible from either a technical or financial standpoint.

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

> [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. (p. 17)

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 individual "spot" exceeds hypothetical acute or otherwise not-to-exceed soil screening levels.

The cited 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, notations added):

> 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 EU [Exposure Unit] and there are actually areas with higher concentrations, then the resulting [maximum concentration left in place] may not be protective. (p. 20)

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. This will always be the case since the mass of soil designated to assess the "maximum concentration" of a contaminant is never defined, and it is extremely unlikely that sampling will ever adequately represent a true mean concentration for a given area at the scale of a discrete sample. 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 in practice 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 practical implementation 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 an individual 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 be impossible to know without knowledge of every individual marble in the bucket.

Recalculation of a mean contaminant concentration based on removal of the "hot spot" data point using data for the 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. In all likelihood, the "hot spot" removed 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 hot spot at 2 out of 20 discrete sample points implies that such hot spots collectively 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 collection of a new, independent set of discrete samples from separate randomly selected points. Even this would not be fully adequate, though, since the representativeness of any single set of samples is unknown. As discussed in the main text of this paper, 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 for *Multi Increment* soil sample replicates (typically triplicates).

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 was acceptably small for any given, discrete sample-size mass of soil that was not tested as part of the investigation. 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.

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

> 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]. (p. 7-11)

As discussed in the main text of this paper, 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 risk assessment and decision making (USEPA 1989a; notation added):

> 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. (p. 7-11)

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 the main text of this report. As discussed, it is ironic that early USEPA sampling guidance emphasized the need to identify sample-sized 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.

6 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 2014). At the writing of this paper, a number of individual offices within the USEPA are implementing "incremental sampling" approaches into their projects. In spite of progress, the USEPA needs to more broadly recognize the inherent error in discrete

sampling methods to characterize soil, and by default sediment, for final decision-making purposes. Clear recommendations for the use of more science-based based and defensible DU-MI investigation approaches to characterize soil and sediment should be adopted nationwide.

References

Gilbert, R.O. 1987. *Statistical methods for environmental pollution monitoring*. New York: Van Nostrand Reinhold Company, Inc.

Hadley, P.W. and Sedman, R.M. 1992. How Hot Is That Spot? *J. of Soil Cont.* 3, p. 217-225.

Hadley, P.W., Crapps, E. and Hewitt, A.D. 2011. Time for a Change of Scene. *Environ. Forensics* 12. p. 312-318.

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

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

Hadley, P.S. and Bruce, M.L. 2014. On Representativeness. *Environ. Forensics* 15(1), 1-3.

Hawaii Department of Health (HDOH), Office of Hazard Evaluation and Emergency Response. 2016. *Technical Guidance Manual*. Honolulu, HI.

Interstate Technology Regulatory Council (ITRC). 2012. *Incremental Sampling Methodology*. Washington, DC, February 2012.

Minnitt, R.C.A., Rice, P.M. and Spangenberg, C. 2007. Part 1: Understanding the components of the fundamental sampling error: a key to good sampling practice. *J. South. Afr. Inst. Min. Metall.* **107**. p. 505-511.

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

Pitard, F.F. 2005. Sampling Correctness - A Comprehensive Guideline, Proceedings of 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*. Proceedings of the Fourth World Conference on Sampling and Blending, Southern African Institute of Mining and Metallurgy, Cape Town, October 19–23, 2009.

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

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

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

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

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

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

U.S. Environmental Protection Agency (USEPA), Environmental Monitoring Systems Laboratory. 1989b. *Soil Sampling Quality Assurance User's Guide*. Washington, DC, March 1989. (EPA/600/8-69/046).

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

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

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

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

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

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

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

U.S. Environmental Protection Agency (USEPA). 2005. *Polychlorinated Biphenyls (PCBs) Manufacturing, Processing, Distribution in Commerce and Use Prohibitions. 40 CFR Part 261*. Washington, DC.

U.S. Environmental Protection Agency (USEPA), National Center for Environmental Assessment, Integrated Risk Information System. 2011a. *IRIS Glossary*. Washington, DC, August 2011.

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

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

U.S. Environmental Protection Agency (USEPA), Superfund. 2015. *Screening Levels for Chemical Contaminants*. Washington, DC, November 2015.