**Hot Spots: Incremental Sampling Methodology (ISM) FAQs**

Prepared by Deana Crumbling, USEPA, 27Mar2014; updated 8Sep2014 (minor typo-type edits) ([crumbling.deana@epa.gov](mailto:crumbling.deana@epa.gov))

**Disclaimer**

This educational FAQ paper discusses the range of scientific evidence that bears on the question of “hot spots” in the presentations about incremental sampling, including the ITRC webinars on ISM (incremental Sampling Methodology). The reasoning expressed in this paper is the author’s interpretation intended to respond to commonly asked technical questions related to ITRC incremental sampling seminars. The responses are not intended to set, alter, or revise existing policy or guidance, but are intended to stimulate discussion within the cleanup community. As with all technical discussion and interpretation, any actions or site decisions should be reviewed and evaluated according to the appropriate regulatory, statutory and policy guidance related to the specific decisions made.

1. **Overview**

When [incremental sampling](#Glossary) is discussed in [webinars](#ITRC_webinar_reference) given through the Interstate Technology and Regulatory Council (ITRC) and in classroom settings, one question is often asked:

**“Won’t hot spots be missed?”**

**The shortest, most direct answer is “yes and no, depending on what you mean.”**

Unfortunately, that answer by itself is not very enlightening. However, no simpler answer will do because the correct answer depends on what the questioner has in mind. The term “hot spot” and the phrase “find a hot spot” mean different things to different people under different sampling scenarios. So giving a meaningful answer first requires knowing the questioner’s thinking. When exploring that thinking, however, a host of ambiguities and engrained misconceptions about soil sampling emerge which need clarification before the hot spot question can be addressed.

This paper tries to address common confusions about soil sampling and hot spots as concisely, yet completely, as possible. Misconceptions about soil sampling and soil data interpretation are nearly universal in environmental cleanup. This paper presents scientific evidence that dispels those misconceptions and puts hot spot discussions on a more productive, evidence-based footing. If you devise soil sampling designs, collect soil samples or analyze them, or review soil data, it is well worth your time to work through this paper. Having an accurate understanding of soil sampling complexities **ensures more effective sampling designs and avoids misinterpretation of soil data results, reducing harmful and costly errors in decisions** about contaminant exposure and site cleanup.

This paper was written to answer Frequently Asked Questions related to hot spots, to be a handy teaching tool, and to stimulate further discussion among practitioners. The paper represents a compilation of common questions received in ISM courses and Internet seminars. Therefore it uses informal language and a conversational style that will hopefully engage the reader. It uses internal hyperlinks and definition pop-ups (bright blue/underlined) that allow the reader to “jump around” in the paper based on the reader’s interests. Therefore, this paper can be read in three ways:

1. A straight-through reading will set the stage with an overview of the many issues raised by hot spots.
   1. Following the introductory overview of the various issues (Sections 1 through 3), detailed discussions follow with the scientific evidence to resolve those issues.
   2. An annotated glossary for important terms appears at the end of the paper. It carries a great deal more detail and explanation than usually found in glossaries.
2. Alternatively, the hyperlinks allow the reader to follow a single issue or discussion thread from its introduction through its detailed discussion before tackling the next issue.
   1. Remember where you were when you use a hyperlink so it is easy to navigate back to where you left off.
   2. Sometimes what appear to be separate questions or topics are actually just two faces of the same coin, and they can be addressed by the same scientific evidence. So different questions may be hyperlinked to the same answers.
   3. If you hover your cursor over linked words or phrases, a brief definition will appear in a pop-up. Usually there is more explanation provided in the Glossary. This is indicated by a “+” sign that appears at the end of the text in the pop-up. If you control-click the link in the pop-up, you will be taken to the Glossary’s opening page. You will have to navigate from there to find the term in the Glossary; then navigate back to where you were in the main body of the paper to resume reading.
3. You can also jump directly to [summary conclusions (Section 4.6)](#Section_4_6) and [Section 4.5.11.1](#Section_4_5_11_1), which condense the information presented in this white paper down to brief statements.

We begin the overview with examining the ambiguities of “hot spot” terminology.

* 1. **Terminology:** **What do you have in mind when you say “hot spot”?**

1. Perhaps you have in mind a significant spill or an anticipated source area with concentrations at least 2-, 5-, 10, or X-times the [decision threshold](file:///C:\Users\DPOWEL02\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.Outlook\O18HYDLU\screening#Glossary) over an area/volume of at least 1-, 5-, 10, or X-[sq.ft](#Glossary)./[cu.ft.](#Glossary)? (See [Appendix K](#Appendix_K) for a discussion of the “hot spot” term.)

OR

1. Perhaps you are thinking that a hot spot is indicated at *all* locations where *any* discrete grab soil sample had a result greater than the decision threshold, even if it is over by just a little (e.g., a lead result of 405 ppm when the decision threshold is 400 ppm) and the volume of soil that actually exceeds the threshold is unknown? See the discussions in [Section 4.5](#Section_4_5).
   1. **Terminology**: **What do you have in mind when you say you want to “find hot spots”?**
2. Are you thinking you want to determine the actual location(s) and soil volumes of any amount of soil having a concentration higher than the decision threshold? Good luck with that! See the answer in [Section 4.5.12](#Section_4_5_12) and discussions throughout [Section 4](#Section_4_0).
3. Do you worry that high concentration increments will be “diluted out” by low concentration increments? See [Section 4.5.7](#dilution).

OR

1. Do you just want to make sure significant areas with high concentration are not missed, and that they are [representatively](file:///C:\Users\DPOWEL02\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.Outlook\O18HYDLU\hot#Glossary) incorporated into an estimation of the true concentration for a specified area or volume? See [Section 4.5.11.2](#Section_4_5_11_2) and [Appendix C](#Appendix_C).
2. Do you realize that [incremental sampling is much, much better than discrete samples](#Section_3_3) for making sure hot spots get incorporated into the statistics? For one real project, calculations show that incremental sampling would have been 33 times more effective at picking up hot spots than the equivalent discrete design. See [Section 3.1](#Section_3_1).

The answers to the above questions, in conjunction with the type of decision you want to make, drive the design of a science-based sampling program that deals effectively with hot spots.

**2.0 Decision goals:** **What is the primary decision for which you are collecting soil data?**

Common project decisions are related to: exposure assessment, remedial design, or demonstrating compliance with an action level after cleanup activities were performed. Different decision scenarios will change the dimensions and methods for determining or accounting for “hot spots.”

**2.1** If you are calculating **exposure concentrations for risk assessment**, first you need to define the exposure unit (EU). When constructing the sampling design to determine the exposure point concentration (EPC), the EU becomes the sampling design’s [decision unit](#Glossary) (DU).

* + 1. EPA risk assessment guidance is very clear that the EPC is calculated as the *arithmetic* *average* (i.e., mean) concentration across an EU.

“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.”(Section 6.4.1, page 6-19, USEPA, 1989)

“As discussed in Section 6.4.1, the concentration term in the exposure equation is the average concentration contacted at the exposure point or points over the exposure period. When estimating exposure concentrations, the objective is to provide a conservative estimate of this average concentration (e.g., the 95 percent upper confidence limit on the arithmetic mean chemical concentration).” (Section 6.5, page 6-24, USEPA, 1989)

In general, if your decision goal is the DU average, you are not concerned with locating hot spots at the time you are determining the average. Your goal is only to make sure “significant hot spots” are included in the determination of the average. In this respect, incremental sampling is greatly superior to discrete sampling, because the default number of increments for a ¼- to ½-acre DU is 30. Thirty increments provide a much higher [sampling density](#Glossary) than the usual discrete sampling design, which increases the probability that significant hot spots will not be missed, but will be incorporated into the incremental sample at the same proportion that they occupy in the field. For more discussion of what makes a hot spot “significant” and how to adjust increment density to ensure their incorporation, see [Section 5](#Section_5).

An exception is when a hot spot could occur in (or cover) a “sensitive area.” A prime example is the existence of a dedicated children’s play area within a yard DU. If the increased activity in the play area can result in greater soil exposure, this area can be split off into its own separate DU. In other words, the design becomes a smaller play-area-DU within the large yard-DU. Each DU is sampled separately with the default of 30 increments. At the ½-acre and smaller areas, the default of 30 increments is less about spatial coverage than about having enough increments to cope with the usual level of short-scale heterogeneity of soils. See [Section 4.5](#Section_4_5) and [Section 5.2.4.2](#Section_5_2_4_2).

However, if you find that the DU average is higher than acceptable, knowing whether contamination is limited to only part of the DU, and knowing which part, may be important to an efficient cleanup that avoids removal and disposal of clean soil. You have a couple of choices.

* If you collected the incremental samples in such a way that the entire DU (all increments) went directly into a single DU-incremental sample, you have no spatial information and a re-sampling of the DU is required to determine whether only a portion of the DU is “dirty.”
* If you anticipated the potential need for cleanup when designing the incremental sampling strategy, you can use a tiered incremental-composite sampling approach that preserves spatial information.
  + That information is only accessed if and when it is discovered that cleanup is needed. If it is needed, that information can be accessed from archived [sampling unit (SU)](#Glossary) samples without remobilization to collect new samples.
  + More information about the this variation on incremental sampling can be found on pages 32 to 34 in the 2011 [User Guide - Uniform Federal Policy Quality Assurance Project Plan Template For Soils Assessment of Dioxin Sites (PDF)](http://www.epa.gov/superfund/health/contaminants/dioxin/pdfs/Dioxin%20UFP%20QAPP%20UserGuide.pdf), which is on the Superfund [Dioxin Toolbox](http://www.epa.gov/superfund/health/contaminants/dioxin/dioxinsoil.html)  website.

* + 1. As instructed by EPA risk assessment guidance, the statistical “upper confidence limit (UCL) on the mean” is used to make sure that the EU’s average is not underestimated. A UCL can be thought of as the sample mean (which is calculated from the sample set) plus a statistical “safety factor” to make sure the estimate of the population mean (or “true mean”) will not be underestimated. If this idea is unfamiliar to you, see the additional discussion in [Appendix F](#Appendix_F). A UCL is calculated by taking into account several considerations that assess how reliable the calculated mean might be. Factors include
* How many physical samples were used to calculate the mean?
  + Estimation of a population mean is more confident when calculated from more samples.
* How variable are the results within the sample set?
  + More variable data lead to a less confident estimate of the population mean.
* What is the statistical distribution of the population that the data came from? If that is not known, what distribution does the data in the sample set have?
  + If data come from a well-defined distribution like a normal, or near-normal distribution, estimates of the mean can be more confident than if the data come from a non-normal (e.g., lognormal or gamma distribution) or [nonparametric distribution](#Glossary).
* How much statistical confidence do you want to have that your safety factor was big enough so that the true mean is not underestimated?
  + Having higher statistical confidence (such as 95%) causes the safety factor to be bigger than having a lower statistical confidence (like 80%).
  + If the relationship between higher statistical confidence and wider uncertainty intervals is new or confusing to you, [Appendix E](#Appendix_E) might help.
* [Appendix F](#Appendix_F) offers more discussion of the UCL using an example to show how changing each of the factors listed above changes the UCL value.
  + 1. There are many ways a UCL can be calculated. The kind of data distribution determines which UCL is the most appropriate one to use.
* A Student’s t-UCL is used for normal or nearly normal distributions which have a more or less bell-shaped curve. This curve can be plotted using a mathematical equation, and the UCL is based on that mathematical model.
* A gamma UCL is used for gamma distributions, which have curves that are skewed (pulled out) to the right, similar to lognormal distributions. Gamma distributions can also be described by mathematical equations.
* A nonparametric distribution is one that does not fit a mathematical model, i.e., cannot be drawn by a mathematical equation. There are a variety of techniques to determine UCLs for nonparametric distributions. A common, simple nonparametric UCL is the Chebyshev UCL. It tends to make higher UCLs than many other UCL calculation methods.

**2.2** If you are sampling to **confirm that contamination has been removed**, you need to know the dimensions of the soil unit over which an action level is based (i.e., the DU). In other words, what is the area, depth and target particle size of the DU over which the decision threshold number applies? Often these values were developed based on exposure, so the DU dimensions may be the same as an EU.

**2.2.1** The answer to “What is the area and depth over which the decision threshold number applies?” *cannot be* “everywhere.” Consider: In a DU having a 30-sq.yd. area and 1-inch depth, there are 1.4 million potential 1-gram analytical samples (such as for metal analysis). It is not possible to confirm that every single 1-gram sample is less than the action level.

**2.2.2** The only viable, scientifically-based option is to take enough physical samples to show that the *true concentration* for the DU (which is the *average* of all 1.4 million samples making up the DU) is *“probably”* less than the action level. (See [Section 4.4.1](#Section_4_4_1) and [Appendix D](#Appendix_D) for discussions of the true concentration/true mean.)

How convinced are you by a mean being “probably” less than the action level? Fortunately, “probably” can be quantified statistically; it is called a “confidence level.” Like in risk assessment (Section 2.1), a UCL on the mean quantifies the word “probably.”

Think of the UCL as a “padded” estimate of the mean. The “padding” or “safety factor” helps ensure that the true population mean is not being under-estimated just because the mean of the sample you happened to pull from the population is lower than the population mean. A 95% confidence limit is traditionally used, although this might be lowered under certain circumstances (through negotiation, analytes with low toxicity, etc.). (See additional discussion in Section 2.1.2 and [Appendix F](#Appendix_F).)

**2.2.3** **A statistical method to assess against a “do-not-exceed” threshold.**

There is a statistical strategy that can be used to assess compliance with a “not-to-exceed” concentration criterion. It is presented in a 2002 USEPA RCRA sampling guidance document (USEPA 2002a) and is discussed in [Appendix G](#Appendix_G) of this paper. However, it cannot actually provide 100% confidence that all potential samples in a DU are less than the decision criteria. The best it can do is 99% confidence that 99% of the potential samples in a population (such as within a DU) are less than the decision criterion. ***But you have to be willing to collect 459 discrete samples from the DU and hope that they ALL are lower than the criterion!*** If none of 459 exceed the “do-not-exceed” criterion…Congratulations! You are a winner! You can claim 99% confidence that 99% of the potential samples in the DU comply with the criterion.

What if 1 of the 459 exceeds? You get one more chance for the 99% brass ring. You can collect another 203 samples and cross your fingers. If none of the additional 203 samples exceed (so you have a total of only 1 exceedance out of 662 samples), you can still claim 99% confidence that 99% of potential samples are less than the criterion.

What if 2 of the 459 exceeds? You have failed to prove that you met a goal of 99% confidence that 99% of the potential samples are less than the decision criterion. So sorry, you lose. Thanks for playing.

If you (or the regulator) are willing to settle for a lower confidence level, such as trying to show at 95% confidence that 95% of potential samples are less than the criterion, you can do a similar spin of the wheel, but start off with only 59 discrete samples, rather than the 459 needed for 99% confidence. [Appendix G](#Appendix_G) contains the rest of the details.

I bet incremental sampling is starting to look pretty good about now!

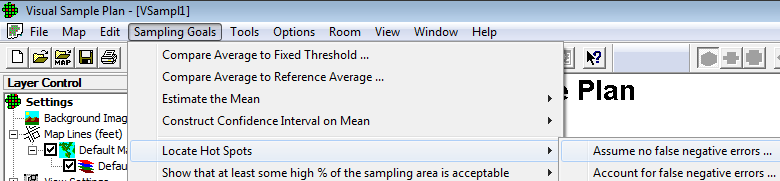
**2.3** If you are **delineating contamination for a remedial design**, you need to define a “remedial unit” (RU), which is the dimensions or volume of soil removed as a unit. The RU is often based on what is operationally convenient to remediate at one time while avoiding inclusion of significant amounts of possibly clean soil which would unnecessarily increase disposal costs. The true concentration of the RU is compared with the decision threshold to determine whether the RU needs removal.

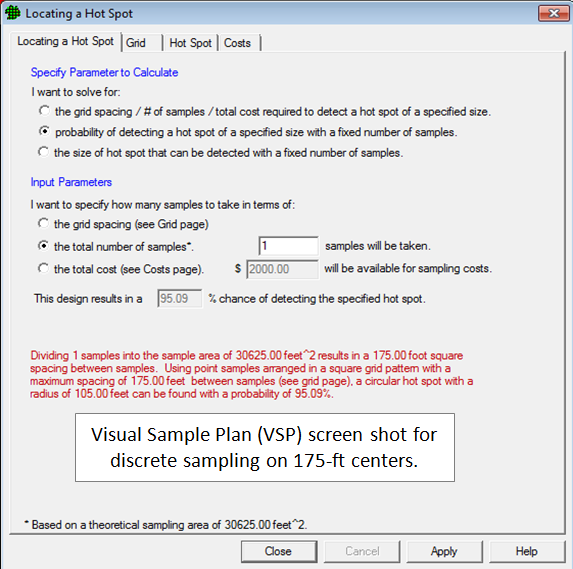
* Depending on the size and nature of the site, an “operationally convenient” RU might be the volume of a 55-gallon drum, the “bite” of a backhoe, a dump truck load, or roll-off bin, etc.
* Each RU is tested to determine whether its concentration exceeds the threshold.
* You want to balance the increased sampling costs of having more, smaller, “surgical” RUs against the cost of unnecessary remediation of clean soil with fewer larger RUs. Look for the “break-even” point.
* To construct a sampling design, the RU becomes the DU for sampling purposes.
* Using the CSM, historical data, visual cues, real-time field analyses or laboratory samples, try to draw DU boundaries to include contaminated areas and exclude clean areas.

**3.0 Are discrete sampling designs “finding hot spots” for you now?**

Consider this real-life example: A project manager wanted to use incremental sampling to sample a large area, but the risk assessor would not agree because they were concerned that “hot spots might be missed.” So a discrete sampling design was developed based on a 175 ft x 175 ft grid cell size, which the risk assessor found acceptable.

**3.1** The discrete sampling design called for a discrete sample to be taken at the center of each 175-ft x 175-ft grid cell. The area of a single grid cell is 30,625 sq.ft. Using the Visual Sample Plan (VSP) Hot Spot module, the size of a hot spot that could be detected by 1 sample every 175 ft can be determined.

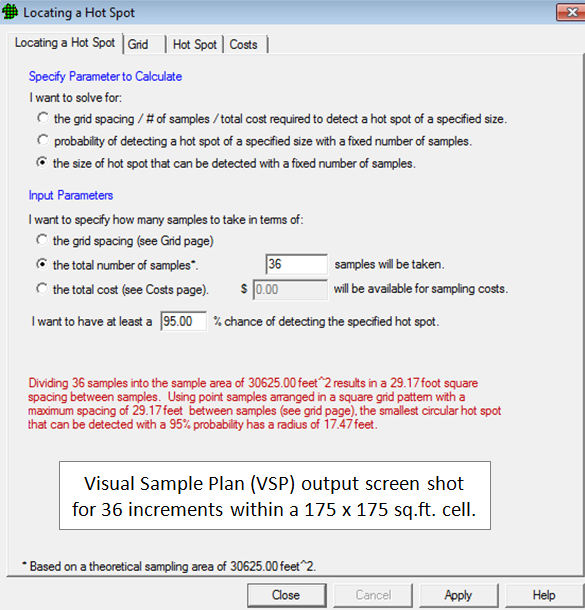
* It turns out that a circular hot spot needs to have a radius of 105 ft (i.e., 34,640 sq.ft. or just over ¾ of an acre) in order to be detected 95% of the time.
* Not surprisingly, the hot spot has to be a bit larger than the area of a grid cell for 95% confidence that at least 1 sample of the grid network will fall within the hot spot area.
* These calculations assume that there are no “false negative” results from samples taken from within a hot spot. As we will see later in this paper, these types of “false negative” sampling errors (a sample collected within a true hot spot gives a low concentration result) are very common as a consequence of short-scale heterogeneity. (See [Section 4.5.11](#Section_4_5_11) and Section [4.5.11.2](#Section_4_5_11_2))
* The VSP Hot Spot module below was used (the one that assumes no false negative errors).
* See the VSP hot spot module’s output, which is in red text, in the screen shot below.
* See more information about using VSP’s Hot Spot modules in [Appendix C](#Appendix_C).
* Continue to the graphic on the next page.



Visual Sample Plan (VSP) screen shot for discrete sampling on 175-ft centers.

**3.2** However, if incremental sampling had been used, it could have been designed to produce the same number of field samples and analyses as the discrete design (1 per grid cell, not counting QC samples).

* But instead of a single grab sample in the center of a grid cell, 36 increments could have covered each 175 ft x 175 ft (30,625 sq.ft.) cell.
* With 36 evenly spaced increments, circular hot spots as little as 17.5 ft in radius (i.e., 960 sq.ft or just 1/45th of an acre) would have been “hit” 95% of the time. (See the VSP screen shot below)
* This assumes no “false negatives” so a direct comparison can be made to the discrete design.
* Although the location of any hot spots would not be revealed, their concentrations are likely to be proportionally incorporated into the incremental sample. The concentration of the incremental sample is much more likely to reflect a grid cell’s true mean concentration (which includes the contribution from hot spots within it) than a single discrete sample ever could.



**3.3** In other words, for this example, the chance of **missing a hot spot with discrete sampling was *32 times higher* (by area) than the chance of missing it with incremental sampling**. Which would you rather have? To have 32 times the probability that a high concentration area will be incorporated into your estimate of the mean (even though you may not know exactly where it is) vs. missing any evidence of a hot spot altogether and falsely assuming there was no hot spot?

**4.0 The Nature of Soil: Misconceptions that affect interpretation of soil data**

What are our expectations for soil samples? As noted by someone with a long history in the field: “In some cases, the interest was in just establishing that soil was, indeed, contaminated. In some cases, soil samples were collected to determine the ‘level’ of contamination, particularly the ‘highest level.’ In many cases, people collected soil samples and sent them to the laboratory hoping that the test results would inform and guide them as to what to do next.” (Hadley, et al 2011)

Practitioners and regulators in the cleanup field may believe that the following two statements are true.

* “Clean” soil will not have any discrete sample results higher than some threshold.
* Even one discrete sample result exceeding some decision threshold indicates there is sufficient contamination to pose a health risk, or the soil in question is “out of compliance.”

So there is sometimes a strong tendency to look at a list of soil results, pick out the ones (or maybe the only one) that exceed(s) some numerical criterion and conclude, “There is some area(s) here where contamination is present over the action level.” There is an equally strong tendency to look at a list of results where none exceed numerical criteria and say, “There is no contamination present.”

To resolve these issues, we need to have a detailed discussion about

* the nature of soil,
* how contaminants behave in soil, and
* how sample collection, analysis and interpretation are complicated by the way contaminants behave in soil.

**Key Points Discussed in Section 4**

* “Particle effects” make soil heterogeneous, which causes each analytical sample to truly have a different concentration.
* Therefore soil data are often highly variable at small spatial scales, so that 2 samples can have widely different results, even if those 2 samples are taken from
  + within the same sample jar, or
  + within a small area/volume of *in situ* soil (samples separated by inches).
* Variable data create opportunities for decision errors about risk and remediation.
* Soil heterogeneity must be proactively managed for soil data to be reliable.
* If you want to preview the summarized conclusions of Section 4, you can jump directly to [Section 4.5.11.1](#Section_4_5_11_1) and [Section 4.6](#Section_4_6).

Contamination concentrations in soils vary widely at three spatial scales relevant to site cleanup.

* **Large-scale heterogeneity** refers to the kind of heterogeneity we are trying to find; that is, differences in concentration between a “contaminated” volume of soil and an “uncontaminated” volume of soil. We want to distinguish the two so the contaminated soil can be treated or removed. These decisions are generally made in spatial scales of linear yards and acres.
* **Short-scale heterogeneity** refers to differences in concentration between samples spaced apart by inches to a couple feet. This is the scale at which [collocated samples](#Glossary) are collected. (See Section 4.
* **Micro-scale heterogeneity** refers to differences in concentration that occur within the same sample jar, so that the analytical results from replicate analytical samples (such as [“laboratory duplicates”](#Glossary)) are significantly different (i.e., different enough to change the decision). This is the spatial scale of subsampling and analysis.

We will tackle micro-scale heterogeneity first.

**4.1 The Difficulties of Analyzing Heterogeneous Matrices**

Take a look at all the different colors of jelly beans in this jar. Now imagine someone took the jar and hid it behind a curtain. Although you saw the diversity of colors before, you can’t see them now. You reach behind the curtain, pluck a jelly bean off the top, pull your hand out and look at a red jelly bean.

Would you tell yourself that ALL the jelly beans in the jar behind the curtain are red just because the one you are looking at this moment is?

Probably not.

Now, pretend you could look at a jar of soil and “see” the arsenic concentration of each soil particle like seeing a color. Red particles have really high concentrations, orange ones are moderately high, through yellow and white on down through green and blue to purple particles which have the lowest arsenic concentration.

Even though you know the particles carry a wide range of concentrations, you pick a blue one off the top and conclude that all the particles in the jar have the same low concentration as that one.

Does this seem like a good idea? Does this sound like defensible science? But this is what is done by practitioners and environmental laboratories all the time when soil heterogeneity is not explicitly taken into consideration. Subsequent discussion will show that this analogy is not as far-fetched as it might sound!

**4.1.1** **The Nature of Soil Creates Conditions that Increase Data Variability**

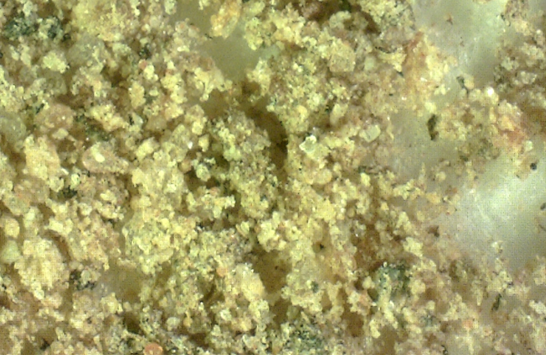
Heterogeneity is the condition of being non-uniform, and there are different kinds of non-uniformity. The jelly bean “particles” were non-uniform in color, although they were pretty uniform (homogeneous) in size, shape and composition (sugar, binder, and food coloring). Jelly beans are exceedingly simple compared to soil particles.

One kind of heterogeneity is called “[compositional heterogeneity](#Glossary),” which is also called “constitutional heterogeneity.” Real soil (which excludes playground sand) is strikingly heterogeneous in composition because it contains a wide range of different minerals. Soil also contains organic carbon, which itself comes in different compositions. As organic material ages in the soil, it undergoes decomposition by insects, fungi and bacteria. Eventually it reaches the “humic” stage where it is rather amorphous and poorly defined, yet seemingly resistant to further breakdown. Humic material (related, not surprisingly, to “humus”) and other forms of organic carbon act like a sponge and are highly absorbent of a wide range of chemical contaminants.

Mineral particles vary greatly in their propensity to bind contaminants. Some, such as quartz or feldspar particles common in sandy soils (see Figure 4.1.1 below), do not bind contaminants very

well. The relatively large particle size of sand particles means that they do not have as much surface area per gram to which contaminant molecules can adhere as do smaller particles. In addition, the electrical and chemical inertness of silicate minerals means that contaminants adhere weakly and wash off easily.

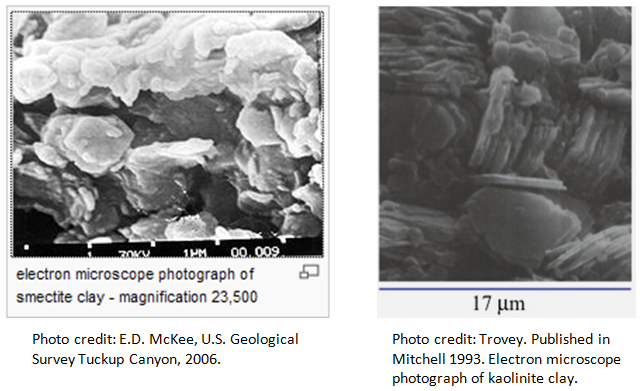
However, some minerals sorb contaminants strongly. [Note: contaminants can stick to the outside surface of particles (called a*d*sorb or a*d*sorption), and may also get inside some particles (called a*b*sorb or a*b*sorption). The words “sorb” and “sorption” cover both mechanisms.] Some mineral particles can sorb very high loadings of contaminants so that the mineral particle is completely covered by contaminant elements or molecules as we will see in [Section 4.1.3](#Section_4_1_3).



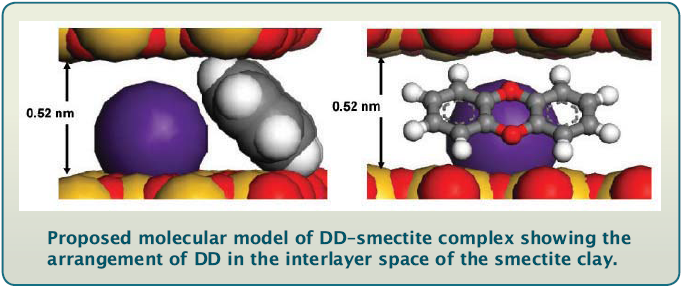
**Figure 4.1.1** A sandy soil may appear fairly uniform to the naked eye, but viewed under a microscope the particles are non-uniform in size, shape, color and composition.

**4.1.2** **Clays sorb contaminants extremely well**

* Particles of clay minerals are hundreds of times smaller than sand particles, so clays have thousands of times more surface area per gram to which molecules can adhere.
* Clay minerals are typically composed of flat plates (see photomicrographs below) that provide additional surface area to sorb contaminants.
* Because of their molecular composition, the flat clay plates are “coated” with electric charges, most of which are negative, although some positive charges line the plates’ edges.
* The negative charges of clay particles attract and hold positively charged ions, such as metal atoms.



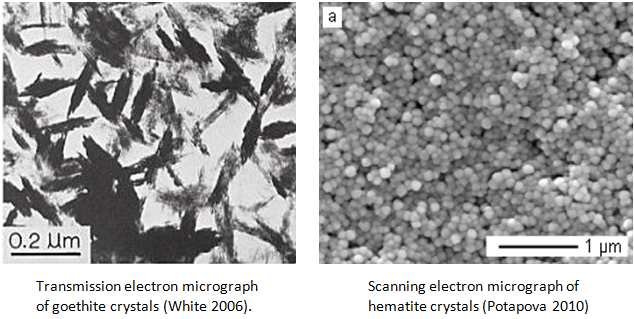
* Clays also attract organic molecules via van der Waals’ forces that encourage organic molecules to work their way between the plates, where they can become trapped.
  + “Van der Waals forces” include attractions and repulsions between atoms, molecules, and surfaces caused by fluctuations in electron orbits such that a slight, temporary electric charge is created on the atom, molecule, or surface.
  + This magnitude of this charge is much less than the charge on an ion and is much less stable. However, because these charges are constantly appearing and disappearing on all atoms and molecules, the interactions can add up to be significant.
* The graphic directly below illustrates how a dioxin molecule can become stuck between two plates of a clay particle.



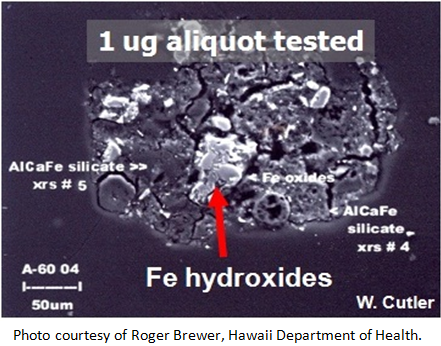
Source: Boyd 2010; DD is dibenzo-*p*-dioxin; smectite is a type of clay mineral. The purple sphere represents a positively charged atom, usually common soil nutrients such as K+, Ca+2 and Mg+2. In soils contaminated by anthropogenic metals, positively charged metal ions, such as Pb+2, also bind to clay particles.

**4.1.3**  **Iron oxides also sorb high levels of contaminants**

The term “iron oxides” is a catch-all phrase for 16 soil minerals that contain iron, oxygen, and often hydrogen in different proportions and molecular arrangements. Depending on soil conditions such as pH, iron oxides take different crystalline forms and possess different electrical charges. Two types of iron oxide particles, goethite and hematite, appear in the two photomicrographs below.



Iron oxide crystals can be smaller than even clay particles. The tiny size/high surface area of iron oxide particles along with their chemical properties create great affinities for both metal and organic contaminants: **“**Given the average concentration in soil, the iron in a cubic yard of soil is capable of adsorbing from 0.5 to 5 pounds of soluble metals as cations, anionic complexes, or a similar amount of organic[s].” (Vance 1994).



The photomicrograph to the right directly visualizes iron (Fe) oxide particles that are coated with arsenic. They are part of a clump of particles, most of which appear dark gray.

The arsenic shows up as bright white in this view under the spectroscopy used to identify elements. The iron oxides were identified in a different spectroscopy scan. There is one large arsenic “nugget” at the red arrow, and several medium and small ones scattered about.

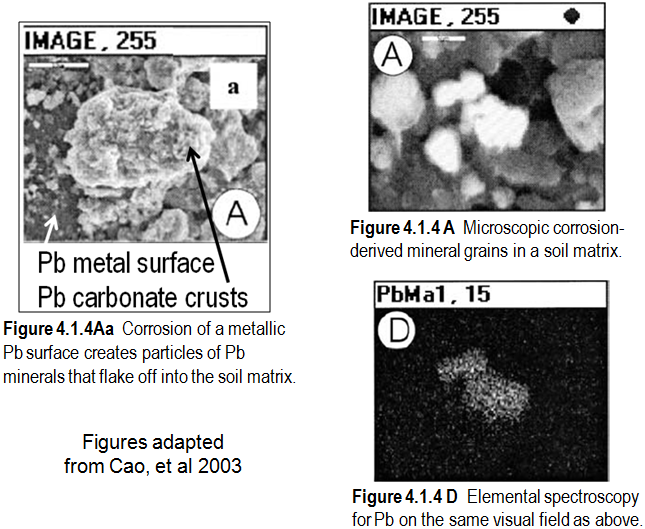
The larger, dark particles are made of a silicate mineral that shows up as dark because arsenic has not bound to its surface.

This photomicrograph dramatically demonstrates how contaminant binding to sorptive mineral grains, such as iron oxides and clays, produces a “nugget effect” for contaminants. Even if contaminants were originally released to the soil in some liquid form (such as pesticide applications), they do not remain in the liquid phase. They selectively attach to soil particles, so they act as particles in soil sample collection and handling procedures. **Contaminant molecules cannot be thought of as “dissolved” throughout soil.** They are NOT homogeneously distributed like contaminants dissolved in water. Instead, contaminants behave as highly concentrated discrete particles because they are bound to discrete soil particles.

Since they behave as particulates, contaminant heterogeneity is very high at the spatial scale at which samples are collected and analyzed. There is a serious mismatch between the scale at which we make decisions and cleanup sites (thousands to millions of tons of soil) and the scale of data generation (1 to 200 or so grams). This situation occurs because the scale of data generation is not representative of the scale of data interpretation and decision-making.

**4.1.4** In addition, contaminants are often released directly in a particulate form. Lead and arsenic are released at firing ranges in the form of dust and fragments created during gun firing and from the impact of bullets with soil. Metals can also be deposited in the form of ash particles from smelters, power plants and mining processes.

Corrosion of bullets in the soil releases lead and arsenic in nugget form, as photographed in Figure 4.1.4Aa. Figures 4.1.4 A and D show corrosion particles having very high Pb content in a “sea” of untainted soil particles carrying very few, if any, Pb atoms.



**4.1.5** Studies that examine the relationship between contaminant concentration and soil particle size consistently find that concentrations for the smaller particle sizes are higher, and sometimes much, much higher, than concentrations for larger particle size fractions.

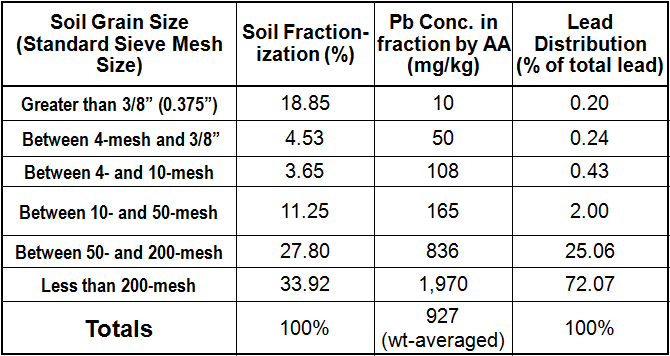
Table 4.1.5 and Figure 4.1.5 below show data from a fractionization study done on lead-contaminated soil from a military firing range (ITRC 2003). Three-quarters of the lead was associated with the smallest particle size fraction measured, which was the less than 200-mesh fraction, corresponding to particle diameters less than 74 µm (0.074 mm).

The general pattern that finer soil particles carry much higher contaminant loads than larger soil particles applies not only for anthropogenic metals contamination in soil, but also for persistent organic compounds such as pesticides, polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs) and dioxins.

A study of organic contaminants in house dust reported: “Fourteen pesticides and all 10 of the target PAHs were detected in one or more of the seven size-fractionated samples…The concentrations of nearly all of the target analytes increased gradually with decreasing particle size for the larger particle [fractions], then increased dramatically for the two smallest particle size [fractions] (4-25 µm and < 4 µm).” (Lewis et al 1999)

**Table 4.1.5**

Lead (Pb) Concentration Increases with Decreasing Particle Size (Data from ITRC 2003)

****

****

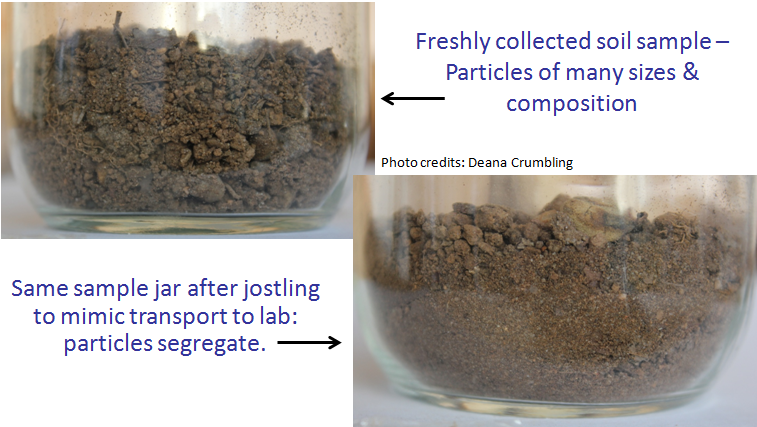
**Figure 4.1.5** This is a bar graph of the data displayed in Table 4.1.5.

**4.2** **Soil Particles Will Segregate at Every Opportunity**

The particulate behavior of contaminants in soil has several implications when analyzing soil samples. Most importantly, particles are subject to “[distributional heterogeneity](#Glossary) (DH).” In contrast

to compositional heterogeneity, DH is not a property of the particles themselves, but is a function of *where* the particles are located with respect to other particles that have a similar composition. DH is a primary cause of [sampling variability](#Glossary) both in the field and in the laboratory. Sampling variability is also called “sampling error,” where the term “error” is used in the statistical sense to mean “variability” or “imprecision.”

A big reason for the high variability often observed between lab duplicates is that the smallest particles (carrying the highest contaminant loadings) settle to the bottom of sample containers as movement of the contents causes larger particles to bump apart, creating gaps through which small particles readily fall.

****

**4.2.1** Anyone who has unloaded soil samples from a cooler after shipment has observed that free-flowing soil samples “[segregate](#Glossary)” during transport. That is, soil particles will separate into layers according to particle size. Small particles work their way toward the bottom, leaving large particles on top, and medium particles in-between. The longer the sample has been moved about, the more distinct the layers become. The only way segregation does not happen during sample transport and handling is if the sample is a wet mud or is completely consolidated (i.e., solid), such as a clod of dried mud or clay.

Any activity that moves soil particles around, including stirring, facilitates segregation. Gravity is the culprit, and it is, as they say, everywhere! **A common misconception** **is that soil samples can be “homogenized by stirring until thoroughly mixed.”**

* “Homogenization” by stirring is impossible, not only because it abets segregation, but also because stirred particles retain all of the original compositional properties which help create concentration variability in the first place. Would stirring the jelly bean jar “homogenize” the jelly beans into a single color?
* Stirring soil makes micro-scale heterogeneity (i.e., distributional heterogeneity within a sample container) worse, not better. Even trying to scoop the fine particles off the bottom to “mix them in” does not work because it exacerbates “[grouping error](#Glossary).” [See panel A in the Glossary definition for “grouping and segregation error,” and see USEPA 2002a and USEPA 2003 for additional discussion of Gy-related sampling errors.]

**4.3** **Common Subsampling Practices Cause Data to be Biased Both High and Low**

All free-flowing soil samples arriving at a laboratory will be segregated to a greater or lesser degree. When segregated soils are sampled, subsampling bias is created. In turn, subsampling bias causes analytical results to be biased as compared to the true concentration for the jar’s contents (see Section 4.4.1 on the next page).

The photo below shows the coarse particles that predominate at the top of a segregated soil sample.



It is extremely common for laboratories to take “grab” subsamples by scooping particles from the top of the jar contents. What happens when subsampling is done this way? Recall that coarser particles carry little, if any, contamination. If coarse particles are over-represented in the analytical sample, the reported concentration will almost certainly be biased low compared to the true concentration that would be obtained from a [representative](#Glossary) mix of particle sizes.

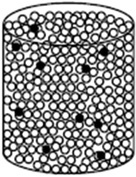
Other common subsampling activities can bias results up or down.

* If digging in the sample with a spoon brought up a subsample from a fine grained layer to the exclusion of medium-sized and larger particles, the analytical result would be biased high.
* Using a spatula with a curved bowl vs. a flat spatula will produce subsamples with different particle size ratios. The bowl shape will retain larger, rounder particles that would roll off a flat spatula.
* When subsamples are being weighed out on a balance and the subsample is getting close to its target weight, it is common to selectively tap finer particles from the spatula into the subsample. This allows the subsample’s mass to be gradually brought to weight without going over by much. But selectively adding smaller particles during weighing creates bias.

Although they may sometimes work in opposite directions, you cannot expect the factors causing subsampling bias to cancel each other out. Rather, each subsample is the product of many uncontrolled variables that collude to create the range of different concentrations observed with [replicate](#Glossary) soil subsamples, such as [laboratory duplicates](#Glossary). A number of common or even “recommended” sample splitting and subsampling techniques do not work because their movement of the soil creates ample opportunity for small particles to settle and then be unevenly distributed among the split portions. This includes cone-and-quartering, fractional shoveling and others (see [Appendix A](#Appendix_A) for a list of techniques no longer recommended because the likelihood of segregation and grouping errors is high). Gy-correct subsampling and splitting techniques are discussed in the [ITRC on-line ISM document, Section 6](http://www.itrcweb.org/ISM-1/6_LABORATORY_SAMPLE_PROCESSING_AND_ANALYSIS.html).

**4.4** **Smaller Lab Subsamples Have Poorer Precision**

Because of the particulate nature of soil, the analytical sample mass is an important factor determining data bias and variability (even if the soil is not segregated!). Soil contamination at “trace concentrations” [i.e., concentrations in the [parts per million (ppm)](#Glossary) and less], can be diagrammed as a relatively few high concentration discrete particles that are “sprinkled” unevenly throughout a bulk matrix of uncontaminated particles. Figure 4.4, which is borrowed from EPA guidance (USEPA 2002a), illustrates this. For simplicity, assume that the black particles all carry the same high contaminant loading, while the white particles carry no contaminant. We will use this “cartoon” of soil particles to understand why small analytical subsamples are more likely than larger ones to misrepresent the true concentration of the sample.



**Figure 4.4** Depiction of the particulate nature of trace amounts of contaminants in soil as a small number of high-concentration “dirty” particles (dark color) present in a “sea” of uncontaminated “clean” particles (light color) (from USEPA 2002a).

**4.4.1** **Determining the true concentration.** First of all, if we want to talk about a “biased” concentration result, we must have some way to determine the true (unbiased) concentration, even if only by a thought experiment. So let’s think about how we could determine the true concentration of the soil in the container of Figure 4.4. To set the stage for our thought experiments,

* Consider all of the soil in the container to be a single [population](#Glossary)
* A concentration result is calculated as the mass of analyte in the soil divided by the mass of the soil.

So, the absolute true concentration in the container can be obtained (theoretically) in two ways:

**Thought Experiment #1**

Weigh all the soil in the container. Let’s say the entire soil population is 1 kg.

* Then extract the analyte from the whole 1 kg at once (as one giant analytical sample).
* Measure the mass of analyte extracted. For our thought experiment, we will assume the extraction and analysis steps are error-free. The mass of analyte is 100 mg.
* Perform the calculation to get the true concentration of the container.
* Note that there is only 1 result because the population was analyzed in a single giant analysis. Therefore, there is no sampling uncertainty or variability because no samples were taken from the population.
* Because the entire population was analyzed, a [UCL](#Glossary) **would not** be calculated.
  + A UCL is used to make an upper-bound estimate of the true concentration based on adding a “[safety factor](#Section_2_1_2)” to the mean of a [sample set](#Glossary). The safety factor takes various uncertainties into account when the characteristics of a small sample set are used to predict the characteristics of the whole population.
  + However, in this thought experiment, we did not take any samples from the population.
* The true concentration is related to the true ratio between the “dirty” particles and the “clean” particles within the container housing the population.

**Thought Experiment #2**

Another way to determine the true concentration is to take the [population](#Glossary) and divide it entirely into individual analytical samples, and then analyze all of them.

* Let’s use 5 grams as the mass of the analytical sample in our thought experiment.
* Since the population is 1 kg (1000 grams), …

… the population is composed of 200 analytical samples.

* Now analyze all 200 samples. As before, assume all procedures are “error-free.”
  + Note that “error-free” in this context does not mean “error-free” in a statistical context. In statistics, the term “error” is often used to mean “variation.” We cannot expect all samples to have the same concentration, so there will be variability (measured as “standard deviation, SD) in the population’s data set.
    - Remember those dark (“dirty”) particles in Figure 4.4? They are free to move around, so it is unrealistic (even in our thought experiment) to expect that each analytical sample will contain the same number of dirty particles.
    - Each of the 200 5-gram samples could have a unique ratio of dirty to clean particles.
  + The concentration for a sample is proportional to the ratio of dirty to clean particles for that sample. A higher dirty to clean ratio produces a higher concentration result.
* Let’s say that the 200 analytical results range between 50 and 150 mg/kg, and the calculated mean for the 200 results is 100 mg/kg.
* Since the entire population is measured as 5-gram samples, the calculated mean for the 200 samples must be the same as the population mean obtained in Thought Experiment #1,.
* Let’s say the variability for the 200 soil results (measured as SD) is 20 mg/kg. This is the SD *for the population* when the analytical samples have a mass of 5 grams.
  + Note that the population SD will not be 20 mg/kg if an analytical sample mass other than 5 grams were used.
  + The population SD for 1000 1-g soil samples will be higher than 20 mg/kg. For a population of 100 10-g samples, the population SD will be lower than 20 mg/kg. (Refer to [Section 4.4.3](#Section_4_4_3) for experimental evidence of this statement.)
* Even though there is variation among the members of the population, there is no sampling uncertainty (i.e., less confidence that population characteristics are accurately predicted because of the amount of variation between sample results) because **all** samples in the population were analyzed.
  + Since all members of the population were analyzed, the mean and SD of the statistical sample IS the mean and SD of the population.
* Again, a UCL is not calculated because we do not need a conservative estimate of the true mean. We know with complete certainty that the true population mean is 100 mg/kg.

**4.4.2** In the real world, soil populations in the field are unimaginably large. There is no way to analyze all of the millions or billions of potential analytical samples within a decision unit (DU). Therefore, we can never absolutely know the true concentration within a DU. We take only a few (3 to 10 or so?) of those billions of potential samples and hope that the mean and SD we calculate from the statistical sample is fairly close to the population’s mean and SD for the [sample support](#Glossary) we are using. [Section 4.4.3 will explain why a soil population SD is higher (i.e., the population is more variable) when smaller sample supports are used.] In other words, when we take a statistical sample, we hope that the sample set will “represent” the characteristics of the population. We hope the mean of our sample set will be close to the population’s mean, the 95th percentile for the sample will be close to the 95th percentile for the population, the SD for the statistical sample will be close to the population SD, etc.

**4.4.2.1** The complication with soil, however, is that the analytical sample mass can and does vary from project to project; and changing the analytical sample mass changes the number of potential samples and the characteristics of the target population. The nature of the decision also changes the target population: quantifying risk contributed by the hand-to-mouth exposure pathway will target a different population than a decision involving the concentration of soil expected to be disposed in a landfill.

* The target soil population for an exposure pathway mediated by soil sticking to skin will be those soil particles having diameters less than 250 micrometers.
* On the other hand, the target soil population for landfill disposal will be “bulk soil,” which is generally defined as soil material with particle diameters less than 2 mm (2000 micrometers) (USDA 2009).

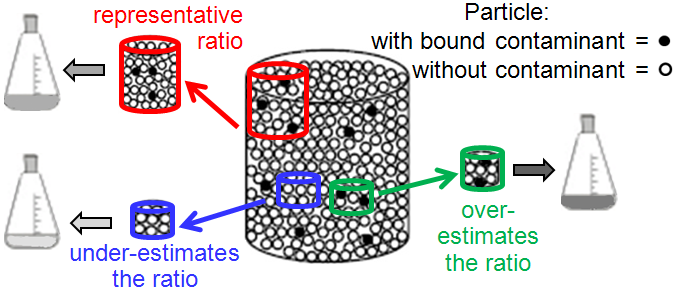
Soil sampling and analysis is not a natural fit for traditional/classical statistics, because **there is no natural unit for the physical soil sample on which measurements are taken.** For most applications of statistics, there are obvious, natural, discrete units that comprise a physical sample.

* If we are counting the frequency of heads in coin tosses, the [sample unit](#Glossary) on which “heads” or “tails” is determined is 1 coin toss. There is no such thing as ½ or ¼ or 2.5 coin tosses.
* If we are doing a medical study looking at the effect of a new drug on blood pressure, the sample unit is the 1 person on whom blood pressure is measured. There is no such thing as a blood pressure measurement for ½ or 3.5 of a person.
* But there is no natural sample unit for soil. The amount of soil collected in the field is whatever is convenient as long as there is enough to run all the analyses.
  + One PAH analysis might take 5 grams in Laboratory #1 and 10 grams in Laboratory #2.
  + If a sample jar containing 50 grams of soil was submitted for PAH analysis to these two laboratories, Lab #1 would analyze 1/10th of the soil sample, whereas Lab #2 would analyze 1/5th of the sample.
  + But it is just as likely that 100 grams, or 150 grams, or 1000 grams of soil might be collected from the DU’s population for submission to a lab.

**4.4.2.2** Since there is no natural soil sample unit, analytical samples can be as small as the analysis will permit. There are strong incentives to collect and analyze small soil samples.

* Small sample jars are cheaper than large ones.
* Since large samples weigh more than small ones, shipping is cheaper for small samples.
* Extracting or digesting large analytical samples requires more of the expensive, ultra-pure solvents and acids used in environmental analysis.
  + This requires larger glassware that is more expensive to buy, wash and store.
  + Sample preparation takes up more space on the lab bench and in the hood (the vented, enclosed space where dangerous chemicals and samples are handled), which means fewer samples can be run at the same time.
  + More liquid and solid waste is generated that may trigger additional RCRA regulations, as well as be more expensive to collect, store and dispose of.
  + The desire for “greener” practices is an incentive for generating less waste.
* Partly in response to the above, and partly because advancing technology lets them do it, laboratory equipment and instrument manufacturers are continually adapting their equipment to use smaller and smaller amounts of sample.

**4.4.2.3** Unfortunately, as analytical samples from heterogeneous matrices get smaller, their results get more variable. The reason goes back, yet again, to that ol’ bugaboo…the particulate nature of soil. Recall the container of dirty and clean particles in Figure 4.4? That container appears again in the next figure. Figure 4.4.2.3 shows what happens when the population is sampled with large and small [sample supports](#Glossary).

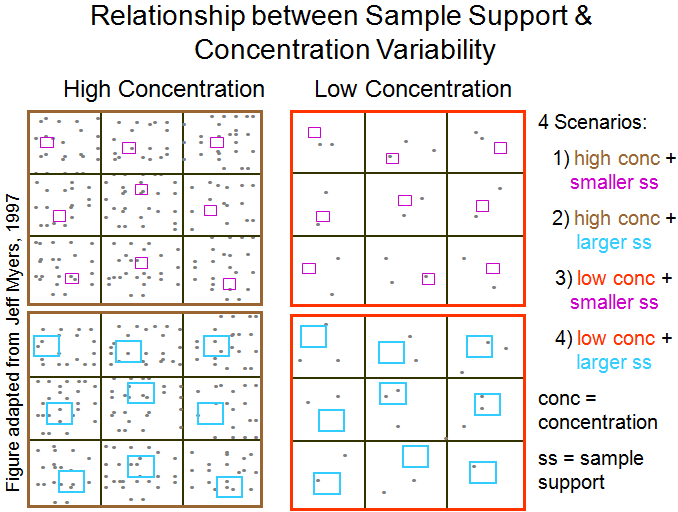


**Figure 4.4.2.3.** Illustration of possible sampling outcomes for a container filled with a certain ratio of dirty to clean particles. Smaller analytical samples (**blue** and **green**) are more likely to misrepresent the true ratio than are larger analytical samples (**red**). When the sample is extracted prior to analysis (represented by the flasks), the concentration of the extract (higher concentrations are represented by darker flask fluid) is proportional to the ratio captured by the sample. (Figure adapted from USEPA 2002a)

Larger analytical samples, illustrated by the **large, red sample at the upper-left of the figure**, are more likely to capture the true ratio of contaminated () to uncontaminated () particles. So the **red sample** has a dirty-to-clean particle ratio (and thus concentration) that is closer to (i.e., more representative of) the ratio and concentration of the population. Smaller analytical samples are more likely to misrepresent the true ratio. Small samples may capture soil holding **fewer dirty particles (see the blue, bottom-left sample**), giving very low or nondetect results after the sample is digested or extracted for analysis (light colored flask). Sometimes a small sample can pick up a **higher ratio of dirty particles (the green sample to the right of the container**). The **smallness of the green sample’s mass** increases the reported concentration even more. Here’s why:

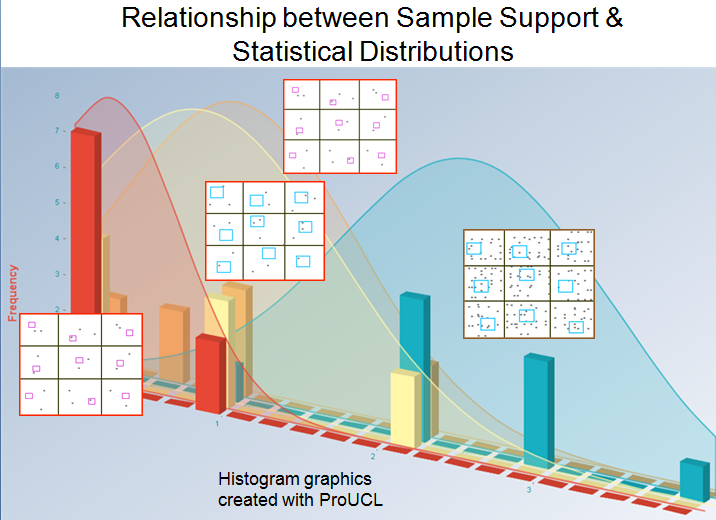
* Imagine if 50 micrograms (a microgram is 10-6 gram) of analyte is extracted from 10 grams of soil.
  + The equation directly below finds that the contaminant concentration [when expressed as milligrams (mg) analyte per kilogram (kg) soil] is 5 mg/kg.
* Then imagine the same 50 micrograms was extracted from one-tenth the original mass (i.e., from 1 gram of soil). Since the sample mass is 1/10th of the former, the latter’s concentration is ten times higher (50 mg/kg).

**4.4.2.4** We saw in Figure 4.4.2.3 that small [sample supports](#Glossary) are less likely to be representative of the true concentration (i.e., are more likely to be biased low or high) than large sample supports. It turns out that the concentration of the matrix (the ratio of dirty to clean particles) is another factor influencing data variability. The interaction between concentration (high vs. low) and sample support (large vs. small) is shown in Figure 4.4.2.4a. In this figure, clean particles are not shown individually. They make up the white background.



**Figure 4.4.2.4a.** Four combinations of relative concentration and sample support are displayed. Variability is measured as the different numbers of dots captured in boxes within a single scenario. A small box (small sample support) combined with a trace concentration (i.e. ppm or less) produces more nondetect results than any other combination. Recall that higher concentrations are produced when smaller sample supports manage to capture some dirty particles.

The corresponding statistical distribution of concentrations for each scenario is shown in Figure 4.4.2.4b below.

****

**Figure 4.4.2.4b.** The combination of small sample supports and trace concentrations (ppm and less) produces many nondetects with occasional high concentrations. This creates a skewed or “lognormal” distribution (**orange curve**) with most results clustered on the left (at very low concentrations) and a right tail (occasional high concentrations. A more bell-shaped distribution (**aqua curve)** is produced when sample supports are large enough to capture a consistent number of particles.

Very small analytical samples are the norm for trace metals analyses. The mass of soil typically digested for metals analysis ranges from 0.5 to 2 grams across laboratories, with most labs using 1 gram. Routine semivolatile organic analyses usually use at least 5 grams, and 10 grams is common. Dioxin analyses usually use an analytical sample mass between 15 to 30 grams.

Let’s use what we’ve learned so far to examine the results of a U.S. Department of Energy study that looked at the relationship between analytical sample support and data variability, and see what this means for using single discrete sample results to find hot spots.

**4.4.3** Back in the 1970s, the U.S. Department of Energy conducted a study that pulls together all the threads discussed in this section. The experiment evaluated the relationship between analytical sample mass and data precision and accuracy (Doctor and Gilbert 1978).

* A very large portion of soil (several kilograms) contaminated with americium-241 (Am-241, a radioactive metal) was obtained and processed by drying, disaggregation, and sieving to less than 10-mesh (i.e., particles larger than 2 millimeters in diameter were removed). This large batch of soil became the population (analogous to the population in Figures 4.4 and 4.4.2.3) that was examined in the study.
* Unlike chemical analysis, radiological analysis is able to analyze a large volume of soil at one time.
  + The experimenters determined that the true activity (a measure of concentration for radioactive elements) of americium-241 for this population was 1.92 nCi/g.
* The experiment then involved taking 20 replicates each of different sample supports.
  + The study used a total of 5 different sample supports (in grams): 1, 10, 25, 50, and 100. However, for the sake of brevity we will only spend time on only three of those: the 1-g, 10-g, and 100-g sample supports.
* All samples were analyzed for radioactivity. Table 4.4.3 provides a summary of the data and Figure 4.4.3.1 displays the statistical distributions produced from sample masses of **1 gram** **(green),** **10 grams (purple)** and **100 grams (red)**.

**Table 4.4.3**.

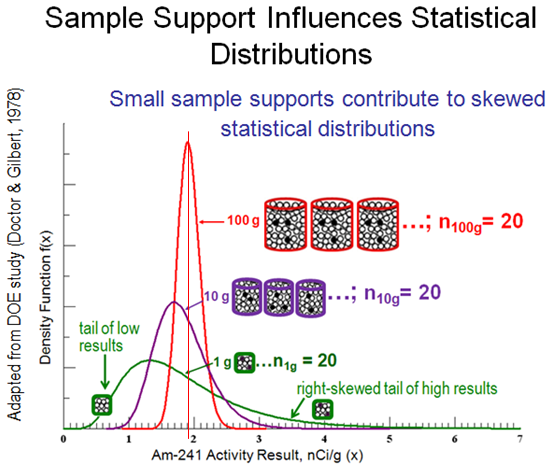
Summary of Am-241 study results for a soil population with true activity equal to 1.92 nCi/g.

|  |  |  |  |
| --- | --- | --- | --- |
| **20 Subsamples of Mass:** | **1-gram** | **10-gram** | **100-gram** |
| Minimum Result (nCi/g): | **1.01** | **1.36** | **1.70** |
| Maximum Result (nCi/g): | **8.00** | **3.43** | **2.30** |
| Range (nCi/g): | **7.0** | **2.1** | **0.6** |
| Distance from mean to minimum: | **0.92** | **0.46** | **0.22** |
| Distance from mean to maximum: | **6.07** | **1.61** | **0.38** |
| Mean (nCi/g): | **1.93** | **1.82** | **1.92** |
| Standard Deviation (nCi/g): | **1.52** | **0.48** | **0.18** |
| CV (equivalent to RSD): | **0.79** | **0.26** | **0.09** |
| Statistical Data Distribution: | **nonparametric** | **gamma** | **normal** |
| ProUCL recommended UCL type: | **modified-t** | **approx. gamma** | **Student’s-t** |
| 95% UCL (nCi/g): | **2.58** | **2.00** | **1.99** |

If you compare the table columns inside the **heavy black border** above, you will observe that

* The means for all 3 sample supports were close to the population mean of 1.92 nCi/g.
* The variability values [measured as standard deviation (SD) and coefficient of variation (CV)] for the 3 data sets were different.
  + The highest variability was seen in the smallest sample support **(1 gram)**, and the lowest variability in the largest support **(100 grams)**.
  + The range between the highest (maximum) and lowest (minimum) result in each data set also decreased from **1-gram** to **100-gram** supports.

**4.4.3.1** If the data sets are plotted as statistical distributions and fitted to mathematical formulas for their curves, each sample support takes a different shape, as shown below in Figure 4.4.3.1.



**Figure 4.4.3.1.**  Statistical data distributions for the **1-gram sample support (green)**, **10-gram support (purple)**, and **100-gram support (red)**. (Figure adapted from Doctor and Gilbert, 1978)

* The **100-g samples, which are plotted in red,** form a tall, slim peak, indicating the 100-g data was very precise; also borne out by the narrow range and low SD in Table 4.4.3.
  + The peak maximum (thin vertical red line) corresponds to the mode (the most common value) of the data set it. It also falls on the mean for that data set, which is the same as the true mean of the population (1.92 nCi/g).
  + The width of the peak runs from about 1.4 to 2.5 nCi/g. It is unlikely for any single result to fall outside of those bounds.
  + The curve is pretty symmetrical, although not perfectly so. There is the slightest right-skew to the curve.
  + Integration shows that 51% of the **100-g** sample results fall below the true mean, and 49% above it. Thus there is almost a perfectly even chance that any single result will fall below or above the true mean. This is what you’d expect with a symmetrical curve centered on the true mean.
    - The fact that the peak is narrow means that low results will not fall far away from the mean (no lower than 1.4, a potential under-estimation of 27%), and high results will fall no farther than 2.5, a potential over-estimation of 30%).
    - Table 4.4.3.1 below will summarize the over- and under-estimation potential for each of the 3 sample supports.
* The **10-g samples, plotted in purple,** form a peak that is wider than the **100-g** peak. This indicates the **10-g** sample supports have less precision (i.e., higher variability), as was observed in Table 4.4.3.
  + Integration shows that 64% of the samples results fall below the true mean. This means there is a somewhat greater chance that any single sample result will under-estimate the mean.
    - The width of the 10-g peak runs down to about 0.9 nCi/g. Not only is it more likely that a single sample will under-estimate the mean, the under-estimation could be as much as 53%.
  + There is some right skew to the curve (i.e., the right side of the curve is pulled out such that the overall curve is not symmetrical).
    - Skew is caused by some samples in the data set having high concentrations that are higher by more than the lower concentrations are low.
    - This is reflected in the data catalogued in Table 4.4.3 as the distances from the mean to the max result (1.61), and from the mean to the min result (0.46).
  + The width of the 10-g peak in the graph runs up to 3.2 nCi/g.
    - Although there is only a 36% chance of over-estimating the mean with a single sample result, when the mean is over-estimated, it could be by as much as 67%.
* Now look at the **green-colored curve representing the 1-g samples**. The green curve is very wide, indicating high data variability (i.e., poor replicate precision).
  + Integration shows that 62% of the samples results fall below the true mean. This means there is a somewhat greater chance that any single sample result will under-estimate the mean, rather than over-estimate it.
  + The curve is very wide, ranging from 0.45 to 6 nCi/g, and the curve is pulled out far to the right into a very distinct right-skewed tail.
    - This indicates that **individual 1-g sample results are somewhat more likely to underestimate the mean, and by as much as 77%.**
    - Although there is only a 38% chance of over-estimating the mean, the right-skewed tail means that when 1-g samples do **over-estimate the mean, it could be by as much as a whopping 213% !**
    - If we consider the range of the 100-g sample results (1.4 to 2.5) to be “near the true mean of 1.92,” then 1-g sample results have only a 44% chance of falling near the true mean. More than half of the individual 1-g samples will have results that are not near the true mean.
    - There is a 20% chance that a single 1-g sample result will be greater than 2.5. Translated to English, this means that 1 out of 5 samples would falsely indicate that the population exceeds the threshold.
    - The over- and under-estimation potential for each of the 3 sample supports is summarized in the table below.

**Table 4.4.3.1.**

Summary of the chance for error and the potential degree of that error when using a single result to estimate the true mean for the Am-241 soil population.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Support** | **Chance of Estimation Error** | | **Amount of Potential Error** | |
| Over-estimation | Under-estimation | Over-estimation | Under-estimation |
| **1-gram** | **38%** | **62%** | **77%** | **213%** |
| **10-gram** | **36%** | **64%** | **67%** | **53%** |
| **100-gram** | **49%** | **51%** | **30%** | **27%** |

**4.4.3.2** **QUIZ!**

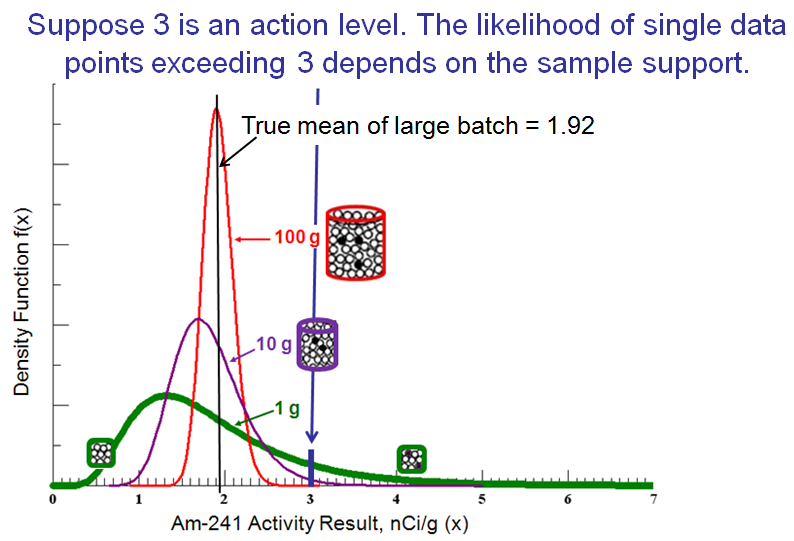
Since you now understand the pitfalls of sampling particulate materials, can you explainwhy wesee such **extreme high and low results in 1-gram samples?** Hint:It goes back to Figures 4.4.2.3, 4.4.2.4a and b. The answers can be found in [Appendix H](#Appendix_H).

* **Now you know why most soil data take “lognormal” distributions.** Skewed distributions are a signal that the sample support is too small to handle the micro-scale particulate heterogeneity of the matrix.
* As observed with the **100-g** sample, as the sample support becomes large enough to control particle effects, the data distributions approach normal.

**4.4.3.3 Just in case you’ve been wondering what all this has to do with hot spots, we are now ready to address that.**

Look at what happens if we add a decision threshold into Figure 4.4.3.1. The **blue spike** on the x-axis at 3 nCi/g in Figure 4.4.3.3 below indicates a theoretical decision threshold.

If a decision about exceedance (i.e., whether or not a hot spot is present) is made on the result of a single sample, the chance that an exceedance will be found **when the true concentration is less than the threshold** is strongly influenced by what sample support is used.

****

**Figure 4.4.3.3.**  Relationship between sample support, data distribution, and the likelihood that some data results will exceed a decision threshold. (Figure is adapted from Doctor and Gilbert, 1978.)

* Examine the data distribution for the **100-g sample support (red curve).** The high precision in the data means that there is practically a 0% chance that a single 100-g sample will exceed the 3 nCi/g threshold (mathematical integration calculates that chance at 0.02%).
* For the **10-g sample support (purple curve)**, its smaller sample support gives less precise data and there is a 1% chance that any single sample will exceed 3 nCi/g.
* Data imprecision and the likelihood of decision error increases to 11% for **1-g samples (heavy green line)**. The chance of falsely identifying a hot spot is 1 in 9, even though the mean of all 20 1-gram results was 1.93 nCi/g, almost exactly the true mean for the population.
  + This is an important point: **although data variability makes it dangerous to make decisions based on a single sample result, if there are enough samples, the highs and lows can balance out and the mean of the data set could accurately reflect the true mean.**
  + **This finding is particularly relevant for using real-time, field portable X-ray fluorescence (XRF) instrumentation for metals analysis. The very small sample support that the XRF can “see” means that data variability can be very high.**
  + Combining incremental sampling and XRF in the field can produce a highly effective sampling design, but sample processing might not be as rigorous as is possible in the laboratory. However all is not lost!
  + Since the XRF uses no consumables per analysis, the only cost for taking many replicate XRF readings is time (1 to 2 minutes or less per reading). Sufficient replicate readings can be used to statistically control for sampling variability when physical control over sample heterogeneity is not as rigorous.
  + The general procedure is outlined in [Appendix I](#Appendix_I).

**4.4.3.4 The number of replicate results required for error balancing so that a representative estimate of the true mean can be attained depends on the variability in the sample set.** The more variability in the data set (i.e., the higher the standard deviation), the more replicates required to get their average reliably close to the true mean. (See discussion in [Appendix J](#Appendix_J).)

The number of results required to achieve this balance can be determined statistically. You need to know how close to the true mean you want to be and how often you want to achieve that closeness. For example, only **4 100-g samples** would be needed for their average to be within 10% (0.2 nCi/g) of the true mean (between 1.72 and 2.12) 95% of the time. But the high sampling error of the **1-g samples** require that **240 sample results** be averaged together to get to the same level of data quality. This leads to a very important conclusion: ***for a heterogeneous matrix, a decision based on an average of many results is much more likely to be correct than a decision based on a single result.***

The take-home message from the Am-241 sample support experiment is clear:

**A decision about the presence of a “hot spot” or the absence of contamination is likely to be made in error if it is based solely on a single isolated sample.**

**Confidence in decisions made on a single discrete soil sample is undermined by within-sample variability, which gets worse as lab subsamples get smaller.**

**4.5 Short-Scale Heterogeneity Causes a Patchwork of Field Sample Concentrations**

Micro-scale, within-sample heterogeneity isn’t the only thing that can lead us astray if decision-making is based on a single soil sample. You say you want to use single sample results to find hot spot locations and to delineate the amount of soil that has a concentration greater than the decision threshold? Ok, but first, let’s look at an experiment that can shed some light on how we might think about accomplishing that task. (If you don’t want to work through the scientific evidence, you can jump straight to [Section 4.6](#Section_4_6) for answers).

**A Quick Look at What Is Covered in Section 4.5**

4.5.2: Explain the experimental design of a field study of soil contaminated with lead- and chromium-based paints.

4.5.3: Review summary statistics for 645 individual samples taken from very small areas, and what we can learn from this.

4.5.7: Answer the FAQ “Won’t hot spot samples be diluted out if they are part of a composite?”

4.5.8: Even background samples are variable when sample supports are small!

4.5.11: If the old way to find hot spots doesn’t work, what should it be done instead?

**4.5.1** This results of this experiment reinforce the fact that [sample support](#Glossary) is a critical factor in determining the magnitude of analytical results. The first half of Section 4 looked at the effect of sample support on analytical samples, which are usually subsamples taken from a jar of soil.

* Because the soil is in a sample jar, differences in matrix particle composition and location are called *within-sample* *heterogeneity*.
* *Within-sample variability* is the term that describes the range of data values observed when multiple subsamples are analyzed from the same heterogeneous sample.

We saw that particles in a jar can move around, carrying their contaminant loads with them while they segregate into layers within the jar (or in a “mixing” pan or any other container), worsening the natural heterogeneity of soil and causing high within-sample data variability.

Now we will look at *in situ* field heterogeneity, called *between-sample* heterogeneity. This heterogeneity causes differences in concentration as you move from spot to spot on the “undisturbed” ground surface or within the subsurface. Sample support plays an important role here also in determining concentration results. This experiment collected small analytical samples directly from the ground surface, so between-sample heterogeneity can be measured directly. There was no subsampling to that could cause within-sample heterogeneity to cloud the issue. However, because the mass of soil taken from the ground surface for analysis was so small (~0.5 g), and the spatial scale for sample collection is so small, micro-scale heterogeneity is still present.

The data and information for this experiment were obtained from the following sources:

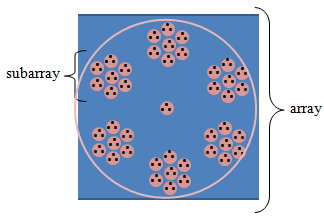
* Joanna Becker’s 2005 Perdue Univ. Ph.D. thesis entitled *Centimeter scale analysis of soil heterogeneities within a long-term, heavy metal contaminated site*, and
* Becker, Joanna, M.T. Parkin, C.H. Nakatsu, J.D. Wilbur and A. Konopka (2006) Bacterial Activity, Community Structure, and Centimeter-Scale Spatial Heterogeneity in Contaminated Soil. *Microbial Ecology Vol. 51*, 220-231.

For her thesis, Becker wanted to determine the micro environment (including toxic metals) that could influence the health of soil bacteria. So she measured lead (Pb) and chromium (Cr) in soil at the centimeter scale. The contaminated soil was “out back” of a former DOT storage warehouse. For years discarded road paint which had been dumped in this area. It was estimated that the dumping began around the 1950s, so paint had decades to age and crumble into dust. White road paint contained Pb and yellow paint contained Cr. This discussion will only cover the Pb results.

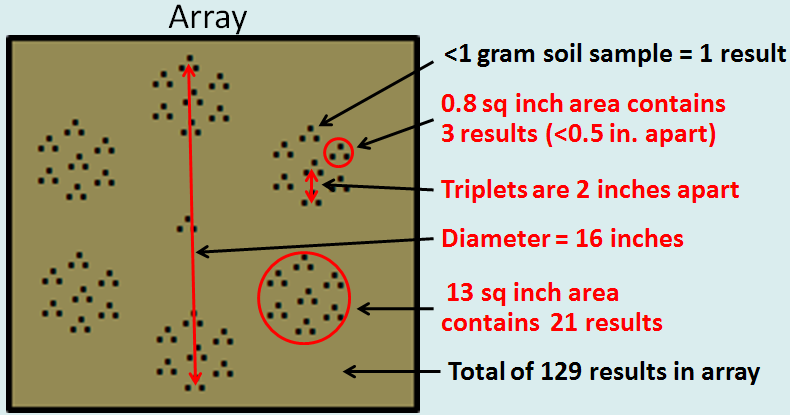
**4.5.2** Becker’s sampling design involved taking single analytical samples (~½ gram each) very close together on the soil surface. The samples were configured as 5 “arrays,” each of which was made up of 6 “subarrays.” Four of the arrays were in the contaminated area. One array was for background concentrations. An array with its 6 subarrays is pictured to the left.

Each subarray contained 7 groupings (a pink circle) of 3 individual sample locations (the 3 dots within each pink circle) where a ½-g of soil was collected at the surface for Pb and Cr analysis by ICP. Sample locations within a triplet were separated by a centimeter (cm, less than a half-inch).

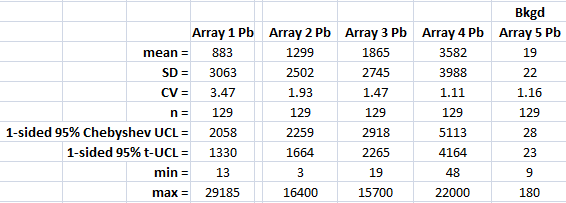
Each subarray was 4 inches in diameter (a circle with an area of 13 sq.in.). A single triplet was located at the center of the array. One entire array was 16 inches in diameter (200 sq.in.).



Dimensions for a single array are shown in the diagram below. The 4 arrays in the contaminated area were lined up side by side, with about 4 inches between adjacent borders. The 5th array was by itself in an uncontaminated area some distance away from the site.



**4.5.3** Before evaluating results of individual samples, let’s look at summary statistics for each array. Each array contains a total of 129 individual ½-gram samples.

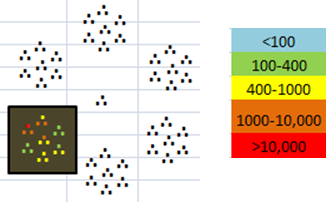


In all cases, the standard deviation (SD) for the array is larger than the mean, indicating that variability cannot be considered “low.” This is reflected in the value for the coefficient of variation (CV, computed as the SD divided by the mean), which are all larger than 1.0 and as high as 3.5.

Notice the range of sample results between the minimum and maximum results for each array. Arrays 1 through 4 show obvious contamination since their means range from 883 ppm to 3582 ppm, as compared to the mean for the background array (19 ppm). Yet these contaminated arrays have some individual samples with values as low as 3 ppm (the “min” for Array 2).

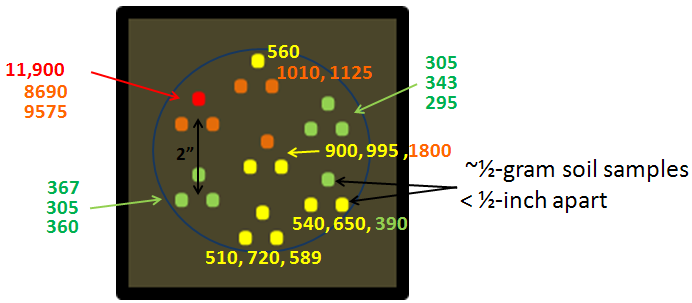
Maximum values (“max”) for Arrays 1 through 4 are at least 3 orders of magnitude (1000 times) higher than its corresponding minimum, and at least 1 order of magnitude (10 times) higher than its corresponding mean. Even the background array (Array 5) had a maximum result that was 1 order of magnitude higher than the mean.

**4.5.4** Now let’s look at the results from a single subarray in an array in the contaminated zone. Array 2 doesn’t have the highest or lowest array concentration or variability, so we’ll look at that one as a moderate example.



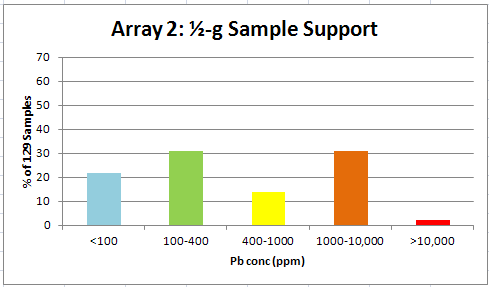
The position of the subarray in Array 2 is shown to the left.

The results of each ½-g sample are color-coded, and the concentration color key is provided. The concentration units are in mg/kg, which is equivalent to [ppm](#Glossary). (Personally I prefer “ppm” over “mg/kg” because it is faster to type and easier to say, so that is what I frequently use.)

Magnifying this subarray allows the 21 ½-g samples to be labeled with their numerical results.

Within this 4-inch diameter circle (a 13-sq.in. area), the 21 results range from 295 to 11,900 ppm. Focus on the center triplet. One of the triplet samples has a result of 1800 ppm, but 1 cm away in either direction the concentration is half of that.

**4.5.5** Tiny sample supports create high variability, and high variability can lead to decision errors if decisions are based on only 1 sample result. **If all of Array 2’s 129 results are averaged together,** **the mean concentration for the 150-sq.in. area covered by the 129 samples is 1300 ppm**. This is definitely over the common screening threshold of 400 ppm for Pb.

A summary of Array 2’s ½-g sample data is shown in the bar graph at left. If only a single ½-g sample were plucked from Array 2, 53% of the time the concentration will be less than 400 ppm (**green** + **blue** bars). **Almost one-quarter of the time (22%), the concentration would be lower than 100 ppm (only the blue bar)!** Within just a 16-inch diameter circle, a tiny grab sample has a 50:50 chance of decision error about whether Pb is greater or less than 400 ppm.

A summary of the chances for individual results to be less than 400 ppm for all four arrays located in the contaminated area are provided in Table 4.5.5. All arrays had the same sampled area (150 sq.in.) and same number of samples (n = 129). Combining these four adjacent arrays gives a total number of 645 samples in a 4 sq.ft. area, for a [sampling density](#Glossary) of 161 samples per sq.ft.

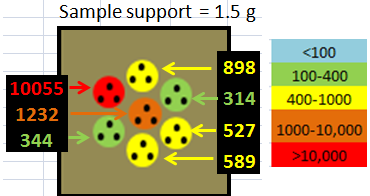
**Table 4.5.5**

Summary of sample results with Pb <400 ppm (“false negative results”) for Becker Study experimental arrays with average concentrations >400 ppm.

|  |  |  |  |
| --- | --- | --- | --- |
| **Array** | **Area Average Pb Concentration (ppm)** | **# of samples <400 ppm** | **% of samples <400 ppm** |
| **#1** | 883 | 97 | 75% |
| **#2** | 1300 | 68 | 53% |
| **#3** | 1865 | 53 | 41% |
| **#4** | 3582 | 4 | 3% |
| **All (n = 516)** | 1907 | 222 | 34% |

Now you might be thinking “the field samples we collect are much larger than ½-g.” And that is true (unless you are taking XRF readings directly on the ground surface). But the areas you are sampling are also much, much larger than 150 sq.in. And without correct sample processing and subsampling of that 100-g or 1,500-g field sample, micro-scale heterogeneity can cause 1-g analytical samples from the jar will have about the same variability as if is they were 1-g grabs straight from the ground.

**4.5.6**  **What happens if we increase the sample support?** Averaging the results of the triplets is the mathematical equivalent to collecting those 3 ½-g samples and pooling them into an analytical single sample with a 1.5-g mass. The graphic below shows what is obtained for the subarray in Array 2 that we’ve been working with.

Instead of 21 ½-g results, we now have 7 1.5-g results, and they are still very variable. Note the pooled triplet with the >10,000 concentration (red):

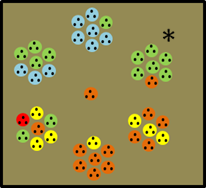
* To its south, it is only 5 cm from a concentration of 344 ppm,
* To its northeast is 898 ppm, and
* To its east is 1,232 ppm.

Within this subarray’s 13-sq.in. area, the average concentration is 1,994 ppm.

* One thing to notice about this larger sample support is that the range of results is a bit narrower (314 to 10,055 ppm) than the range for the ½-g samples (295 to 11,900 ppm).
* Another thing is that a decision based on a 400-ppm threshold for this 13-sq.in. area (mean of 1,994 ppm) has a 2 out of 7 (28%) chance of being made in error if just 1 1.5-g sample is grabbed, just slightly down from the 30% chance when the 21 ½-g samples are considered.

Going from a ½-g to 1.5-g sample support doesn’t seem to improve things very much. So let’s look at another jump in sample support. To do that, we have to go to the level of the entire array.

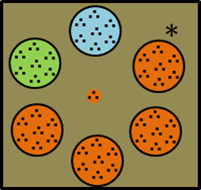
This graphic below shows color-coded results for the 43 1.5-g pooled samples for Array 2 in its entirety. The subarray we were working with is located in the lower left.

*  Concentrations for the 43 pooled samples range from 5 to 10,055 ppm, as compared to the 3- to 16,400-ppm range of results for the 129 ½-g samples.
* The chance of decision error for the entire array (which has a mean of 1,300 ppm) using the 1.5-g sample supports is still about 50:50 (51%), since 22 out of the 43 samples have concentrations less than 400 ppm.
* Also note that even over a circle only 16 inches across (just over a foot) in a known contaminated area (a known hot spot!), there appears to be a pattern where higher concentrations are grouped in the lower half of the circle.

However, imagine this small spatial scale on your typical site. Would knowing that half of a 16-inch diameter circle is clean help you in your cleanup design? Would you try to avoid remediating a clean area of 100 sq.in. (about 0.7 sq.ft.) when the 0.7 sq.ft. of contaminated soil next to it is remediated?

**4.5.7 Increase the sample support to 10.5 g.** If we average all 21 individual results within a subarray, the sample support goes up to 10.5 g (21 x ½-g). The area covered by each 10.5-g sample is a 13-sq.in. subarray. Array 2 now looks like the figure at the top of the next page.

* Using a 10.5-g sample support decreased the chance of decision error for the entire array to 2 out of 6 (about 30%).
* The range for the 6 subarray results is 67 to 1,994 ppm, much improved over the 3 to 16,400 ppm range for the 129 ½-g samples.

You may have noticed the black asterisk in this graphic and in the one before.

* If you look at the previous depiction of Array 2, the asterisked subarray has **6 green triplets** and **1 orange triplet**.
* The **6** **green** triplets have concentrations that range between 141 and 177 ppm (the 18 individual samples making up those triplets range from 130 to 220 ppm).
* The **1** **high** triplet has a concentration of 8,100 ppm (the 3 individual samples making up that triplet were 16,400, 2,800 and 5,100 for an average of 8,100 ppm).
* Despite being outnumbered by 6 low concentration triplets, the high triplet pulled the average concentration of the subarray to 1,295 ppm. Since it is over 1,000, the subarray is **orange** in the graphic to the left.

This feeds into another common question about incremental sampling:

**Won’t high concentration increments be “diluted out” by low concentration increments?**

There are several responses to this:

1. Since a “high concentration increment” may actually only represent a few grams of soil, you don’t want to “find” that spot and be mislead into thinking that many kilograms of soil have that same concentration.
   * There is no need from a risk standpoint to remove a few grams of high concentration material. Although the concentration may be high, the contaminant mass is very, very low because the mass of soil is very small. (See [Section 5.0](#Section_5))
2. When you are trying to determine the average concentration of a DU, you want the extreme highs and lows to cancel out so you can get at the true mean.
3. Don’t underestimate the power of a few high increments to pull up the concentration for a whole incremental or composite sample.
   * Field studies comparing composite sampling to discrete sampling have consistently found that when “significant hot spots” (see [Section 5](#Section_5)) are present, incremental and composite sampling are better at picking them up.
   * Discrete samples are likely to miss hot spots completely (see also [Section 3.0](#Section_3_0)), so the mean of the discrete sample set is usually less than the result of an incremental sample (which more accurately reflects the true mean) when hot spots are present.
   * If the discrete samples only pick up low concentration soil, even the UCL on the discrete sample set may be less than the incremental sample result.

Back to Array 2, it is clear that as sample supports get larger, (i.e., more similar to the spatial scale over which a decision is made), the chances are better that a single sample will have a concentration more representative of the entire decision unit. This agrees nicely with what we saw in the Department of Energy Am-241 study in [Section 4.4.3.](#Section_4_4_3)

**4.5.8** **Let’s not forget the background array.** Even when samples are taken from an area that is clearly not impacted by site contamination, variability is present. If you look back to the chart of array results in Section 4.5.3, you will see that:

* The mean of the 129 samples from the background array was 19 ppm,
* Yet individual sample results ranged from 9 to 180 ppm.

A count of the 129 Array 5 sample data find that most results are less than 20 ppm (n = 106), but 4 results were over 100 ppm. The 180 ppm result was part of a triplet that had 25 and 32 ppm as the other two results.

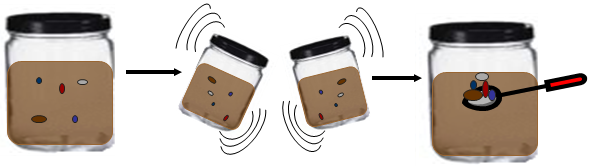
**The inescapable conclusion is that particle effects occur even in “background” soils. This may be due to natural mineralogy or anthropogenic releases, some of which have been going on for millennia (in the “Old World”), and centuries or decades (in the “New World”). Tiny contaminant laden particles can travel great distances on air currents.**

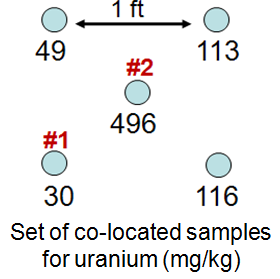
**The chance inclusion of a high concentration particle can greatly raise an individual sample result if its sample support is very small. But an occasional “hot” particle is not terribly important in the context of a DU composed of zillions of mostly “cold” particles.**

**4.5.9 Every Analytical Soil Sample is a Composite!** It has probably already occurred to you that since soil samples are composed of particles, every analytical sample is a composite of many mineral and organic carbon particles with each carrying different contaminant loads. No matter whether the analytical support is 0.5 g or 30 g, or whether the sample is digested, extracted, or read directly (like with [XRF](#Appendix_I)), **every** single sample result is an average of the individual concentrations of the particles within the analytical sample.



In the case of subsampling from a jar, formation of the analytical sample brings together a chance arrangement of particles with a certain average concentration. Disturb the sample, and a different chance arrangement with a different concentration, is formed.



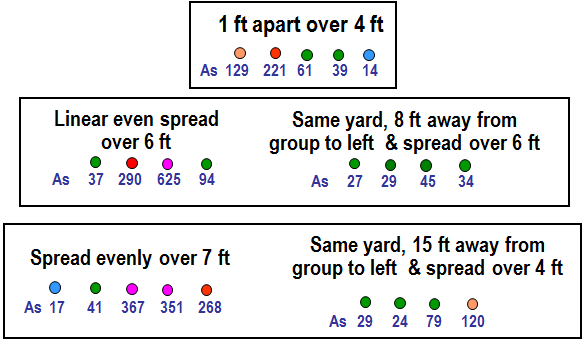
**4.5.10** S**hort-scale heterogeneity at more macro levels.** The Becker study focused on extremely small (centimeters) short-scale between-sample heterogeneity. What about more normal scales of several inches to feet?

The figure at the right shows the results of collocated surface soil samples for uranium.

* The four corner samples are 1 foot apart on the sides.
* The center sample is about 8 inches from its neighbors.
* Compare the results numbered as **#1** and **#2**.

If a sampler picked up a discrete soil sample from position #1, the decision about whether the area is a hot spot would be completely different from the decision made if the sampler had picked the sample from position #2.

The groups of sample results shown below are transects collected from several residential yards. The contaminant measured is arsenic (As, in mg/kg).

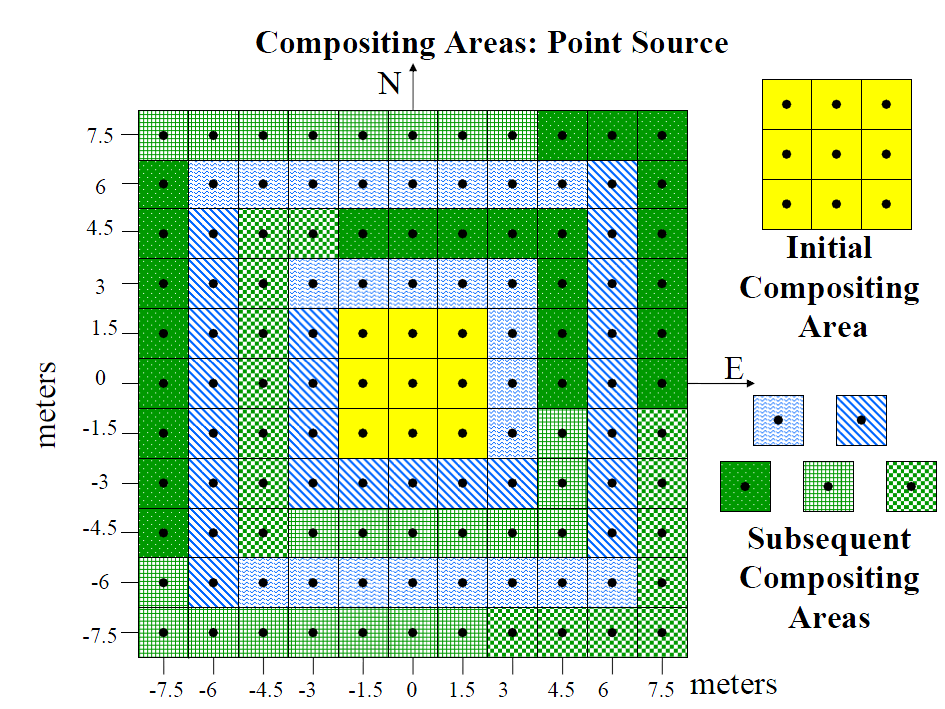


**If based on the results of a single soil sample, or very few soil samples, decisions about “clean” and “dirty” can be a matter of chance), since chance governs where the sampler kneels down to collect that all-important sample!**

**4.5.11** Now that we have considered the degree of spatial variability that is possible with small sample supports, what does that mean for our ability to find and delineate hot spots?

***One of the take-home messages is that you can be in the middle of an actual hot spot, yet individual small samples taken there may have low concentrations.*** Over the area covered by the 4 arrays located in a definitely contaminated area, the average concentration for those 516 samples was 1,907 ppm. Yet 222 of those 516, or 43% of the individual samples, had results less than 400 ppm (see [Section 4.5.5](#Section_4_5_5))! In VSP Hot Spot module terminology, this is a “false negative error rate” of 43%.

**4.5.11.1** **What’s a soil sampler to do?**

1. Think in terms of “averages,” not point concentrations or grab sample results.
2. An average can be sought over a relatively large area, such as a DU, for risk and compliance decisions.
3. But an average can also be sought over much smaller areas, like 1 or so square feet. This is helpful when we are trying to find contaminant boundaries for designing or guiding a cleanup.
   * Since we know that discrete samples are easily biased by unrepresentative pockets of soil, instead of betting all your chips on a single grab sample result, use a small composite instead.
   * [No less than 5 increments](#Appendix_L) should be used in the composite. One possible configuration for a 5-point composite is a 1 ft x 1 ft square with increments in the 4 corners and 1 in the center.
   * Even better, if circumstances allow (vegetation, soil compaction, etc.), all the soil within the 1-sq.ft. area might be scraped up to the target depth to create the sample.
   * As a form of field sampling QC, it is a very good idea to test the adequacy of the composite design (the area the composite covers, and the number and sample support of increments).
     + How repeatable or reliable are the results you obtain from composites collected within an area you think should have about the same overall concentration throughout?
       - If you place two or three 5-point composites right next to each other, do they give similar results?
     + If not, you may need to enlarge the composite’s area and increase its number of increments to exert better control over the degree of short-scale heterogeneity that is present.
     + For example, you might go to a 9, 12, or 16-point composite over a larger area, such as a 2 ft. x 2 ft. area (4 sq.ft.).
     + You are looking for a composite design that suppresses the short-scale “noise” enough so that the spatial signal that is important to your decision-making can be detected.
     + Field sampling QC gives you the evidence you need to know whether the signal you see is real and, therefore, your decisions will be correct, or whether short-scale heterogeneity is creating a signal that really isn’t there.
     + If you enter an area with a different soil type, or where the CSM indicates contaminant heterogeneity might be different, you would want to recheck the repeatability of your composite design and adjust as necessary.
     + Your field sampling QC is vital to establishing the representativeness of your composite data, and as such, should be described in the QAPP. (See [Appendix J](#Appendix_J) for more information.)
4. If the variability in composite replicates is still too high, determine the spatial scale over which you want to detect a consistent concentration trend.
   * For example, what is a convenient removal unit size? If you are using a back-hoe, a convenient removal unit might be a 5-foot wide extension of an excavation.
   * Then set up 5-ft wide [SUs](#Glossary) that are sampled with closer to 30 increments.
   * Remember, the key to reducing short-scale heterogeneity (so you can find trends and patterns at the large scales meaningful to risk determination and cleanup) is to increase the sample support!
5. Composite configurations don’t necessarily have to be square and compact. If working outward from a source area to delineate a hot spot, bent or linear composites may be useful. For example, EPA’s TSCA PCB guidance recommends the configuration to the right to determine the boundaries of a spill area. [(USEPA 40 CFR Ch. 1 §761.283)](#References) Detecting concentration trends along roads or streets might need long, but narrow SUs that parallel the road. Use what makes sense in the context of your CSM and your decision goals.

**4.5.11.2** If designing sampling plans that are to ensure that hot spots of a certain size are included in incremental samples, you must keep in mind the high “false negative error rate” of samples falling within hot spots (recall [Section 4.5.5](#Section_4_5_5)). If using VSP’s Hot Spot modules to determine the spacing of increments or discrete samples in order to catch a hot spot of a certain size, it is wise to use the [VSP Hot Spot module](#Appendix_C) that takes “false negatives” into account. “False negatives” in this context refer to individual samples that pick up “clean” soil even if it is within the bounds of a legitimate hot spot. Unless you have data to the contrary, the safest strategy is to set the “false negative error rate” at 50% when using small sample supports (such as 2-inch cores). VSP will not allow a false negative error rate greater than 50%.

This “false negative error rate” for samples located in a hot spot is yet another reason not to make decisions about “clean” based on single data results. Even sampling in the middle of a hot spot does not guarantee the sample result will be high!

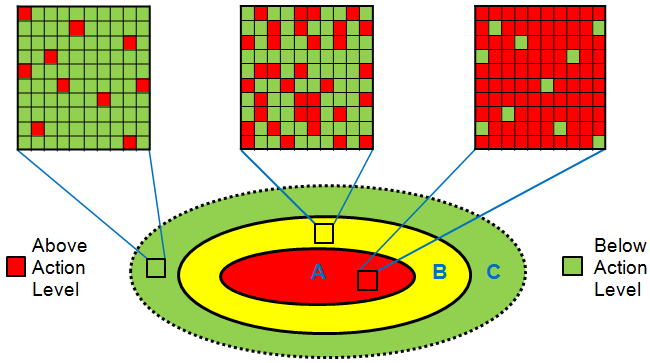
**Sampling error due to soil heterogeneity can cause discrete samples to miss hot spots in 2 ways:**

**1) The low sampling density does not place a sample inside a hot spot, and**

**2) A grab sample taken from within a hot spot has a misleadingly low result.**

**4.5.12** Finally, we are ready to answer the original question: Do you think you can find actual locations and delineate the amount of soil at each location that has a concentration greater than the decision threshold using discrete/grab samples?

* If you stayed with the discussion thus far, you may have concluded that small groups of soil particles, such as ½-g analytical samples, are heavily influenced by the presence or absence of heavily laden particles, as explained throughout this Section. **Therefore contaminant concentrations on micro- and short-spatial scales are patchwork.**
* With patchwork contamination**, the concentration of any single sample is not meaningful**, since it does reliably not predict the concentration of patches around it.

**Figure 4.5.12** Illustration of short-scale variability in contaminant concentration for all potential field samples inside (A) a high concentration zone (such as a spill area), (B) a transitional moderate concentration zone around the spill proper, and (C) compliant soil with only isolated pockets of migrated contamination. Adjacent **green** and **red** blocks illustrate short-scale soil heterogeneity. (Figure adapted from ITRC, 2013)

**4.5.12.1** If results for single individual samples are not meaningful, what is?

**Since all soil samples are composites, all concentration results are averages! The only meaningful concentration value for a heterogeneous soil matrix is an average! Your job is to figure out the volume of soil that needs to have its average determined.**

The average is the value that USEPA’s risk assessment guidance recognizes as important: “As discussed in Section 6.4.1, the concentration term in the exposure equation is the average concentration contacted at the exposure point or points over the exposure period. When estimating exposure concentrations, the objective is to provide a conservative estimate of this average concentration (e.g., the 95 percent upper confidence limit on the arithmetic mean chemical concentration).” (Section 6.5, page 6-24, USEPA 1989)

**4.5.12.2** Another argument for the average concentration being the meaningful measure is given in [Appendix D](#Appendix_D), which relates to the concept of “representativeness” and the “true concentration.” Think back to the Becker Pb study and the 516 samples collected from the 4 arrays placed in the contaminated area. The average over those 516 samples was 1900 ppm, yet the individual concentration results for tiny ½-g clumps of soil particles ranged from 3 ppm to 29,000 ppm. A total of 222 samples had concentrations less than 400 ppm.

If you had all 516 data points in front you, which one of them could you consider to be “representative” of the soil area covered by those samples? 3 ppm? 29,000 ppm? Samples with concentrations less than 400 ppm? Wouldn’t you consider a sample representative if it was close to the mean? There are in fact, a few samples out of those 516 that are close to 1900 ppm. But the chances are small that you’d pick one of them up if you take 1 small grab sample from this area which altogether covers about 4 sq.ft.

**4.6** **That leads us to a grand summary of Section 4.** Environmental practitioners, policymakers, regulators, and problem owners need to recognize the following facts:

* Every soil sample is a composite of particles carrying different contaminant concentrations.
* Because soil is a heterogeneous matrix, sample concentrations exist as a patchwork, such that the concentration of a single sample cannot be relied upon to predict the concentrations of samples around it.
* Therefore, soil discrete/grab samples often give potentially misleading results when a single sample is interpreted in isolation.
* If a true hot spot does exist, you have a poor chance of finding it using traditional discrete/grab sampling strategies having low sampling densities. Incremental sampling is much better at capturing hot spot concentrations in a representative way. (See [Section 3.0](#Section_3_0))
* The mean for a designated volume of soil (the DU) is the most meaningful measure of concentration. The DU size should be selected very carefully based on the CSM and the project decision goals.
* Since we can’t measure the concentration of a DU as a whole, the mean concentration needs to be determined from multiple samplings across the DU (See [Appendix D](#Appendix_D)).
* The concentration and variability of soil samples is governed by their [sample support](#Glossary).
* The ability to produce an accurate estimate of the DU mean depends on having sufficient sampling locations to control for the degree of matrix heterogeneity, as determined by statistical calculations.
* Realistic budgets cannot afford the number of discrete samples that this requires.
* Therefore, incremental sampling is the most appropriate technique for estimating the true concentration of a DU.
* Mathematical tools (such as the spreadsheet pictured in [Appendix B](#Appendix_B)), in conjunction with a solid [CSM](#Glossary) can help determine what size and concentration of potential hot spots are a legitimate concern.
* [VSP](#Appendix_C) is a tool that can be used to determine the increment spacing for a grid design that can ensure that significant hot spots will be incorporated into the incremental sample in the correct proportion (i.e., the proportion in the incremental sample matches the proportion in the field).
* In addition to serving as a means to calculate UCLs, replicate incremental samples (i.e., multiple incremental samples from a single DU) serve an important quality assurance function addressing the question: Was the number of increments sufficient to control the degree of heterogeneity within the DU?

**5.0 What Makes A Hot Spot “Significant”?**

**A Quick Look at Section 5: Deciding When a Hot Spot is “Significant”**

5.1: Use a story about a lost fishing sinker to explore the idea of hot spot “significance.”

5.2: One way a hot spot’s significance might be assessed is by determining the consequences to decision-making if it were missed.

* In other words, if a hot spot is missed, the DU concentration would NOT cross an action level.
* If it is “caught,” the DU concentration WOULD cross an action level.

5.3: Another way “significance” might be assessed is by the imprecision allowed in lab duplicate QC.

* “Allowable” data imprecision is reflected in what %RPD for lab dups is acceptable (per the QAPP) or is accepted (as evidenced by using the associated data for decision-making purposes).
* To be consistent, the acceptable difference in data results stemming from missing a hot spot should be commensurate with the acceptable difference between lab dups.

5.4 Return to the lead sinker story to explore

* The importance of QC when interpreting XRF results.
* Using VSP to predict how many samples are needed to decrease uncertainty in the estimate of the mean.
* The potential financial and database consequences of misinterpreting a high grab sample result to represent a significant soil area/volume, when in fact, it represents only a few inconsequential grams.

We are most concerned with finding “significant” volumes of contaminated soil at large spatial scales (tons of soil), as opposed to finding individual, isolated 1-, 10- or 100-gram pockets with elevated concentrations. So let’s look at how we might determine what constitutes “significant” vs. “insignificant” volumes of contamination.

Let’s explore the topic in the context of a story.

**5.1**  Say the owner of a certain residence is a fisherman. One day he accidentally drops a fishing weight (or “sinker”), which is a hunk of solid metallic lead (Pb), in his front yard on his way to his truck. It lays there in the grass for several years, getting stepped on and pressed into the soil and hidden by the grass. Over time, the sinker develops some [surface corrosion](#Figure_4_1_4) that releases some soluble lead and insoluble lead mineral particles into the soil under and around the sinker (Cao et al 2003). One day his wife is digging up dandelions and she finds the sinker and removes it.

A little pocket of soil in a single yard location is now contaminated with lead, so it could be called a “hot spot.” But is it a “significant hot spot? In other words, if the yard was sampled as part of an urban lead study, would missing a 4-cu.in. (4-sq.in. area x 1-in. depth) hot spot significantly change a decision or an estimation of how much lead contamination is in the yard?

Say the overall concentration for the yard (excluding the little hot spot) is 200 ppm, and say that the little hot spot has an average concentration of 2210 ppm. Is that hot spot significant? Certainly a result over 2,000 ppm sounds worrisome. But consider: the actual mass of Pb contained in that 4-cu.in. hot spot is 258 mg. The mass of Pb in the entire yard (excluding the hot spot) is 54 g (53,900 mg to be exact). The total Pb in the yard is now 54,160 mg. The effect on the overall yard concentration from the additional Pb is that the 200 ppm yard concentration goes to 201 ppm. Would you consider a concentration change of 1 ppm to represent a significant hot spot? Probably not.

**5.2.**  So, how big of a change in yard concentration could be considered “significant”?

* Would a concentration change of 200 to 250 ppm be “significant”? It would take 58 of those sinker hot spots in the yard to change the concentration that much.
* How about a concentration change from 200 to 300 ppm? There would need to be 116 2,210-ppm sinker hot spots to change the concentration that much.
* What about the concentration going from 200 to 450 ppm? It would take 286 sinker hot spots to raise the concentration clearly over the 400 ppm lead screening level.

What do you consider “significant”? Perhaps your answer depends on how close concentrations are to a decision threshold? It is relatively easy to get consensus among different parties about what a significant hot spot is when the discussion is grounded in getting close or exceeding a threshold. It is much harder to find consensus when determination of the size and concentration of a “significant” hot spot seems arbitrary, or falls back on “I’ll know it when I see it.”

Therefore, a useful tool for negotiating hot spot discussions might be spreadsheets set up to assess the effect on DU concentration by theoretical hot spots of different sizes and concentrations.

* These spreadsheets are fairly easy to set up; an example appears in [Appendix B](#Appendix_B).
* They add a semi-quantitative basis to negotiations by projecting when hot spots might become “significant” from a decision-making perspective.
* Having some historical information (like general contaminant concentrations in the area) is very helpful because it allows initial site conditions to be constrained.
* A carefully thought out CSM is also vital to constrain likely sizes and concentrations of potential hot spots. If some of the required inputs are completely unknown, a pilot study may be useful to provide some field-verified, objective inputs.

**5.3.** Another way to evaluate what regulators consider “significant” is the amount of data variability (i.e., data uncertainty or data imprecision) they are willing to accept.

* “Laboratory duplicates” (“lab dups”) are a routine QC check that measures only the data variability that comes from the laboratory subsampling and analysis process (it does not account for data variability in the field).
* It is common for regulator-approved quality assurance project plans (QAPPs) to state that it is acceptable for laboratory duplicate QC for soil samples to be as high as 50% [RPD](#Glossary).
* RPDs may be set that high because that is what practitioners observe in their soil data. This strategy simply ignores the information that the high RPD provides. This is actually not the way QC acceptance criteria should be set.
* What are the implications of a 50% RPD?

**5.3.1** The following bullets demonstrate what the data imprecision represented by 50% RPD means in terms of data results. Approved QAPPs that allow for 50% RPD indicates agreement that it is acceptable that

* A sample jar with a true soil concentration of 300 ppm can be reported as low as 180 or as high as 500 ppm.
* A jar with a true concentration of 400 ppm can be reported as low as 240 or as high as 670 ppm.
* A jar with a true concentration of 600 ppm can be reported as low as 360 or as high as 1000 ppm.

**When lab duplicates specific to a site have high RPDs, all samples from that site should be suspected as having high subsampling variability. Regulators need to understand that the probability of making decision errors about “hot spots” or “clean” areas rises as the degree of data imprecision increases.**

Table 5.3.1 examines the relationships between different lab dup RPD values and the potential for decision error when the sample jar’s true concentration is near a decision threshold.

**Table 5.3.1**

The effect of **subsampling imprecision** (as measured by the [RPD](#Glossary)) on the reliability of a **single** analytical result for lead (Pb), and the likelihood of making **incorrect decisions at a 400 ppm** screening level. The typical tolerable decision error rate is 5% or less (i.e., 95% or higher statistical confidence, equivalent to a “1 in 20” chance).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **True soil Pb conc. in sample jar (ppm)** | **RPD for lab duplicate QC** | **Potential result extremes** | | **Chance of decision error** |
| **Low end (ppm)** | **High end (ppm)** |
| 300 | 50% | 180 | 500 | 21% (1 in 5) |
| 350 | “ | 210 | 583 | 34% (1 in 3) |
| 450 | “ | 270 | 750 | 37% (1 in 3) |
| 500 | “ | 300 | 833 | 29% (1 in 3) |
| 600 | “ | 360 | 1000 | 20% (1 in 5) |
|  |  |  |  |  |
| 300 | 35% | 210 | 427 | 16% (1 in 6) |
| 350 | “ | 246 | 500 | 29% (1 in 3) |
| 450 | “ | 316 | 640 | 33% (1 in 3) |
| 500 | “ | 350 | 712 | 23% (1 in 4) |
| 600 | “ | 420 | 855 | 16% (1 in 6) |
|  |  |  |  |  |
| 300 | 20% | 245 | 367 | 9% (1 in 10) |
| 350 | “ | 286 | 428 | 19% (1 in 5) |
| 450 | “ | 368 | 550 | 23% (1 in 4) |
| 500 | “ | 409 | 612 | 15% (1 in 7) |
| 600 | “ | 491 | 733 | 9% (1 in 10) |
|  |  |  |  |  |
| 300 | 10% | 271 | 332 | 4.5% (1 in 20) |
| 350 | “ | 317 | 387 | 10% (1 in 10) |
| 450 | “ | 407 | 498 | 13% (1 in 8) |
| 500 | “ | 454 | 553 | 7% (1 in 15) |
| 600 | “ | 543 | 663 | 4.6% (1 in 20) |

**5.3.2** **How does subsampling imprecision relate to the significance of hot spots?**

If 50% RPD for lab dups indicates subsampling variability can change the decision by the “luck of the draw,” *and if this is acceptable*, then we must consider the difference between 240 and 670 ppm (400 ± 50% RPD) to be acceptable (i.e., insignificant)!

What does this imply for hot spot “significance”? For a residential yard, the difference in concentration represented by 50% RPD is equivalent to changing the overall yard concentration by 67%. So if the amount of data imprecision represented by 50% RPD is acceptable, it follows that it is also acceptable to miss hot spots that are capable of raising a “baseline” concentration by 67%.

To put this in numerical terms,

* First, consider a yard with a baseline concentration of 200 ppm (i.e., the true mean concentration for the volume of soil considered to be “the yard” is 200 ppm). Now, add a hot spot in the yard that contains enough lead to raise the overall yard concentration by 67%. An increase of 67% causes the “true” (i.e. the overall) yard concentration to go from 200 ppm to 334 ppm [200 + (0.67) x (200) = 334].
  + If the sampling design detects the hot spot in the same proportion as it exists in the yard, the estimated mean concentration of the data set (as represented by the UCL) should be near, or a little above, the yard’s true concentration (i.e., the true mean) of 334 ppm.
  + But if sampling misses that hot spot, the conclusion would be that the yard concentration is around 200 ppm.
  + Since missing the hot spot does not change the decision of whether the yard concentration exceeds the 400 ppm screening level, some may consider the difference between 200 and 334 ppm to be insignificant.
* Now consider another yard with a baseline yard concentration of 300 ppm. A hot spot that increases that yard’s lead concentration by 67% brings the concentration to 501 ppm [300 + (0.67) x (300) = 501]. If 50% RPD for lab dups is considered to be acceptable (i.e., is not significant) for that yard’s data, then to be consistent, the difference between 300 ppm and 500 ppm for the yard concentration is not significant either. So missing that hot spot is not a cause for concern.
  + However, many would disagree and find the difference between 300 and 500 ppm to be significant because it involves crossing a screening threshold.
  + If the hot spot is missed, the conclusion is that the yard’s concentration is 300 ppm (rather than the true concentration of 500 ppm), causing a decision error at the 400 ppm the screening level.

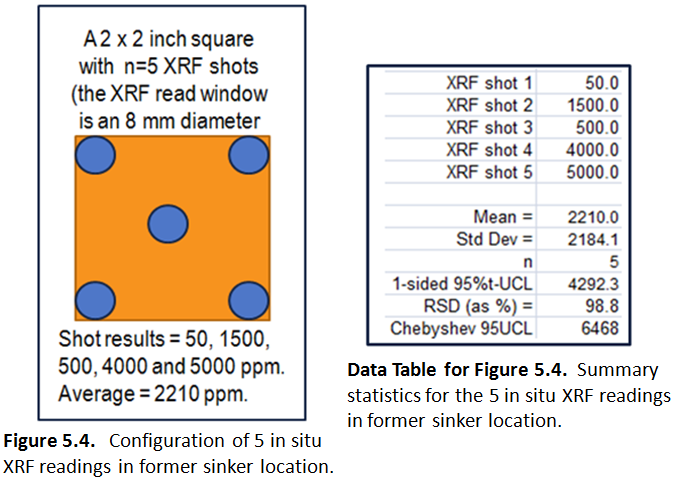
The point is that **it is inconsistent and contradictory to be very concerned about missing hot spots (that could cause a decision error) if you are not also concerned about the effect of poor precision in laboratory duplicates (which could also cause a decision error).**

Both problems have a similar effect: increasing the likelihood of making a decision error based on misleading sample results. In the case of poor subsampling precision, the result is misleading because it is interpreted as representing a large volume of soil in the field, when in fact, the result represents only the concentration of the 1- or 10-gram analytical sample. The result doesn’t even represent the true concentration of the jar’s contents!

**5.4** **Misinterpreting Micro- or Short-scale Variability: A Cautionary Tale**

Let’s return to our Pb sinker story. Say the fisherman’s wife is an analytical chemist, and she happens to have a field-portable X-ray fluorescence (XRF) instrument that can take *in* *situ* readings of soil metals. Just for grins she takes 5 shots on a 4-square inch box (4 corners and the center) on the ground over the sinker location (Figure 5.2.4 below). Her XRF results and some summary statistics are given in the accompanying data table.

Her XRF has an analytical [sample support](#Glossary) area (the window) that is a circle 8 mm in diameter. The X-rays penetrate about 1-2 mm. This amounts to about ½ gram of soil. So a single XRF reading represents the average lead content for all the individual soil particles within that ½ gram of soil the XRF “sees.”



Her 5 *in situ* Pb readings are 50, 1,500, 500, 4,000, and 5,000 ppm. How should she interpret the data? Should she

* Think that the XRF is malfunctioning because the results jump all over the place?
* Take the highest value and say it applies to entire front yard (8 x 8 ft or 64 sq.ft.)?
* Take the highest value and say it applies to the 4-sq.inch area?
* Take the average of the 5 readings (2,210 ppm), and say it applies to the 4-sq.inch area?

Let’s think about it.

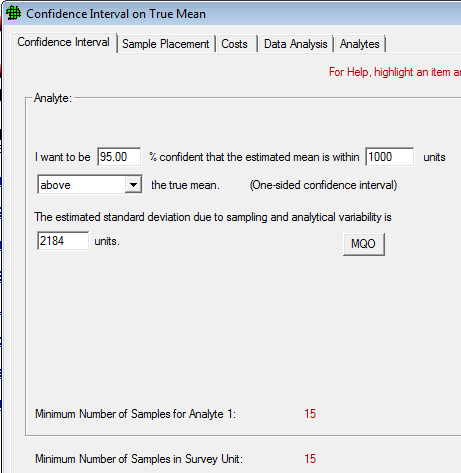
* She knows her XRF is not malfunctioning.
  + She always runs her control sample set (a blank and three reference soils with Pb concentrations equal to 50 ppm, 500 ppm, and 3,242 ppm) before and after analyzing any unknown samples. She plotted the reference soil results on their control charts and saw that they were well within their acceptable ranges.
  + She also ran back-to-back duplicate shots (XRF was not moved between them) to check instrument precision. The instrument precision was 0.9 [%RSD](#Glossary), which is very good.
* She knows that soil Pb contamination from a corroding sinker would show a “nugget effect,” because tiny corrosion particles will have a very high Pb concentration (Figures in [4.1.4](#Figure_4_1_4)) in a “sea” of soil particles with very low Pb concentrations. If no nuggets are seen in the XRF window during an XRF reading, the XRF will report a low concentration. If many nuggets are captured in the shot, the concentration will be very high. So high variability for multiple XRF readings taken over the contaminated soil in slightly different locations is not unexpected.
* She would not assume the maximum concentration represented her entire 64 sq.ft. front yard because she didn’t sample the entire front yard, and her [CSM](#Glossary) is that soil Pb contamination would be localized to a soil volume immediately around the corroding sinker.
* She would not apply the highest value to the whole 4-sq.in. area, again because she knows that nugget effects can skew data sets.
* She would take the average the 5 readings (2,210 ppm), and apply it to the 4-sq.inch area.

**5.4.1** Given the high variability in the data set, how confident would she feel that 2,210 ppm is close to the true mean? One way to assess that confidence is to calculate a UCL for the n = 5 data set.

* Using the Student’s t-UCL assumes a normal distribution, which the n = 5 data set is not. However for the sake of illustration using the [VSP freeware](#References), we’ll stick with it.
* The one-sided 95% t-UCL is 4,290 ppm.
* Recall that the mean is 2,210 ppm.

The fact that the UCL value is nearly 2X the average of the data set is an indication of high uncertainty in this estimate of the mean. Reasons for this large uncertainty are

* There are only 5 data points, and
* The standard deviation (SD) is high. The SD = 2,184 ppm, which is nearly as high as the mean (2,210 ppm).
* Since she is a curious person, she wonders how many XRF shots would be needed in order to calculate a more confident estimate of the mean, say an upper estimate (a UCL) that is only 1,000 ppm higher than the calculated average, rather than the current 2,080 ppm (4,290 - 2,210 = 2,080).
* She opens the freeware program, Visual Sample Plan (VSP), and uses the module for calculating a “Confidence Interval in a True Mean.” (see the VSP screen shot below)
* She enters 95% as the desired statistical confidence level.



* She enters 1,000 as the maximum allowable distance between the estimated mean (which is the UCL) and the true mean (which she does not know).
* She only cares about the UCL. She doesn’t care about the LCL (the lower bound of the estimate of the mean), so she selects the word “above” in the drop down menu and the phrase “(One-sided confidence interval)” pops up.
  + If she wanted to know the full range where the true mean might reside, she would be interested in the UCL and the LCL at the same time. In that case she would select “above and below” from the

drop-down menu and the phrase “(Two-sided confidence interval)” would pop up.

* She puts the SD (2,184) into the last box.
* VSP reports she would need a total of 15 data points to increase statistical confidence to the point where the UCL is 3,200 ppm.
  + This assumes that the SD does not change much after collection of the 10 additional data points.
  + If the SD increases for the n = 15 data set, she might not meet her goal of having the UCL within 1,000 ppm of the calculated sample mean.

**5.4.2** One day she gets a call from the City asking permission to access her property to take soil samples for a neighborhood soil Pb screening. She agrees, but forgot to tell them about the 4-cu.in. pocket of sinker-contaminated soil.

* The City health department-sponsored urban Pb investigation sampled her yard one day when she wasn’t there to tell them about the sinker location. Since the front yard is so small, the City contractor decided only 1 grab sample was needed to “characterize” those 64 sq.ft., and he happened to take the sample in the former sinker location.
* A 100-gram grab sample was collected. It was an aggregated dry clay clod that was slid undisturbed into the sample jar and sent to the lab. The lab tech dug his 1-gram analytical sample off the top of the uppermost clod where many Pb-laden corrosion particles happened to be. The result was 6,000 ppm.
* A laboratory duplicate happened to be done on this same sample. The lab tech dug this second subsample from the edge of the clod where there were very few Pb nuggets. The result was 200 ppm.
* The large difference between these laboratory duplicates was noticed by the data validator, who dutifully qualified all 17 field sample results in this same analytical batch as estimated (i.e., J-qualified). The 17 samples included the 6,000 ppm sinker result along with 16 samples from different yards in the same city block.
  + In practice, J-qualified data are treated like data without qualifiers, because there is no quantitative uncertainty associated with “J.”
    - So there is no mathematical way to deal with the “estimation” of J-qualified data.
    - The uncertainty implied by the J could be just a little, or it could be a great deal.
    - There is no way of knowing unless the reason for the J-qualifier, and its quantitative implications, are retained with the data when it goes to the data user and into databases
    - Without this information, it is very difficult to judge whether J-qualified data are suitable for a secondary data use.
  + There are several reasons why a J-qualifier might be applied. The reason for the J-flag for this data set is explained in the data validator’s report. But after data have been labeled as “validated,” the qualifier may be lost or go unnoticed.
  + Therefore, the particular reason for this J-flag (extreme subsampling variability at an RPD of 187%) is lost to data users as soon as the 6,000 ppm J-flagged result was efficiently uploaded into the City’s database.
  + Along with 200 other Pb results from the same neighborhood, the “6,000 ppm J” data point was uploaded without the intervention of a knowledgeable human who might question the wisdom of loading such an extreme result or who might investigate the J-qualifier.
* Since it is difficult for a human to evaluate hundreds and thousands of data results, the City has efficiently programmed the database to generate an average Pb concentration on a block-by-block basis. The database was also programmed to automatically calculate a one-sided 95% t-UCL on a block average without checking whether the Student’s t-UCL was the appropriate UCL for that data distribution, which is a no-no.
* All of the other yard data for that block ranged between 80 and 500 ppm.
  + If the 6,000 ppm result was not there, the block’s mean would be 348 ppm, and the UCL on that mean would be 377 ppm (i.e., less than the City’s 400 ppm screening level).
  + With the 6,000 ppm result included, however, the block average is falsely elevated to 490 ppm and the UCL to 730 ppm.
* The computer tags this block for attention because the mean and UCL both exceed the 400 ppm screening level.
  + The data set is examined, and the “6,000 J” result is noticed.
  + The Health Department asks the City to task the contractor to investigate the high result.
  + Another work plan, which included a Field Sampling Plan, QAPP and safety plan, were prepared for the new field work. Although the contractor only had to cut and paste from the original work plans, the City was still charged for creating the “new” work plan.
  + More contractor and staff time was used to review the QAPP by both the contractor and City organizations.
* A couple months later, the contractor revisited the yard, and took 4 grab samples: one from each quarter of the 64-sq.ft. yard.
  + Several weeks later, the four grab sample results came back from the lab with results ranging between 80 ppm to 150 ppm.
  + The contractor explained to the Health Department that “soil heterogeneity” can cause results to differ, and this might be the reason for the discrepancy, although it could have been a “lab error.” The Health Department’s concerns over that yard were alleviated, and the “alarm” note in the database was removed.
* These four new results were put in the queue for uploading to the database, but the data entry technician did not receive any instructions to remove the 6,000 ppm result.
* Now the database contains five results from that one yard, while there is only one result per yard for the rest of the block. In the future, when someone calls up the average and UCL for that block, the database will again alert on the block’s statistics since the average and UCL are now 456 and 674 ppm, respectively.

[This is the end of the Hot Spot FAQ text. The Annotated Glossary, References and Appendices follow.]

**ANNOTATED GLOSSARY** for terms, acronyms, and concepts as used in this paper

(Note: italicized words indicate that they also have an entry in the Glossary)

**Aliquot** - Term used to refer to a portion of a liquid solution or solid matrix that is taken OUT for analysis. “Aliquot” is somewhat synonymous with “subsample.”

* “Aliquot” is often used incorrectly in the site cleanup industry to refer to an “increment.” This is inappropriate and confusing because these terms are NOT synonymous.
* “*Increment*” is the proper word for field samples which are ADDED together (or pooled) to form a *composite sample* or an *incremental sample*.

**Analytical quality**  - The degree to which evidence demonstrates that all steps in an analytical process were performing with acceptable bias and precision.

* Major steps in the analytical process include
  + Sample preparation, usually extraction (in the case of organic contaminants) or digestion (in the case of metal/metalloid contaminants),
  + Extract cleanup (if required), and
  + Analysis of the extract by an instrument.
* Quality control (QC) checks associated with each step demonstrate whether its performance was within acceptable limits.
* “Quality” is judged by the ability of the data to be used for their intended purpose. Since there are many types of decisions which vary in their need for data rigor, analytical quality that is acceptable for making one decision may not be acceptable for making a different decision.
* When QC checks associated with a data set are out of their control limits, analytical quality may be characterized as “poor” or “unacceptable” if the analytical process step is sufficiently out of control so that the data results are excessively uncertain with respect to the intended decision-making.
* If documentation for all QC (or the QC for critical process steps) is lacking, the data are said to be of “unknown” quality. It is USEPA policy that only data of known quality will be used (USEPA 2002b, pages 2 and 15).
* Although subsampling is generally performed in the laboratory, and the laboratory performs the QC check called “laboratory duplicates,” subsampling precision is usually considered part of sampling quality rather than analytical quality.

**Analytical sample**  - The portion of a soil sample submitted to the laboratory that actually undergoes extraction or digestion to dissolve target analytes into a liquid that can be injected into an instrument for measurement. The term is interchangeable with the term “analytical subsample.”

**Analytical variability** - Imprecision in data results attributable to the analytical process of extraction/digestion of an analytical sample, (possibly) cleanup of the ensuing extract, introduction of the extract into the analytical instrument, and the operation (calibration, signal stability, maintenance, etc.) of the analytical instrument itself. The degree and sources of analytical variability are measured by *analytical quality* control (QC) checks. Variability is typically measured in terms of standard deviation (SD), relative standard deviation (*RSD*) or relative percent difference (*RPD*).

**Area of influence** - The area of soil surrounding a sample that is considered to have the same concentration as the sample. The term was introduced in EPA regulations concerned with *PCBs*. This concept is equivalent to the “sampling unit” concept, which more explicitly considers soil volume, rather than just area. Refer to the Glossary entry for “*Sampling unit*” for additional details.

**bgs**  - below ground surface

**Bulk soil** - Generally, native soil that has not been sieved. However it may contain components that are not considered to be “soil” in its strictest sense, such as twigs and other macro plant fragments, living creatures (insects, worms, etc.), stones larger than 2 mm, man-made debris, or trash.

**Bulk soil sample** - A soil sample expected to be representative of native “soil.” “Soil” is defined as mineral and organic material that is less than 2 mm in diameter (USDA 2009), which is achieved by sieving with a 10-mesh sieve. A bulk soil sample contains both the fine [usually considered the less than 60-mesh fraction (<250 µm or 0.25 mm)] and the coarse [fraction between 60- and 10-mesh (0.25 to 2 mm)] fractions. The soil fraction that does not pass through a 10-mesh sieve is often referred to as “over-sized” material.

**Collocated samples**  - Soil samples collected a few inches to a few feet apart. The expectation is that collocated sample concentrations will be similar until an interface between a contaminated area (such as a “hot spot”) and uncontaminated soil is reached. Unfortunately, this expectation is seldom realized because soil contamination can be highly heterogeneous at the scale of field sample collection. Contaminant deposition is frequently in the form of atmospheric particles (bullets and explosives and their fragments and dusts, paint chips, ash, etc.) or droplets (sprayed pesticides, etc.), and therefore very “spotty” at short scales. Even if contamination deposition was originally fairly uniform (such as a liquid spill), over time small scale environmental processes will create heterogeneity by

* Moving soil around (wind, water, burrowing animals, foot traffic, landscaping and construction activities, etc.), or
* Altering local contaminant fate and transport (biodegradation, organic matter that changes local redox or pH conditions, geology or mineral make-up of the soil, etc.) that concentrates contaminants in certain mineral or organic fractions.

It is typical to find that differences in contaminant concentrations range over one, two, or even several orders of magnitude for samples collected a foot or even a centimeter apart (See Becker study discussion in Section 4.5). This means that the result from one sample may not reflect the true concentration for a larger volume of soil surrounding that sample. Single sample results can be much higher, or much lower, than the true concentration of the surrounding soil. Collocated samples provide a measure of short-scale heterogeneity so that the sampling design, or data interpretation, can take it into account.

**Composite sample** - A sample prepared by combining a relatively small number of increments into a single field sample that is then processed the same as an *incremental sample* would be to control micro-scale heterogeneity. Composite samples are collected for a different purpose than incremental samples, and are comprised of fewer increments (less than 30). Since it has less than 30 increments, the “mean” for the area covered by a composite sample is not expected to be as accurate as a mean produced by an incremental sample. Common purposes for composite samples include

* Reducing the effect of short-scale heterogeneity on a “discrete” sample result by collected no less than 5, and preferably at least 9, increments over a 1- to 4-sq ft area.
* Sampling of individual “sampling units” (SUs) within a decision unit (DU) to preserve some spatial information about high-concentration areas (i.e., locate “hot spots”).

**Conceptual site model (CSM)**  - A written or pictorial representation of an environmental system and the biological, physical, and chemical processes that determine the transport of contaminants from sources through environmental media to environmental receptors within the system. (ASTM E 1689 – 95). The CSM should include the contamination release mechanisms; fate and transport mechanisms and how long they have been acting on the contamination and creating short-scale heterogeneity; what receptors are present; what exposure pathways may exist; and what, if anything, will be done about the contamination and/or intact exposure pathways.

**Correct samples**  - Term used in Gy-based *Theory of Sampling* to designate samples for which representativeness is known and documented. The analytical data from correct samples can be relied upon to support correct decisions. See also *Representativeness (assessing)*.

**cu.ft.** - cubic feet - a measure of volume

**Data quality** - Data quality refers to the fitness of data for its intended use. Since soil data come from soil samples, data quality must include sampling quality (see Glossary entry) as well as analytical quality (see Glossary entry). The analysis can be perfect, but if the sample was “wrong” (perhaps degraded or mislabeled or the wrong particle sizes analyzed), the data quality is “bad” in that the results can give misleading information.

**Data uncertainty** - A lack of confidence that data can be used for a particular application. Data are uncertain when

* There is insufficient documentation that explains how samples were collected, processed, or subsampled (i.e., *sample representativeness* is unknown).
* There is insufficient documentation of QC that documents sources of sampling and analytical variability, so that the bias and precision of the data are unknown.
* QC documentation shows that data is too biased, or data precision is too poor, to support confident decisions.
* Data uncertainty is always present to some degree for soil data, however, data uncertainty becomes excessive when the risk of incorrect decisions exceeds tolerable levels.

USEPA’s data quality policy is that “EPA will produce data of known and documented quality and non-EPA organizations will submit data of known and documented quality to EPA.” (USEPA 2002b, page 2).

**Data validation** - Data validation is a process examining the analytical side of data generation. The validation process may look at the laboratory QC checks for sample extraction and digestion and for extract cleanup. Data validation always includes evaluation of the laboratory QC used to evaluate instrument calibration and performance, and may check mathematical calculations. In this way, data validation is important to establish the precision and bias of the pure *analytical* process. However data validation does little to establish data representativeness or estimate data uncertainty.

* At most, the effect of micro-scale heterogeneity is estimated by laboratory duplicates and matrix spike/matrix spike duplicates, as long as the duplicates and matrix spikes were performed on *your* site-specific soil, and not reported from some other project’s soils in the same analytical batch as yours.
* The effect of short-scale heterogeneity may be estimated *if* you collected collocated samples from your Site.
* If lab duplicates, matrix spike/matrix spike duplicates or collocated samples show “excessive” variability, the data validator may qualify associated data as “estimated.” However, there is no quantitative information included that might help you understand how much decision uncertainty is created by the “estimated” data. So “estimated” data are usually used exactly the same as unqualified data.

**Decision support**  - The spatial dimensions of the decision unit (DU), including the target particle size, which define the soil *population* of interest.

**Decision threshold** - Any type of numerical value used for an exceedance/non-exceedance decision, such as a “screening level,” an “action level,” a “cleanup level,” a “criterion,” etc.

**Decision unit (DU)**  - A volume of soil, defined by location, length, width and depth, over which a specific decision is made. The specifics of the decision is what determines the decision unit.

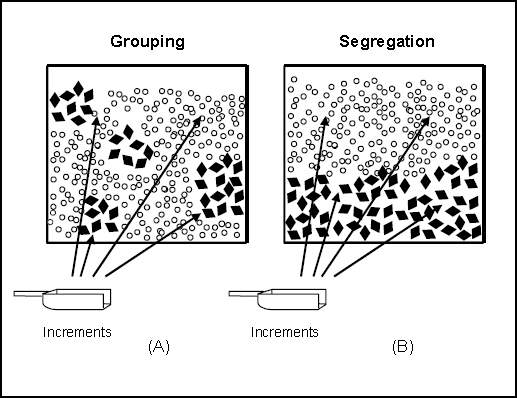
* “Define the smallest, most appropriate subsets of the *population* (sub-populations), waste, or media to be characterized based on spatial or temporal boundaries. The boundaries will define the unit of waste or media about which a decision will be made. The unit is known as the decision unit.” (USEPA 2002a, page 38).
* Because soil contains a wide range of particle sizes, and some measured properties (such as contaminant concentration) depend on the soil particle size analyzed, the target particle size must always be designated when the DU is defined. (USEPA 2002a, page 58).

**Disaggregate (a soil sample)**  - To break apart clumps of soil particles that are simply stuck to each other, including dried lumps of clay (which made of individual microscopic particles). Disaggregation is often performed by crushing soil clods using fingers, a hand-operated mortar and pestle, a coffee grinder, a rubber mallet, etc. Disaggregation does NOT involve particle size reduction, which is the breaking apart or crushing of individual solid particles by milling (See “*Milling*” entry). Although the term “grinding” is often used with the same meaning as “milling,” grinding is a more ambiguous term (See “*Grinding*” entry). Soil that is “ground” in a mortar and pestle or coffee grinder will be disaggregated, but not milled. Those types of “grinders” do not generate enough force to break apart typical mineral and hard rock particles (sands, pebbles, etc.). However, mills are able to reduce solid rock particles to the consistency of flour.

**Discrete soil sample** - Refers to a soil sample obtained from the parent matrix by scooping or coring from a single location. May also be termed a “grab sample,” especially if the sample has been collected without consideration of a statistically-valid sampling design or a representative *sample support*. Grab samples are almost always “*incorrect samples*.”

**Grinding (of a soil sample)** - The term “grinding” is ambiguous unless more information is provided (see *Disaggregate* above) Soil that is “ground” in a coffee grinder or in a hand-operated mortar and pestle will only be disaggregated because they do not have the force needed to fracture non-friable mineral and rock particles (sands, pebbles, etc.). When “grinding” is used in the context of particle size reduction, the term “milling” is probably more appropriate. See the *Milling* entry below.

**Grouping and segregation error** - Grouping and segregation error is one of the 7 sampling errors described in the publications of Pierre Gy. The “grouping” phenomenon can be present in the field or in a sample. It is caused by the “tendency for some particles to associate into groups of like particles due to gravitational separation, chemical partitioning, differing moisture content, magnetism, or electrostatic charge. Grouping and segregation of particles can lead to sampling bias.” (USEPA 2002a)

The figure to the left (from USEPA 2002a) depicts particle grouping (in panel “A”) and segregation (in “B”). Analytical samples formed from just one group of particles in “A” would yield biased results. Particle segregation in “B” could result from gravitation separation (e.g., during sample shipment). If the contaminant of interest was associated with only one class of particle (for

example, only the black diamond shapes),

then a sample collected from the top would

result in a different concentration than a sample collected from the bottom, thus biasing the sample.

**Gy-compliant (procedure)**  - Sample collection, sample processing, sample splitting, and subsampling procedures that comply with the *Theory of Sampling* by using activities that minimize sampling errors, in particular, fundamental error, grouping and segregation error, and increment delimitation and extraction error. Using Gy-compliant procedures for all steps of the sample collection and analysis process produces “*correct samples*” for which representativeness is known. See the references (USEPA 2002a, USEPA 2003 and Petersen et al 2005) for more complete discussion and graphic illustrations of Gy sampling errors.

**Heterogeneity** - The condition of being non-uniform in composition or other properties, such as particle size, shape, color, density, etc. Heterogeneity is split into 2 parts: Compositional Heterogeneity, CH (also called Constitutional Heterogeneity by some) and Distributional Heterogeneity (DH).

**Heterogeneity (Compositional), CH** - “CH is inherently dependent on the composition, shape, size, density etc. of the particles or fragments making up the lot. If a great overall composition-wise difference between the individual fragments exists, the constitution heterogeneity is large, but if the fragments are more homogeneous CH is lower. The total contribution to heterogeneity is never nil, however, as that would be the case of all fragments being strictly identical. Mixing and blending does not change CH. The only way to alter the constitution heterogeneity of any given material would be by crushing (comminution) or by other methods changing the physical properties of a sample. The reduction of the average grain-size is the dominating factor in reducing CH by such means.” (Petersen et al 2005)

**Heterogeneity (Distributional), DH** - DH is dependent on the spatial distribution of particles. “For the sake of illustration, imagine a lot consisting of black and white spheres and with significantly different grain-size distributions. If all the black spheres are to be found at the bottom of the lot and the white spheres are more to the top, the system displays a very high DH. If on the other hand the spheres were to be well mixed, the system DH would be significantly reduced.” (Petersen et al 2005)

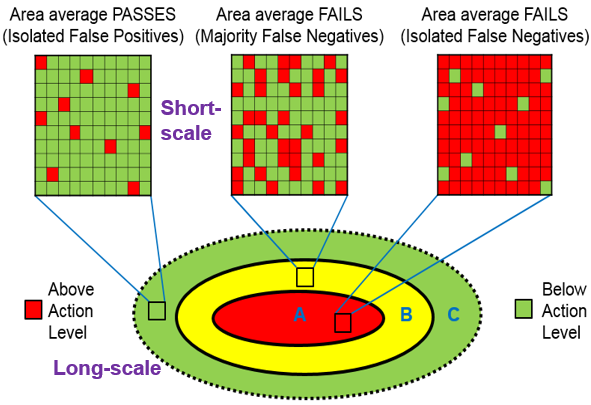
* **What is short-scale heterogeneity?** A: Differences in concentration at spatial scales that are too small to be important to decision-making. The scale is too small to allow separation of “clean” vs. “dirty” soil during cleanup, and too small to be meaningful to the receptors identified during risk assessment. (Note: the “meaningful” spatial scale can be vastly different depending on the receptor, such an earthworm vs. a human resident vs. a fox.) The only way to handle short-scale heterogeneity is to average it over the DU, which represents the spatial scale at which heterogeneity becomes important to decision-making.

As illustrated in the figure below (ITRC, 2013), individual grab samples (represented by the smallest square blocks) can have concentrations that differ from the overall (i.e., average) concentration for an area. So reliance on single grab sample results for making risk or cleanup decisions can be erroneous if the single grab sample gives a misleading result.

* **What is long-scale heterogeneity?** A: The spatial scale at which heterogeneity becomes important to decision-making. It is the scale of exposure units and cleanup designs. If DUs are set correctly, it is the difference between DU concentrations. If a few dirty DUs are intermixed with mostly clean DUs, it is the scale at which it is efficient and cost-effective to clean up the dirty DUs while leaving the clean ones in place.

As illustrated in the figure below (ITRC, 2013), differences in the true (i.e., average) concentrations for defined areas are called long-scale (alternatively, “large-scale”) when the defined areas are large enough to be relevant to risk assessment and cleanup actions (such as digging up soil with a shovel or backhoe, or *in situ* treatment with amendments).

* **Why is it important to control the effects of short-scale heterogeneity?** Because short-scale heterogeneity creates “noise” in data that can mislead you into thinking it represents long-scale heterogeneity. It can cause you to think the DU concentration is either falsely high or falsely low.



(Figure from ITRC, 2013)

**Incorrect samples** - Term used in Gy-based *Theory of Sampling* to designate samples for which representativeness is not known and cannot be determined. No matter how good the analytical quality is, data from these samples are not reliable to support decision making. See also *Representativeness (assessing)*.

**Increment**  - A portion of matrix that is pooled with other portions to form a *composite sample* or an *incremental sample*. “An increment is defined as a partial sample unit that, when combined with other [increments], provides a final sample. This procedure is known as composite sampling.” (Petersen et al 2005)

* **This term should be used instead of the term “*aliquot*,”** which actually has the opposite meaning. An “increment” is something added in or added together, an “aliquot” is something taken out, like a portion of extract taken from a flask to inject into an analytical instrument.

**Incremental sample**  - A sample prepared by combining many increments (at least 30 for “typical” analytes and “typical” contamination scenarios) into a single field sample. The entire field sample is then rigorously processed and subsampled so that subsampling error caused by micro-scale heterogeneity is controlled.

* The purpose of incremental samples is to produce a reasonably accurate estimate of the true mean concentration for the area covered by the incremental sample (usually ¼- to 1 acre).
* For highly heterogeneous analytes, more than 30 increments may be required. The exact number of increments can be determined by measuring the short-scale heterogeneity (as standard deviation) for the increment’s sample support, and calculating “n” using statistical formulas.
  + For some scenarios, such as firing and bombing ranges, 50 to 100 increments per DU (with DU area less than 1 acre) may be needed to control extreme short-scale heterogeneity.
* The number of increments per DU can also be determined based on the increment density needed to ensure that hot spots of a minimum specified area will be confidently included in the incremental sample.
* For DU areas up to 3 acres, more increments should be used to maintain an increment density of no less than 30 per acre.
* For DUs with areas larger than 3 acres (such as some types of exposure units), a DU can be split into SUs (each having areas of 1-acre or less). Statistically based selection and sampling of SUs can be done to create a set of SU results for which the data distribution, mean, standard deviation, and confidence limits can be calculated to characterize the DU.
* Incremental Sampling Methodology (ISM) was developed by the ITRC ISM Team, and guidance is posted at [www.itrcweb.org/ISM-1](http://www.itrcweb.org/ISM-1)

**Laboratory duplicates**  - A laboratory QC check where two separate subsamples are taken from the same soil sample jar and analyzed. When all QC checks downstream of subsampling are in control, yet the laboratory duplicates show poor precision (as measured by a high RPD or RSD), the problem lies in the heterogeneity of the soil sample and the failure of sample processing and subsampling techniques to reduce micro-scale heterogeneity.

**Micrometers** - One millionth of a meter, or one thousandth of a millimeter (0.001 mm) or one twenty-five thousandth of an inch. A micrometer is also called a “micron.” Abbreviated as µm.

**Milling (of a soil sample)** - Mills are able to fracture and crush hard particles to sizes as little as 37 micrometers, which is the size of plant pollen. The terms “pulverization” and “comminution” are synonymous with “milling.” The types of mills commonly used with soil samples include various types of laboratory grade ball mills and ring and puck mills. Usually, up to 500 - 800 grams of soil can be milled at one time. Particle size reduction from 2 mm to less than 75 µm can be achieved in one to two minutes, which is enough to reduce subsampling variability for many analytes to less than 10 %RSD for subsample masses of 2 grams (and up). When analytes could be destroyed or volatized by heat, soils are milled in 2-5 short pulses separated by 2 minutes of cool-down time.

* In a field study of soil metals, milling of upland soil samples from less than a 2-mm particle size to less than 75 µm reduced subsampling variability from 37 %RSD (n = 5, averaged across 17 metal analytes) to 7 %RSD (n = 5, averaged across 17 metal analytes) (unpublished data from USEPA Region 3 Ft. Eustis study, DU-4).
* In the same study, silty marshland soils (DU-6) showed little change in subsampling variability from before-milling (average of 8 %RSD) to after-milling (average of 7 %RSD) because the particle size for the un-milled silty samples was already very small in relation to the analytical sample mass (2 grams).

**Parametric/nonparametric statistical distribution**  - The statistical distribution of a data set is assessed by a frequency plot (similar to a histogram) of the data. (If frequency plots and histograms are unfamiliar, Wikipedia is a good resource.) Frequency plots can take any number of shapes.

* The shape is called “parametric” if the curve can be plotted by a mathematical formula. Examples of parametric data distributions commonly observed with environmental data sets are normal distributions (which is another name for bell-shaped curves), lognormal distributions, and gamma distributions.
* Nonparametric data distributions cannot be fit to a mathematical formula.

**Polychlorinated biphenyls (PCBs)** - A family of compounds sharing a common carbon skeleton of two linked benzene rings. Individual family members differ in the number and arrangement of

chlorine atoms attached to the rings and is called a “congener.” There are 209 congeners in the PCB family. Until banned, mixtures of PCB congeners, trade-named Aroclors, were most commonly produced as fire suppressants that were added to oils used in electrical equipment. PCBs also had a number of other uses.

**Population (of soil)** - A volume of soil that shares a common characteristic. The project decision defines the soil population of interest. For example,

* Background concentrations may define a population of un-impacted soil. Concentrations higher than background may define a population of impacted soil.
* A soil containing more than 80% sand may define one soil population; soil containing between 50 and 80% sand may define another population; between 20 and 50% sand may define a third population, and less than 20% sand may define a fourth soil population as one moves from a beach area to upland areas.
* The shared characteristic can be as simple as the particles are all in the same container.

**Population (of potential soil samples)**  - ALL potential soil samples within a *decision unit* or other defining boundary. If a soil sample is considered to be 100 grams of soil in a jar, and the decision unit is a mass of soil weighing 2000 kg, then 20,000 potential soil samples make up the population defined by the decision unit.

If within-jar heterogeneity is high so that 100 1-gram subsamples from the same jar give very different results (i.e., a single 1-gram subsample cannot be expected to be representative of the entire 100 grams), then it would be more appropriate to use the analytical sample mass to determine the number of samples in the population. For the 2,000 kg DU, there would be 2 million potential 1-g analytical samples making up the population of soil samples. The true concentration of the entire population (i.e., the DU) is equal to

* The single result obtained if the entire DU could be analyzed in a single analysis, and
* The average of all 2 million analytical samples making up the population. This value is called the “true mean.” Note that there is no uncertainty present for the true mean, so a *UCL* would never be calculated on a true mean.

**ppm** - parts per million; ppm is shown to be equivalent to mg/kg by the following (remember that same units can cancel when on opposite sides of division sign):



**Replicate samples** - Samples that undergo the exact same procedure(s) in order to measure the precision of the procedure(s). Two replicates are commonly called “duplicates,” three replicates can be called “triplicates,” four replicates can be called “quadruplicates,” and so on. Common types of replicates used as QC in incremental-composite sampling designs are listed below. (Also see *sampling quality*)

* Field replicates (i.e., replicate incremental samples): Two or more incremental samples are collected from the same area/volume of soil, using the same number of increments, but from off-set increment collection locations.
  + Field replicates are collected to measure the ability of the field sample support (the number and mass of increments) to represent the true mean of the area/volume being sampled.
  + Variability between field replicates represents the total data variability and captures the variability in the entire chain of sample collection, processing and analysis processes.
  + To isolate the degree of field heterogeneity, the other two main sources of data variability (sample processing/subsampling variability and analytical variability) must be measured and subtracted from the total data variability.
  + If field heterogeneity is too high, more increments per incremental sample are needed in order for the incremental sample result to be closer to the true mean.
* Subsampling replicates: Two or more subsamples taken from a single field sample. By necessity, subsampling replicates include the variability from sample processing (sieving, milling, etc.) and from the subsampling process (grab subsampling vs. incremental subsampling).
* Analytical replicates: This measure of variability should be obtained from repeated measures of a laboratory control sample (LCS) over the course of several analytical runs.
* In order to perform the math of partitioning variability among its sources, all measures of precision should be expressed in terms of relative standard deviation (*RSD)*.

**Representative soil sample** - Key features of a representative soil sample:

* “A representative sample is one that answers a question about a decision unit with an acceptable level of confidence. This requires a complete understanding of the data quality objective (DQO) process, selecting the appropriate sampling design and strategy, and including proper quality controls to assess sample representativeness…A sample that is representative for a specific question is most likely not representative for a different question.” (Ramsey & Hewitt, 2005).
* “ ‘Representative sample’ means a sample of a universe or whole (e.g., waste pile, lagoon, ground water) which can be expected to exhibit the average properties of the universe or the whole.” (40 CFR 260.10)
* A sample that is representative of a *decision unit* will allow confident extrapolation of the sample’s result to apply to the DU as a whole.
* “[T]he sample mass [must be] commensurate with the heterogeneity of the material to be sampled. Only representative samples in this context will ensure that the conclusions from the chemical analysis and subsequent data analysis will be reliable.” (Petersen et al 2005)

**Representativeness (assessing)**  - “It is not possible to ascertain whether a particular sample is

representative from any kind of inspection or characterization of the sample itself. Only a full qualification of the sampling process can lead to recognition of representative samples.” (Petersen et al 2005). Therefore, assessing sample representativeness must include

* Full understanding of the specifics of the decision(s) the data were collected to support;
* Full understanding of the *decision unit* that the sample is supposed to represent;
* A good understanding of seven sampling errors described in Gy theory. (See *Theory of Sampling*.)
* Full knowledge of the *sample support* and the mechanics involved with sample (or increment) collection and handling in the field, processing of the sample to obtain the *target particle size*, and subsampling to obtain the *analytical sample*.
* QC results that measure the between-sample variability in the field.
* QC results that measure the between-subsample variability in the laboratory.
* QC results for laboratory control samples (LCSs).
* Those three kinds of QC data can be used to partition total data variability into its three main components: field variability, sample processing and subsampling variability, and *analytical variability*. See also *sampling variability*.
* See also the discussion in [Appendix J](#Appendix_J).

**Representativeness (of a soil sample)** - The degree to which a soil sample shows the same properties of the volume of soil from which it came, which is called a sampling unit. Depending on the context, soil sample representativeness can be expressed in more specific, operational terms, such as

**1)** For hotspots: the proportion of mass from “hot spots” IN THE SAMPLE is the same (or nearly so) as the proportion of “hot spot” mass IN THE FIELD decision unit or sampling unit;

**2)** Defining the *target particle size*: the ratio of target soil particle types (sizes and composition) in the sample is the same (or nearly so) as the ratio of the same throughout the entire decision unit or sampling unit;

**3)** In terms of a true concentration: a measured property of the *analytical sample* gives the same data result as the property if it were measured for the entire DU or SU in a single analysis, which would be the same result that would be obtained if all millions or billions of potential analytical samples within the DU/SU could each be measured and then those results averaged to find the true mean for the DU/SU.

**Representativeness (of an analytical sample)** - The ratio of soil particle sizes in the *analytical subsample* is the same as the ratio in the entire container holding a fully processed field sample. It is unlikely that the analytical subsample will represent the field sample unless processing and subsampling is performed according to specified procedures [which are determined by the decision(s) the data are too support]. These procedures will minimize particle segregation and grouping in the container so that each particle in the processed field sample has an equal chance of ending up in the analytical sample.

**Responsible party (RP)** - A person or entity with cleanup or institutional control responsibilities under a cleanup program.

**RPD** - relative percent difference; a measure of imprecision when only two replicate procedures (i.e., duplicates) were performed.

* The most common formula for RPD is to subtract one replicate from the other, and divide that difference by the average of the two replicate results. The fractional result is then multiplied by 100.
* If the fractional number is not multiplied by 100, the value is called a relative difference (RD).
* Usually the absolute value is used so that RPDs are expressed as positive numbers.
* Since “percent” is in the name, an RPD value is typically rounded to a whole number.
* It is mathematically impossible for an RPD value to be higher than 200.
* An RPD cannot be calculated directly from three or more replicates, however, an RSD (see below) from three or more replicates can be converted into an RPD value by the following formula: RPD = RSD x 100 x sqrt(2)
  + note: “sqrt(2)” is Excel notation for the square root of 2

**RSD** - relative standard deviation; a measure of imprecision when two or more replicate procedures were performed

* The formula for RSD is the standard deviation (SD) for the replicates divided by the replicates’ mean.
* RSD is always a positive number, and expressed as a decimal.
* When RSD is multiplied by 100, it becomes %RSD and may be rounded to a whole number. For example, an RSD of 0.253 can be expressed in terms of %RSD as 25.3 or 25.
* An RPD value from a duplicate pair can be converted into an equivalent RSD or %RSD value by the following formulas: RPD/100/sqrt(2) = RSD or RPD/sqrt(2) = %RSD.
* An RSD can be calculated directly from two replicates using a statistical calculator or a spreadsheet to calculate the standard deviation, and then divide by the mean of the two values.
* **NOTE! You cannot directly compare RPD and %RSD values.** The same numerical value for anRPD and RSD do not represent the same level of precision.An RPD of 20% represents better precision than a %RSD of 20 (which is the same as an RSD of 0.20). For the same level of precision, an RPD is a higher number than the %RSD. An RPD of 20% is the precision equivalent of 14 %RSD (i.e., an RSD = 0.14). A %RSD of 20% is the same precision as an RPD of 28%.
* If you want to compare precision between two sets of replicates, one set having two replicates and the other set having three or more replicates, and you have the actual data, you can use a spreadsheet to directly calculate the RSD for each set. You can directly compare the precision of the two data sets using their RSDs.
* If you only have summary data as a mix of RPD and RSD values, use a conversion equation to convert to either all RPDs or all RSDs before assessing which has the better precision. Conversion to RSD is the more convenient option.
* Note that an average of several RSD values is NOT the simple mathematical average of those values. Because of the squared relationship between SD and variance (the SD is the square root of the variance), RSD values must first be squared to their variance equivalents. Then average the variance equivalents. Finally, take the square root of that average to get the average RSD.

**Sample (physical)**  - An individual or single portion of the population from which data is gathered.

**Sample (statistical)**  - A set of measurements from a group of physical samples that were collected in such a way that the properties of the population can be deduced.

**Sample support** - “The weight, shape (length, width and height dimensions), and orientation of a sample…The appropriate size of a sample (either volume or mass) can be determined based on the relationship that exists between the particle size distribution and expected sampling error.” (USEPA 2002a, page 94).

* “[T]he sample mass [must be] commensurate with the heterogeneity of the material to be sampled. Only *representative samples* in this context will ensure that the conclusions from the chemical analysis and subsequent data analysis will be reliable.” (Petersen et al 2005)
* The sample support must mirror, to the extent possible, the “*decision support*” (i.e., the dimensions of the decision unit, including the target particle size). So, if the *decision unit* is 6 inches deep, then the sample support of grab samples or of increments must be 6 inches deep (such as a 6-inch core).
* If the *target particle size* in the decision unit is everything less than 250 micrometers, then sample processing must produce analytical samples comprised only of native particles (i.e., after disaggregation and before milling) less than 250 micrometers.

**Sample unit** - (Contrast with *sampling unit*) The unit upon which a measurement is made. For example, if the average height for a population of white men is desired, the thing that is measured is the height of an individual man that is part of the specified population. No matter how the measurement is made (metric or English scale, backed against a marked wall or with a tape measure, etc.) the sample unit is still a man. For almost all statistical scenarios, the unit upon which the measurement is made is obvious: a coin toss, a car wreck (for car insurance statistics), a test score (for education assessment), etc. The sample unit cannot be divided and still produce the measurement.

This is not the case for many environmental measurements, especially a concentration measurement. There is no natural, obvious soil or water unit upon which measurements are made. Any amount soil or water may be collected in the field, as long as there is enough to perform the analysis. If 10 grams is required to perform an analysis, at least 10 grams must be collected, but usually much more is collected. There might be enough for 5 or 10 or 50 analyses. Then, if the sample gets sent to a different lab, they may only require 5 grams to run the same analysis.

**Sampling quality**  - The degree to which evidence demonstrates that all steps related to acquiring [representative](#Glossary) samples from the field, and preserving that representativeness through laboratory subsampling were performed with acceptable bias and precision in the sample collection, processing, and subsampling.

* Major steps in the representative sampling process include
  + sample collection from the DU,
  + sample processing to obtain the target particle size (usually involves soil drying, disaggregation and sieving),
  + determining the appropriate subsample mass given the maximum particle size in the material to be subsampled,
  + possibly additional sample processing (i.e., grinding/milling) if the target particle size is too large with respect to the intended analytical sample mass, and
  + [representative subsampling](#Glossary) done either by manual incremental subsampling or mechanical sample splitting (as with a rotary sectorial splitter).
* Quality control checks on representative sampling include
  + QC Replication of incremental samples from the same DU on all, or some, of the project DUs.
    - Although primarily a measure of precision, replicate incremental samples also provide some reassurance that sample collection in the field was not biased.
    - If all downstream processes show good performance (as measured by their QC checks), and the replicate increment sets are spaced so there is uniform coverage of the DU (show fig of spaced & non-spaced), the chance is very low that three or more replicate incremental samples will give the same biased value.
    - If the number of increments are too few to effectively manage the field heterogeneity, then the incremental results are more likely to disagree, rather than all converge on the same wrong result.
  + QC Replication of sample splitting between parties (if performed)
    - This replication answers the question: “Did both parties receive the “same” samples?”
    - If field samples are to be split between laboratories (such as the regulator’s lab and the RP’s lab), sample processing should be as complete as possible before samples are split.
    - Sampling splitting should be done using an appropriate technique recommended by [USEPA’s Subsampling guidance](#Appendix_A) (USEPA 2003)
    - The splitting process should be replicated for at least one sample of the group.
      * If a mechanical splitting mechanism is used that generates all splits at the same time, and two parties are receiving splits, then at least 6 splits should be generated: three random splits go to Party A, and three to Party B.
      * If a manual splitting process generates splits one at a time, the two parties should alternate receiving sequential splits.
    - If the analytes of interest are not labile and holding times will not be violated, some splits can be archived until data on the others is obtained.
      * Archived splits can be analyzed only if disagreement between the parties’ data sets needs to be resolved.
  + QC Replication of sample splitting between different analyses (if performed)
    - If field samples need to be split for different analyses (such as between chemical analysis and toxicology work), as much sample processing should be performed as possible before sample splitting. “As much as possible” means that processing should be performed up to the point where further processing would compromise the representativeness of the sample for certain tests.
    - Remaining heterogeneity is measured by replicating the splitting process similar to splitting between parties discussed above.
  + QC Replication of subsampling
    - At least one set of triplicate subsamples should be collected from one of the first samples submitted to the laboratory.
    - Subsample replication measures both the adequacy of sample processing and the adequacy of the subsampling technique.
    - If the triplicate precision is acceptable, subsequent subsampling replication can take the form of laboratory duplicates at their usual frequency.
    - If a different matrix is encountered such that the adequacy of sample processing for that matrix is unknown, the first sample of that matrix should undergo triplicate subsampling.
  + Based on QC replication at critical sample handling steps, data variability can be partitioned

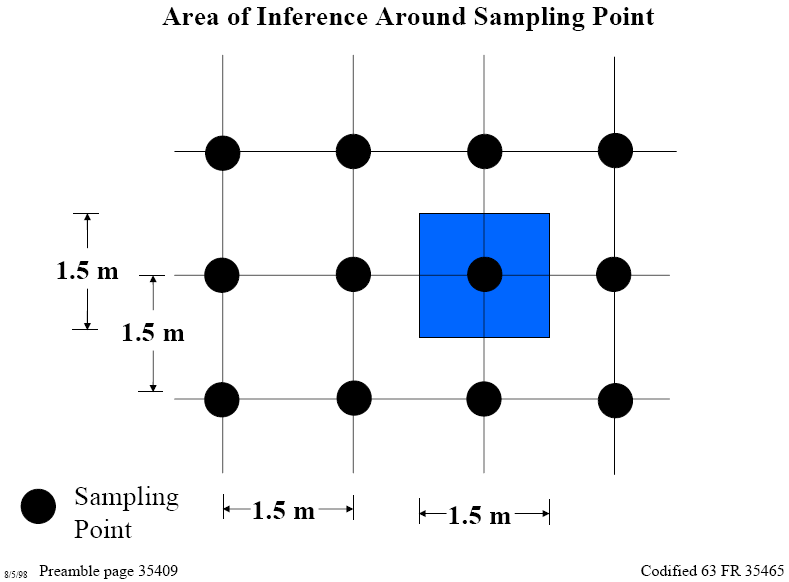
**Sampling density** - The number of discrete samples or increments per area or volume of soil. For example:

1. 10 discrete samples/acre is a lower sampling density than 30 discrete samples/acre.
2. 30 discrete samples/acre is the same sampling density as 30 increments/acre.

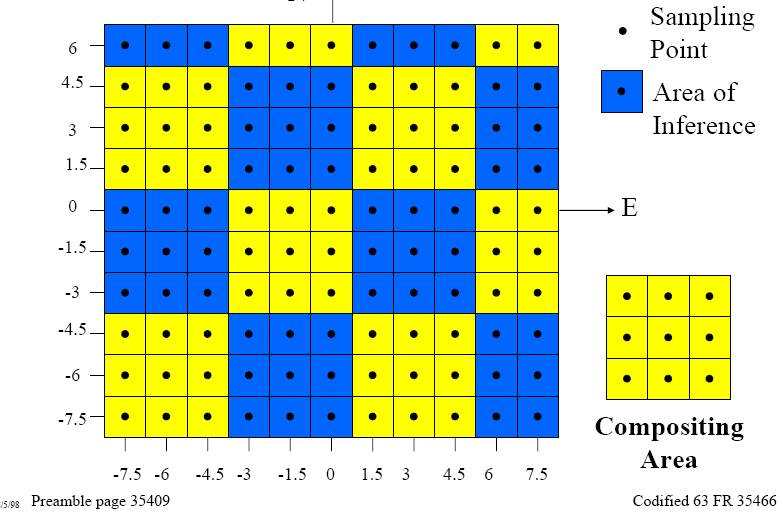
**Sampling error** - See *sampling variability*

**Sampling unit (SU)**  - (Contrast with *sample unit*) The area or volume of soil intended to be represented by an individual sample result. The single sample result can be from a *discrete sample*, a *composite sample*, or an *incremental sample*. This contrasts with a *sample unit*.

* A sampling unit is the same concept as the “area of inference” discussed in USEPA’s PCB regulations (USEPA 40 CFR §761.283(d) and 63 FR 35466) which states (for discrete samples): “Analytical results from an individual sample point apply to the sample point and to an area of inference [around the sample point as shown in the graphic below].” (40 CFR §761.298)



* When interpreting discrete data results, the PCB regulations say: “For grid samples which are chemically analyzed individually, the PCB concentration applies to the area of inference as described in § 761.283(d).” (40 CFR §761.298)
* For composite samples the PCB regulations (USEPA 40 CFR §761.283(d) and 63 FR 35466) say: “The area of inference [for] a composite sample is the total of the areas of the individual [increments] included in the composite.” (Illustrated by the graphic below using 9-point composites.)



* When interpreting composite data results, the PCB regulations say: “For grid samples analyzed as part of a composite sample, the PCB concentration applies to the area of inference of the composite sample as described in § 761.283(d) (i.e., the area of inference is the total of the areas of the individual samples included in the composite).” (40 CFR §761.298) See also 63 FR 35409. **Note that the composite sample’s result is NOT divided by 9 (the number of increments) for comparison to the decision threshold!**
* An SU will be the same as the DU when the DU is sampled using an incremental sample composed of increments taken from across the entire DU.
* An SU will be smaller than the DU if the DU has been partitioned into subunits which are sampled separately with their own set of increments.
  + When a DU is divided into several SUs as a means of preserving some spatial information, all SUs are sampled.
  + When large DUs (>~5 acres) to very large DUs (100s of acres) are expected to have fairly homogeneous concentrations at the spatial scale of ½- to 1-acre, they may be divided into ½- to 1-acre SUs for statistical sampling.
    - In other words, a random statistical selection of >8 SUs from the DU are incrementally sampled using ~30 increments per SU.
    - The >8 SU results are averaged and a UCL calculated to estimate the mean concentration of the DU.
    - If there is too much uncertainty in the UCL, additional SUs may be sampled and added to the statistical data set to calculate a more confident estimate of the DU mean.
* An SU is never larger than a DU.

**Sampling variability** - Imprecision in data results due to unmanaged *heterogeneity* within the sample matrix. The term “sampling error” is synonymous with “sampling variability.” The word “error” is commonly used in statistics to refer to variability or imprecision. Sources of sampling variability include

* Short-scale heterogeneity (see *collocated samples*);
* Large-scale heterogeneity in the parent matrix (i.e., the soil in the field) at the scale at which hot spots and source areas are sought for identification and removal (see also section 3.1 of this paper);
* Insufficient mass of the field sample in relation to the maximum particle size in the DU;
* Improper sample handling or processing of the field sample that causes segregation of particles [see *Representativeness (of an analytical sample*)] prior to sample splitting or subsampling;
* Insufficient mass of the analytical sample in relation to the maximum particle size in the field sample at the time of subsampling for analysis; and
* Improper subsampling techniques.

The degree and sources of sampling variability are measured by *sampling quality* control (QC) checks. Variability is typically measured in terms of standard deviation (SD), relative standard deviation (*RSD*) or relative percent difference (*RPD*).

**Slabcake subsampling** - An entire, fully processed sample is spread out in a pan (such as a foil-lined cookie sheet) to about ¼- to ½-inch depth. As many as 30 small increments are taken from the full thickness of the slabcake and combined to form the analytical sample. This is a *Gy-compliant* procedure for subsampling matrices composed of solid particles.

**sq.ft.** - square feet - a measure of area

**Target particle size fraction**  - The particle size fraction of soil that is representative of the decision to be made. Selecting the correct target particle size is important because the concentration of soil contaminants generally increases as the particle size fraction analyzed decreases (i.e., the old proverb: “the contamination is in the fines”). This is best illustrated by examples:

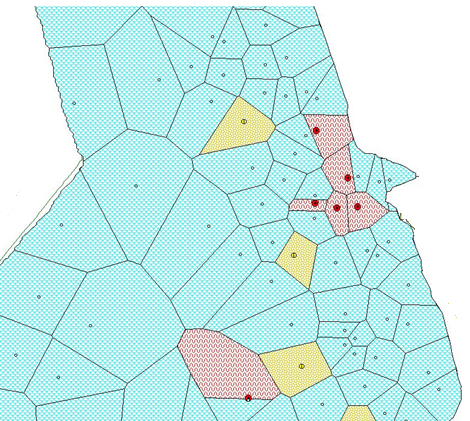
1. When analyzing soil to determine LANDFILL DISPOSAL, the target particle size fraction is that defined as a “*bulk soil sample*” [i.e., all particle sizes less than 2 mm, which is what passes through a 10-mesh sieve (USDA 2009)], since that is representative of the soil that would be placed in the landfill.
2. When analyzing soil to determining EXPOSURE risk by an accidental hand-to-mouth pathway, the target particle size fraction is recommended to be smaller than 250 *micrometers*, which is what passes through a 60-mesh sieve (USEPA 2007). This recommendation was based on research on what particle sizes are likely to adhere to skin. Some researchers suggests that a size smaller than 250 µm should be used, such as less than 175 µm (an 80-mesh sieve).

**Theory of Sampling (TOS)** - “TOS shifts the focus from the fallacy of looking upon sampling as a mere materials handling matter, towards a scientific, quantifiable and well-documented process of exactly how to obtain representative samples…” (Petersen et al 2005)

* “TOS states…‘**Correct samples’** (truly *representative samples*) only originate from a qualified sampling process, in contrast to ‘**incorrect samples’**—which are all samples that cannot be documented to be representative in the above fashion.” (Petersen et al 2005)
* “[Incorrect samples] are worthless in the context of particulate sampling, as one is in principle unable to even assess the magnitude of sampling errors present; neither is there any way to compensate for these errors anywhere in the ensuing analytical process.” (Petersen et al 2005)

**Thiessen Polygons**  - A method of determining the “*area of influence*” for irregularly spaced discrete samples. Lines are placed ½-way between adjacent data points, and oriented perpendicular to a line connecting the two points. (See Brassel and Reif, 1979)

* Within a Thiessen polygon, all the space within that area is closer to its centered sampling point than it is to any other sampling point.
* Although a useful technique in geographic information systems (GIS), and often used to display contamination data, it encourages the odd notion that the soil concentration is the same as the sample result for only 5 ft. in one direction (because the next sample in that direction is 10 ft. away), while the same concentration holds for 50 ft. (10 times further) in the other direction simply because the next sample location in that direction is 100 ft. away.



**UCL** - Upper confidence limit on the mean. A statistical way to derive an upper estimate of the mean. The UCL is calculated by adding a “safety factor” to the mean obtained from the sample set. The “safety factor” takes into account the number of samples used in the calculation, the variability in the sample results, and how confident you want to be that your estimate of the mean does not underestimate the true mean. There are different kinds of UCL calculations depending on the statistical distribution of the data. See [Section 2.2.2](#Section_2_2_2) and [Appendix F](#Appendix_F) for more details.

**Van der Waal’s forces** -Van der Waals forces include attractions and repulsions between atoms, molecules, and surfaces caused by fluctuations in electron orbits such that a slight, temporary electric charge is created on the atom, molecule, or surface. The magnitude of this charge is much less than the charge on an ion and is much less stable.

**REFERENCES**

ASTM E 1689 – 95 *Standard Guide for Developing Conceptual Site Models for Contaminated Sites*. Reapproved in 2008.

Becker, Joanna (2005) Perdue Univ. Ph.D. Thesis *Centimeter scale analysis of soil heterogeneities within a long-term, heavy metal contaminated site.*

Becker, J., M.T. Parkin, C.H. Nakatsu, J.D. Wilbur and A. Konopka (2006) Bacterial Activity, Community Structure, and Centimeter-Scale Spatial Heterogeneity in Contaminated Soil. *Microbial Ecology Vol. 51*, 220-231.

Boyd, Stephen A. (2010) Sequestration of Dioxin by Clays, Superfund Research Brief No, 183, March 3, 2010, page 2. <http://tools.niehs.nih.gov/srp/researchbriefs/view.cfm?Brief_ID=183>

Brassel, K.E. and D. Reif (1979) A Procedure to Generate Thiessen Polygons. *Geographical Analysis Vol 11*, Issue 3, pp 289-303. July. Available for free at

<http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1979.tb00695.x/pdf>

Cao X., L.Q. Ma, M. Chen, D.W. Hardison, Jr., and W.G. Harris (2003) Weathering of Lead Bullets and Their Environmental Effects at Outdoor Shooting Ranges*. Journal of Environmental Quality Vol 32*, No. 2, 526-534.

Doctor, P. G. and R.O. Gilbert (1978) *Two studies in variability for soil concentrations with aliquot size and distance*. In Selected Environmental Plutonium Research Reports of the Nevada Applied Ecology Group (NAEG), eds. M. White and P. Dunaway, 405–449. Las Vegas, NV: U. S. Department of Energy.

Hadley, P., E. Crapps, and A. Hewitt (2011) Time for a Change of Scene. *Environmental Forensics*, 12:312-318. Free article access at: <http://www.tandfonline.com/doi/abs/10.1080/15275922.2011.622344>

ITRC (Interstate Technology and Regulatory Council) (2013) ITRC Training Webinar Slides: *Soil Sampling and Decision Making Using Incremental Sampling Methodology - Part 1* (Dec 3, 2013). Available at <http://www.clu-in.org/conf/itrc/ISM_120313/> Also, the second part of this 2-part series, *Soil Sampling and Decision Making Using Incremental Sampling Methodology - Part 2* (Feb18, 2014) is available at

<http://www.clu-in.org/conf/itrc/ISM_021814/>

ITRC (Interstate Technology and Regulatory Council) (2013) *Incremental Sampling Methodology: Representative Sampling, Confident Decisions.* A web-based document available at <http://www.itrcweb.org/ism-1/>

ITRC (Interstate Technology and Regulatory Counsel) (2003) *Characterization and Remediation of Soils at Closed Small Arms Firing Ranges.* January. Available on-line at <http://www.itrcweb.org/SMART-1.pdf>

Lewis R.G., C.R. Fortune, R.D. Willis, D.E. Camann, and J.T. Antley (1999)Distribution of Pesticides and Polycyclic Aromatic Hydrocarbons in House Dust as a Function of Particle Size. [*Environmental Health Perspectives Vol 107, No. 9, September.*](http://www.ehponline.org/cgi-bin/findtoc2.pl?tocinfo=Environmental%20Health%20Perspectives@107@9@1999) <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1566465/>

McKee, E.D. (2006) Photomicrograph: Scanning electron microscope photograph of Supai Group clay minerals: possible smectite flakes from the Manakacha Formation in Tuckup Canyon; Magnification = x 23,500. Grand Canyon National Park, Arizona. U.S. Geological Survey. *This work is in the public domain in the United States because it is a work prepared by an officer or employee of the United States Government as part of that person’s official duties under the terms of* Title 17, Chapter 1, Section 105 of the US Code.

<http://libraryphoto.cr.usgs.gov/cgi-bin/show_picture.cgi?ID=ID.%20McKee,%20E.D.%20%20316>

Mitchell, J.K. (1993) *Fundamentals of Soil Behavior*, 2nd edition, John Wiley & Sons (Chapter 3).

Myers, Jeffrey C. (1997) *Geostatistical Error Management: Quantifying Uncertainty for Environmental Sampling and Mapping*. NY, NY: Van Nostrand Reinhold.

Petersen, L., P. Minkkinen and K.H. Esbensen (2005) Representative sampling for reliable data analysis: Theory of Sampling. *Chemometrics and Intelligent Laboratory Systems Vol. 77*, pages 261-277. <http://www.sciencedirect.com/science/article/pii/S0169743904002448>

Potapova, E., I. Carabante, M. Grahn, A. Holmgren, and J. Hedlund (2010) Studies of Collector Adsorption on Iron Oxides by in Situ ATR-FTIR Spectroscopy*. Industrial & Engineering Chemical Research Vol. 49*, No. 4, 1493-1502.

Ramsey C. A. and A. D. Hewitt (2005) A methodology for assessing sample representativeness. *Environmental Forensics* 6:71-75. <http://www.tandfonline.com/doi/pdf/10.1080/15275920590913877>

Richard O. and Pamela G. Doctor. 1985. *Determining the Number and Size of Soil Aliquots for Assessing Particulate Contaminant Concentrations.* Journal of Environmental Quality Vol 14, No 2, pp. 286-292.

U.S. Department of Agriculture (USDA) (2009) USDA Natural Resources Conservation Service Soil Survey. *Field and Laboratory Methods Manual, Soil Survey Investigations Report No. 51*, Version 1.0, National Soil Survey Center, Lincoln, NE (Rebecca Burt, ed.)

<ftp://ftp-fc.sc.egov.usda.gov/NSSC/Lab_References/SSIR_51.pdf>

USEPA 40 CFR Chapter 1. §260.10 Hazardous Waste Management System General - Subpart B - Definitions. <http://www.gpo.gov/fdsys/pkg/CFR-2013-title40-vol27/pdf/CFR-2013-title40-vol27-sec260-10.pdf>

USEPA 40 CFR Ch. 1 §761.283 Polychlorinated Biphenyls (PCBs) Manufacturing, Processing, Distribution in Commerce, and Use Prohibitions - Subpart O - Determination of the number of samples to collect and sample collection locations. Available at <http://www.gpo.gov/fdsys/pkg/CFR-2007-title40-vol30/pdf/CFR-2007-title40-vol30-part761-subpartO.pdf> For graphical illustrations, see also <http://www.epa.gov/epawaste/hazard/tsd/pcbs/pubs/subpartmopr.pdf> and <http://www.epa.gov/epawaste/hazard/tsd/pcbs/>

USEPA 63 FR 35409 and 35466, 29Jun1998 <http://www.gpo.gov/fdsys/pkg/FR-1998-06-29/pdf/98-17048.pdf>

USEPA (1989). *Risk Assessment Guidance for Superfund, Volume 1: Human Health Evaluation Manual (Part A)—Interim Final*. Washington, DC: Office of Emergency and Remedial Response. EPA/540/1-89/002. December. <http://www.epa.gov/oswer/riskassessment/ragsa/index.htm>

USEPA (2002a) *RCRA Waste Sampling Draft Technical Guidance: Planning, Implementation, and Assessment.* Office of Soil Waste. 530-D-02-002, August. <http://www.epa.gov/osw/hazard/testmethods/sw846/pdfs/rwsdtg.pdf>

USEPA (2002b) *Overview of the EPA Quality System for Environmental Data and Technology*. Washington, DC: Office of Environmental Information. EPA/240/R-02/003. November. <http://www.epa.gov/quality/qs-docs/overview-final.pdf>

USEPA (2003) *Guidance for Obtaining Representative Laboratory Analytical Subsamples from Particulate Laboratory Samples*. EPA/600/R-03/027. November. [http://www.clu-in.org/download/char/  
epa\_subsampling\_guidance.pdf](http://www.clu-in.org/download/char/epa_subsampling_guidance.pdf)

USEPA (2007) *Short Sheet: Estimating the Soil Lead Concentration Term for the Integrated Exposure Uptake Biokinetic Model (IEUBK),* Office of Solid Waste and Emergency Response (OSWER) 9200.1-78, September. <http://epa.gov/superfund/lead/guidance.htm#estim>

USEPA (2011) *User Guide - Uniform Federal Policy Quality Assurance Project Plan Template for Soils Assessment of Dioxin Sites,* Office of Superfund Remediation and Technology Innovation, September. <http://www.epa.gov/superfund/health/contaminants/dioxin/pdfs/Dioxin%20UFP%20QAPP%20UserGuide.pdf>

Vance, David B. (1994) Iron – The Environmental Impact of a Universal Element. *National Environmental Journal*, May/June. 1994 Vol.4 No. 3 page 24-25. <http://2the4.net/iron.htm>

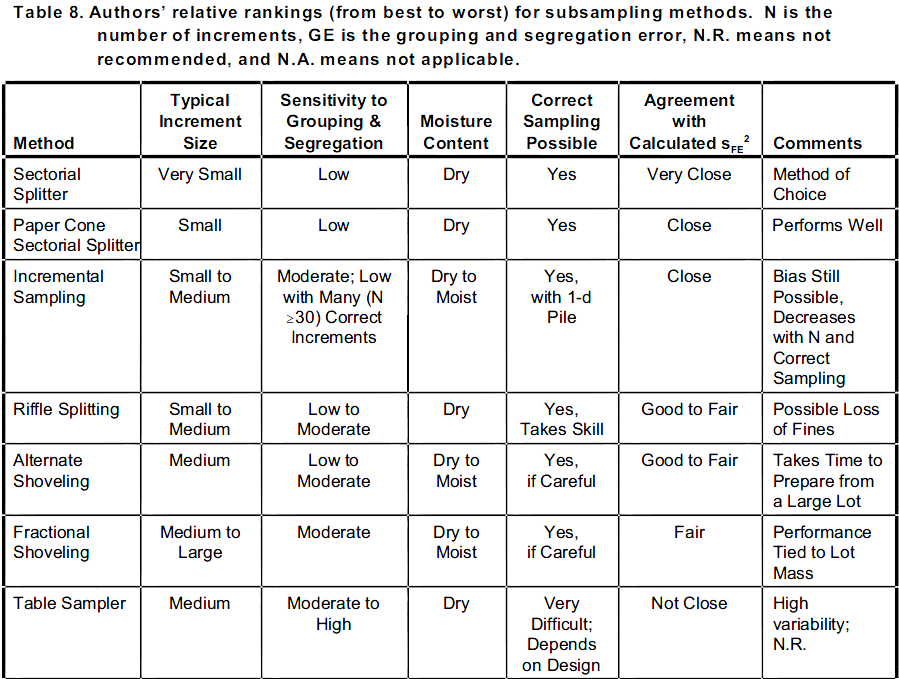
Visual Sample Plan (VSP) v. 6.5. Software is developed by the federal government and is free. Download from <http://vsp.pnnl.gov/>

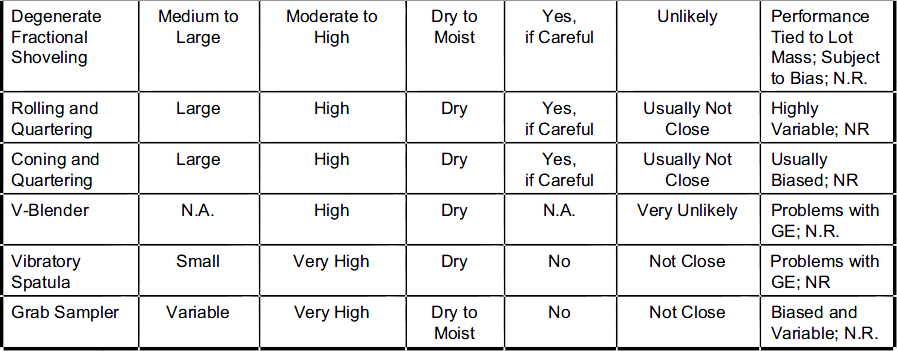
White, R.E. (2006) *Principles and Practice of Soil Science: The Soil as a Natural Resource*. 4th edition. Blackwell Publishing, Malden, MA, USA.

**Appendix A**

**Ranking of Subsampling and Mixing Methods**

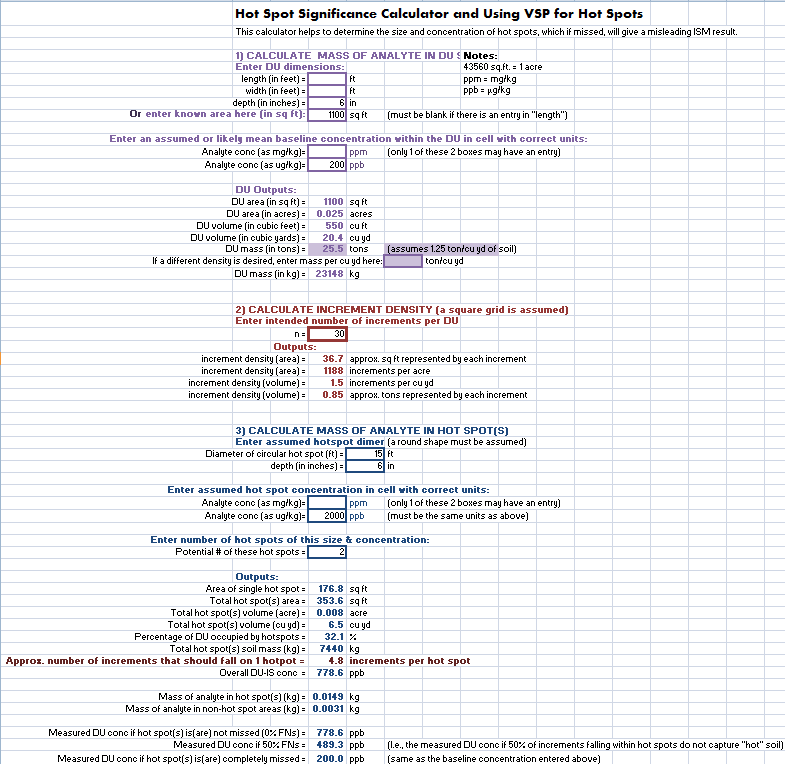
**Table from USEPA 2003, page 72.**





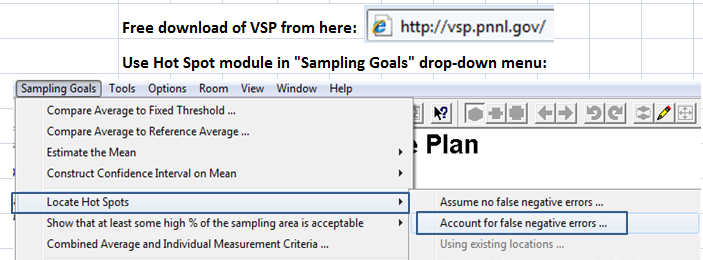
**Appendix B**

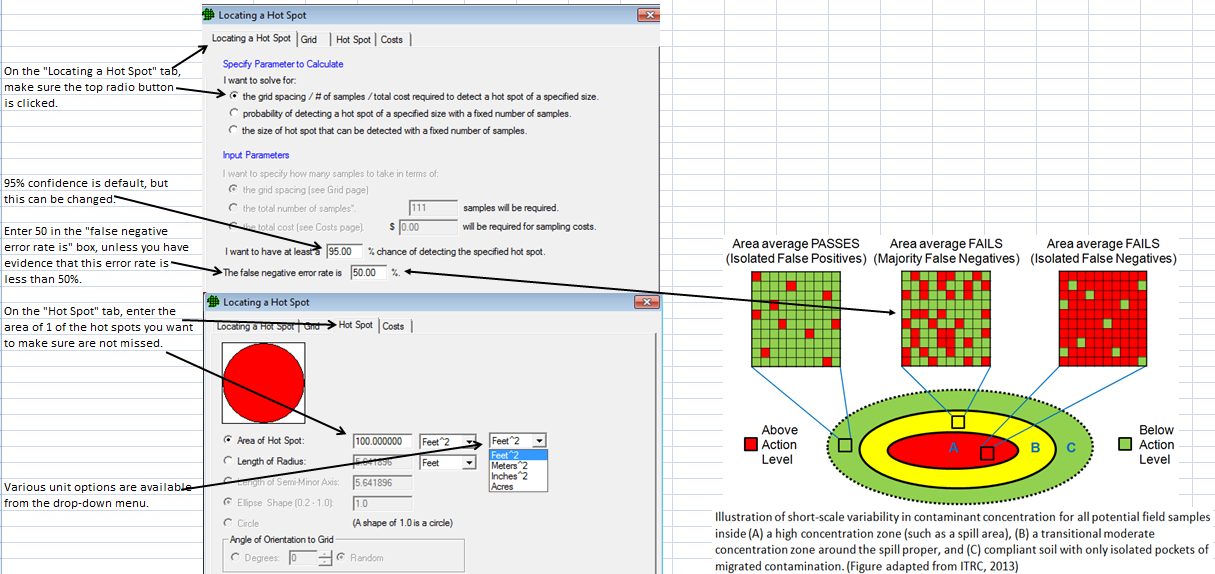
**Screen Shot of an Excel Calculator for Assessing Hot Spot Significance**

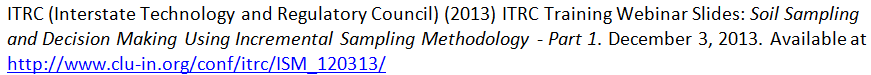


**Appendix C**

**Instructions for using VSP to determine the spacing of increments to detect a hot spot of the given area**

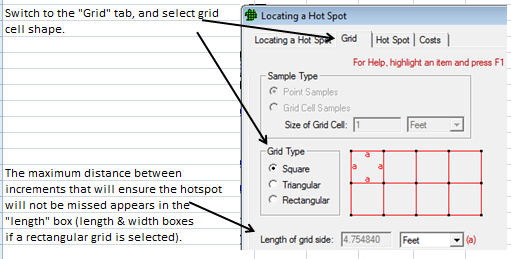
****

****

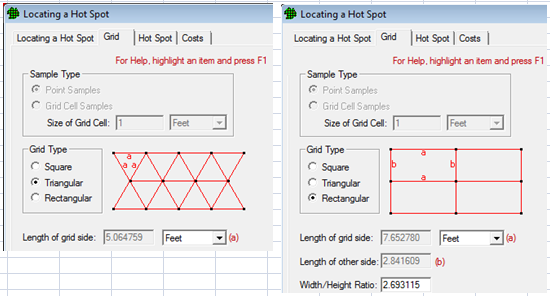
****

**Appendix C, continued**

**Instructions for using VSP to determine the spacing of increments to detect a hot spot of the given area (continued)**

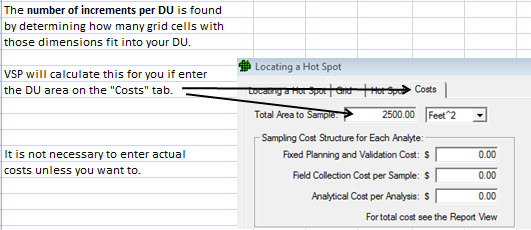
****

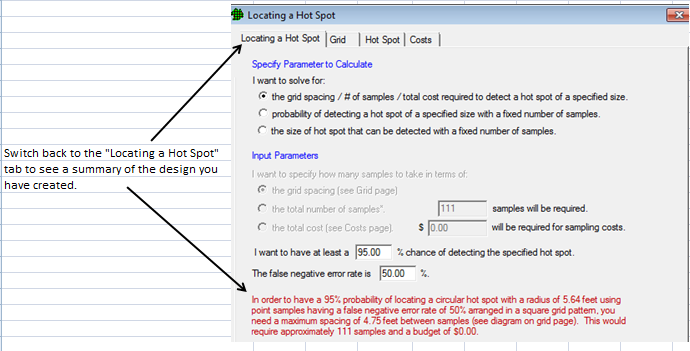
**Other options for increment spacing:**

****

**Appendix C, continued**

**Instructions for using VSP to determine the spacing of increments to detect a hot spot of the given area (continued)**

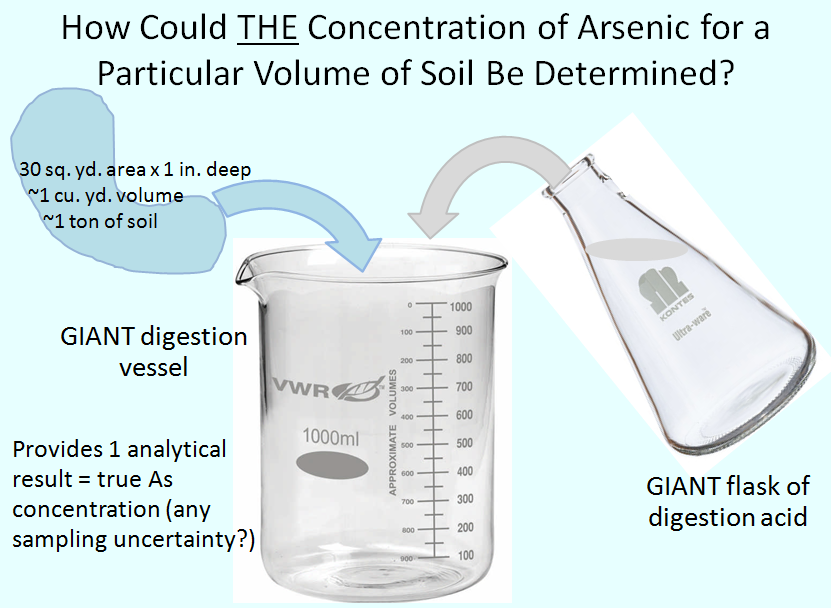
****

****

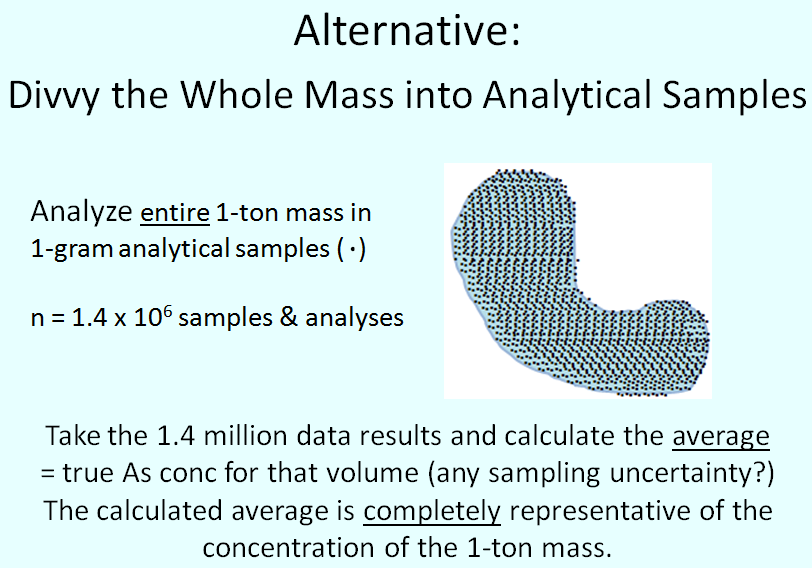
**Appendix D**

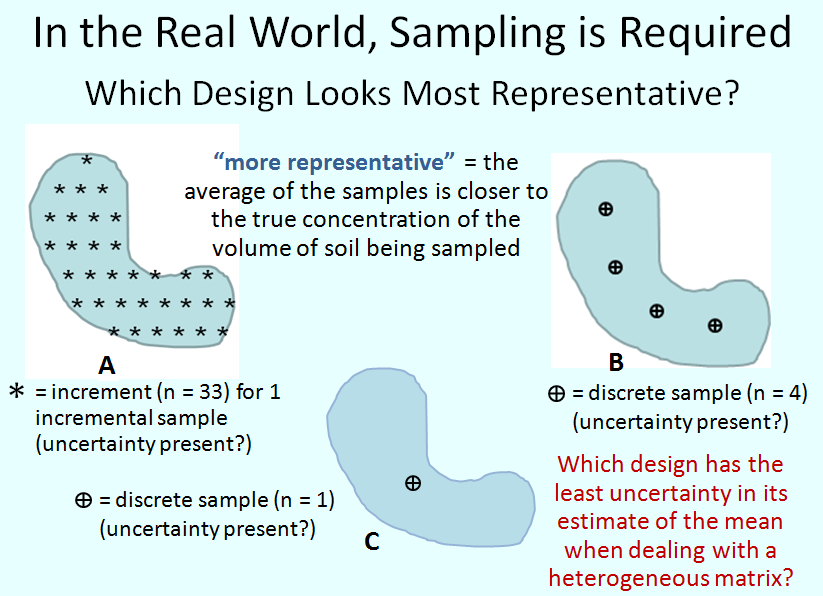
**A Thought Experiment**

What is the “True Concentration” for a Given Soil Volume (e.g., a DU)?

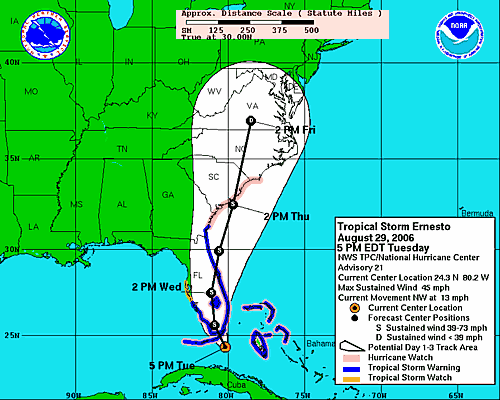


**Appendix D, continued**





**Appendix E**



**Figure E.**  Illustration of why having higher statistical confidence has larger (wider) bounds. The white “tear drop” around the central tendency track (black line) defines a fixed statistical confidence level around the “best guess” line. It might be 95% or whatever meteorologists use. The width of the white band reflects the amount of uncertainty in where the hurricane might go. Note a couple of things:

* The width of the uncertainty bound is narrow where the prediction is just a few hours ahead of where the storm is now, because there is more confidence of where the storm could track.
* The more hours out the prediction, the more uncertainty in the prediction and the wider the band.
* Imagine that the uncertainty band represents 80% confidence. If you wanted to have *more* confidence (like 95% confidence) that you have bounded the possible area where the storm might go, would you narrow the uncertainty band or widen it?

**Appendix F**

**Why is a “safety factor” used to calculate an estimate of the population mean?**

Imagine 10 soil [(physical) samples](#Glossary) of 100 grams each are collected from a soil unit that is 30 square yards in area and 1-inch deep (which is the example DU in [Appendix D](#Appendix_D)).

* The volume of this soil unit is just under 1 cubic yard, and it weighs about 1 ton (2000 pounds or 909 kilograms).
* Each 100-g sample represents 0.0001 (or 0.01%) of the entire DU mass.
* The volume of 10 100-g samples represents 0.1% of the DU mass.
* There are 9,090 100-g samples that make up the DU population of 100-g physical samples.

Our sample set (i.e., the statistical sample) represents 10/9,090 or 1/909 or about 0.001 (one-one thousandths, or 0.1%) of the population of 100-gram samples. Knowing that soil is very heterogeneous, would you expect that the average of the 10 samples (i.e., the sample mean) will be the same value as the average of all 9,090 samples (i.e., the true population mean)? Of course not! The 10-sample data set provides just one estimate of the population mean. You would expect that a different set of 10 samples will give a different estimate of the population mean.

If you wanted to guard against the possibility that your estimate of the population mean is lower than the actual value, what could you do? You could add a “safety factor” to the sample mean and use that as your estimate of the population mean. This “padded” estimate of the mean is called the upper confidence limit or UCL. So the UCL is a high-end estimate of the population mean that is based on the sample mean and some other characteristics of the data set (as discussed below in this Appendix).

Just as you can make a high-end estimate of the population mean, you can make a low-end estimate. The low-end estimate is called the lower confidence limit or LCL. Generally, it is calculated by taking the same safety factor as was used to compute the UCL, but this time subtracting it from the sample mean. This generalization will not hold true if the statistical confidence level desired for the UCL (often 95%) is different from the confidence level acceptable for the LCL. Sometimes a 90% or 80% confidence level is used if an LCL is used for environmental decision-making. The safety factor for 95% confidence is smaller than the safety factor for 90% statistical confidence.

In addition to the confidence level, characteristics of the data set used to compute the safety factor are the number of samples (n), the amount of variation among values in the data set (the standard deviation, SD), and what kind of statistical distribution the population is expected to have (usually predicted from the shape of the data set’s histogram).

Below is a graphic to demonstrate how each data characteristic either enlarges (widens) or reduces (narrows) the safety factor, more formally called the “mean-to-UCL width.”

**Appendix F, continued**

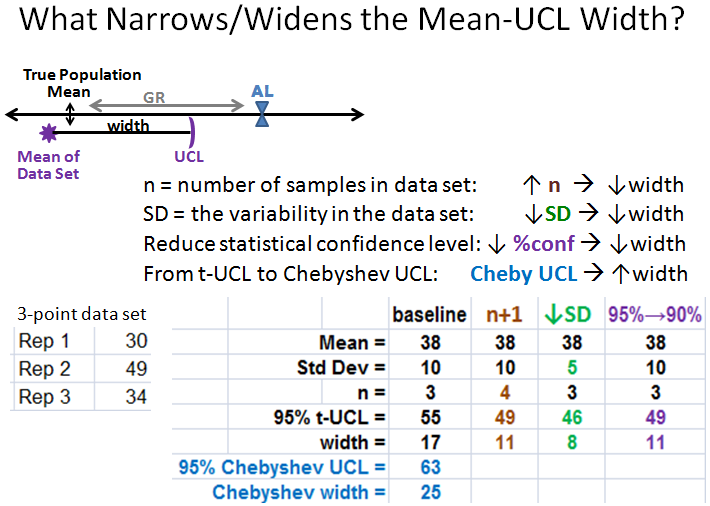


Figure F. A visual representation of a UCL for a data set in relation to its calculated mean. A UCL is calculated AFTER data are collected. A visual representation of the gray region (GR) in relation to the true mean and action level (AL). The gray region is an important concept for using statistical calculations to determine the number of samples BEFORE data are collected. It can be thought of as the numerical distance between the true mean and the AL, which means it is the interval into which a UCL can fall without exceeding the AL. “Width” is the numerical distance between the mean and the UCL.

The table in Figure F shows the effect on the UCL of changing one factor at a time from the baseline statistics for the presented 3-point data set. The changes are

* “n+1”: changes n from 3 to 4 in the equation used to calculate the UCL. The t-UCL goes from 55 to 50.
* “**↓**SD”: changes the SD from 10 to 5 in the equation used to calculate the UCL. The t-UCL goes from 55 to 46.
* “95% →90%”: reduces the confidence level a little. The t-UCL goes from 55 to 49.
* A Chebyshev UCL is higher (and thus the mean-UCL width is wider) than a t-UCL of the same confidence level.

**Appendix G**

**A Statistical Way to Assess “Do-Not-Exceed” Criteria**

The EPA guidance document, *RCRA Waste Sampling Draft Technical Guidance,* presents a statistical method to assess compliance with “not-to-exceed” concentration criteria (Section 5.5.2 and Appendix G of USEPA 2002a)

This is a non-parametric, volume-independent statistical calculation that says that if a randomly selected 59 out of those 1.4 million samples are analyzed, and all 59 results are below the action level, then there is a 95% probability that 95% of all potential samples are below the action level. However, note that 95% probability that 95% are below the action level still leaves (for the 1.4 million samples example) open the chance that somewhere around 70,000 of the 1-gram samples (5%) are greater than the action level.

If the 59 random samples are analyzed initially, but 1 of them is greater than the action level, then another 34 samples must be analyzed. If none of those samples exceed, i.e., there is still only 1 original exceedance out of a total of 93 samples, then there is 95% probability that 95% of all potential samples are below the action level.

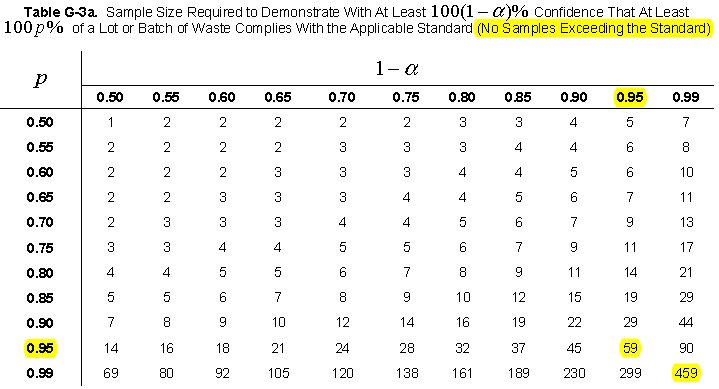
Note that the “59” number does not depend on how many samples make up the population, as long as the number is “many.” In other words, it doesn’t matter if the population is made of 1.4 million samples, or 20 million samples. As long as your first 59 samples all have results less than the action level, you can assert there is 95% probability that 95% of all potential samples are below the action level.

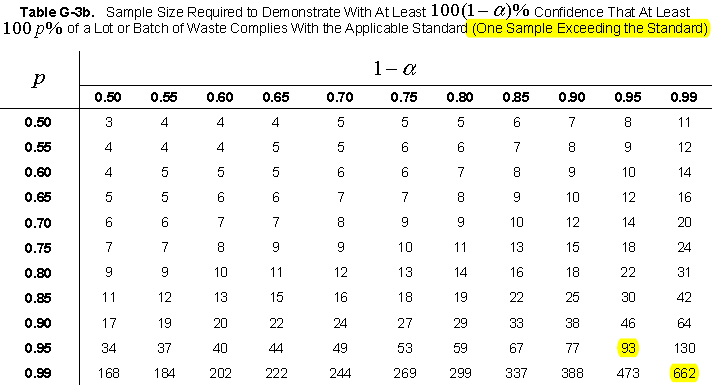
The same statistical model says that if you want 99% confidence that 99% of the 1.4 million samples are below that action level, then you must analyze at least 459 samples, and all of them must be below the action level. Note that still leaves room for 14,000 samples (14 kilograms of soil) to be greater than the action level. (USEPA 2002a, pages 83 and 266)

The bottom line is that you can never be 100% sure than all potential analytical samples are below the action level.

The two tables below were copied from the RCRA Sampling guidance referenced above.

**Appendix G, continued**





**Appendix H**

Answers to the QUIZ question posed in Section 4.4.3.1:

Since you now understand the pitfalls of sampling particulate materials, **can you explain why we see extreme high and low results in 1-g samples?**

* The **1-g samples with** **results close to zero** didn’t capture very many particles loaded with americium-241, so the “dirty-to-clean” ratio was very low (see the **blue sample** in the lower left of Figure 4.4.2.3).
* But a few of the 1-gram samples captured more than the true ratio. Combined with their small sample support (recall the calculations in Section 4.4.2.3), the higher ratio translates into **very high results**, creating the **green curve’s right-skewed tail.**

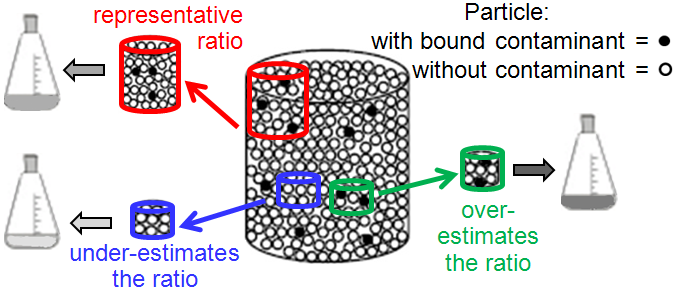


Figure 4.4.2.3

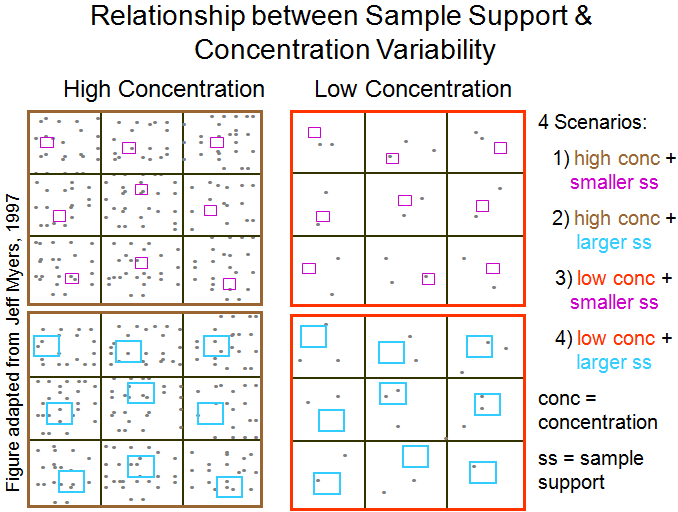


Figure 4.4.2.4a

**Appendix H continued**

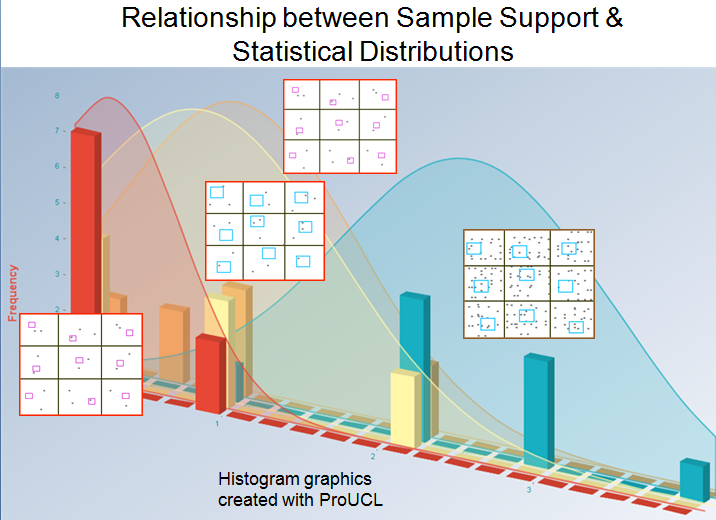


Figure 4.4.2.4b

**Appendix I**

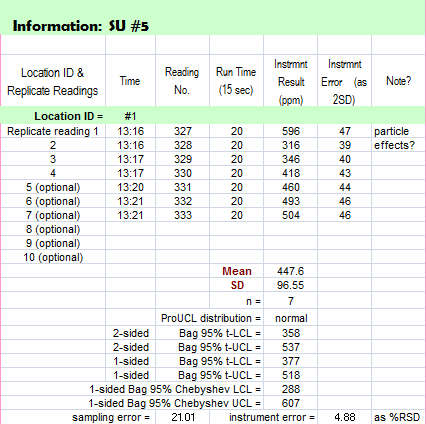
**Managing Sampling Variability Using XRF**

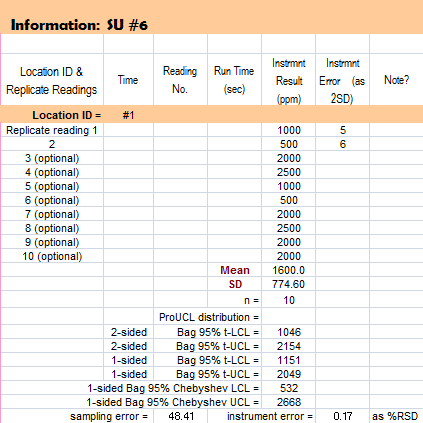
Since X-ray fluorescence (XRF) instruments require no consumables per analysis, the only cost for taking many replicate XRF readings on a single sample is time (1 to 2 minutes or less per reading). Replicate readings can be used to statistically control for sampling variability when physical control over sample heterogeneity is not as rigorous.

A pre-programmed spreadsheet to process the XRF data in real-time is highly recommended to determine with decision confidence is achieved. Screen shots of sections of such a spreadsheet appear below the text.

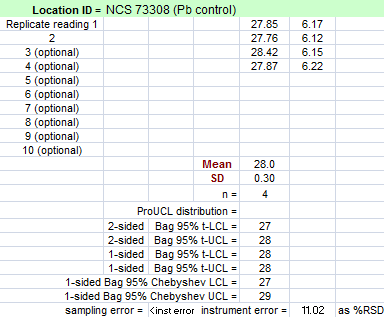
* Disaggregate a dry to mostly dry incremental soil sample.
* Sieve to target particle size.
* Place processed soil in a clear plastic bag with a lot number that has previously been checked to ensure no interference with the passage of X-rays through the plastic.
* Lay the bag flat with minimal handling (avoid particle segregation as much as possible).
* Take two widely spaced XRF shots on the top side of the bag.
* Flip the bag over and take two shots on the other side.
* Enter the XRF readings and instrument error into a pre-programmed XRF spreadsheet that automatically calculates statistics on entered readings, along with the [RSDs](#Glossary) for both instrument and sampling “errors” (variability). (The spreadsheet can be obtained from Deana Crumbling, this paper’s author.)
* If the mean is below the decision threshold, but the UCL is above, there is statistical uncertainty about whether the true mean actually exceeds the threshold. Take additional readings in the same pattern as before (2 per side while alternating sides) until a confident decision can be made.
  + The mean rises to exceed or nearly exceed the threshold, or
  + The UCL comes down under the threshold.
* If the mean is above the threshold but the LCL (lower confidence limit) is below it, there is statistical uncertainty whether the true mean is above the threshold. Take additional readings until decision confidence is achieved.
* If decision confidence cannot be reached after a total of 10 readings on the bag due to elevated sampling variability, improve sample processing procedures to reduce sampling error.

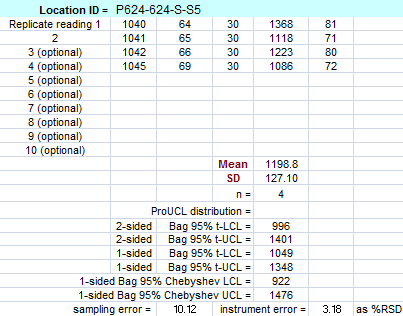
**Appendix I, continued**

****

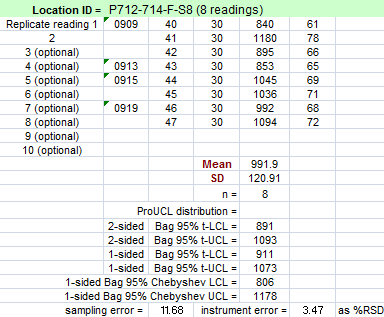
****

**Appendix I, continued**

****

****

**Appendix I, continued**

****

**Appendix J**

**Assessing Data Representativeness**

**Knowing sample representativeness is as important to correctly interpreting soil data as knowing that a samples came from the correct Site.**

“It is not possible to ascertain whether a particular sample is representative from any kind of inspection or characterization of the sample itself. Only a full qualification of the sampling process can lead to recognition of representative samples.” (Petersen et al 2005).

**The items below are part of “a full qualification of the sampling process.”**

* How many grab samples or increments were collected over what area/volume of soil in the field (i.e., the [sampling “density”](#Glossary))?
* How were the sample locations selected: randomly? An even-spaced grid? Judgmental? Something else?
* Was the sampling design statistically based? If yes, what was the statistical goal:
  + Determining the mean and [UCL](file:///C:\Users\DPOWEL02\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.Outlook\O18HYDLU\fudge#Glossary) for an area?
  + Determining a proportion?
  + Establishing compliance with a not-to-exceed criterion?
* If a form of compositing was used, how many increments were combined into a composite sample or an incremental sample?
* What was the mass of soil collected for each field sample?
* How were field samples processed prior to subsampling: Were they dried? [Disaggregated](#Glossary)? Sieved (and, if sieved, what particle size was retained for analysis)? [Ground](#Glossary) or [milled](#Glossary)?)
* What analytical mass was used, and how was the subsampling performed (e.g., grab subsampling by scooping off the top of the jar vs. incremental subsampling of a [slabcake](#Glossary))?
* What is the variability (i.e., imprecision) between replicate (i.e., collocated) field samples?
* What is the variability (i.e., imprecision) between subsampling replicates (such as laboratory duplicates)?

All of the above considerations affect sample representativeness.

* Some determine whether analyte concentrations will be tend to be high or low.
  + For example, concentrations for many contaminants tend to be much higher in fine-grain fractions than in [bulk soil](#Glossary).
* Some determine how [variable](#Glossary) the data will be, and what the likely sources of data variability are (field heterogeneity? subsampling error in the lab?).
* Any secondary use of soil data needs access to this information to determine whether the data are appropriate for that use.
* Archiving of data results in databases must include this metadata.

Knowing that the data have passed [data validation](#Glossary) is of no use if there is no documentation that explains what the data represent in terms of “[decision support](#Glossary).” Soil data can be misinterpreted because they are considered to be of “good quality” after data validation, but sample representativeness is completely unknown!

**Appendix K**

**Terminology: What is a soil “hot spot”?**

The term “hot spot” appears in some EPA and state guidance documents, often without a definition. There are definitions in at least two EPA guidance documents:

* In EPA’s 1996 Soil Screening Guidance, a hot spot is described as “a small portion of the [exposure area that] has very high levels” [USEPA (1996) Soil Screening Guidance: Technical Background Document Superfund. EPA/540/R95/128 May 1996].
* In EPA’s 2002 RCRA Waste Sampling Technical Guidance, a hot spot is a “strata that contain high concentrations of the characteristic of interest and are relatively small in size when compared with the total size of the materials being sampled (ASTM D 6009-96).” (USEPA 2002a)

However, these guidance documents do not clearly discuss how to develop the metrics of what constitutes a “hot spot” or how to include hot spot detection in up-front project planning.

**What happens when there is no project-specific definition of a hot spot?**

Hot spots can, and should, be defined according to the mass of analyte they contain as it contributes to the total mass of analyte in the DU.

* A pre-programmed spreadsheet (see [Appendix B](#Appendix_B)) can be used to perform calculations that set the minimum size (i.e., volume, but often expressed in terms of area) and concentration that would be considered a hot spot.
* Without a scenario-specific definition of what is, or is not, a hot spot, likely each party has in mind a vague, but different, impression of what they think a hot spot is, and what should be done about it.
* Because it is uncomfortable to try to communicate vague impressions among a group of people, especially people who see themselves on opposing sides, there is a strong incentive to NOT tackle hot spot definition before data are collected.
* Because of high variability, discrete sample results can raise more questions than they answer.
  + Variability can make it hard for a regulator concerned with the health of people and the environment to feel confident that “nothing has been missed” when signing off that a project is satisfactorily completed.
  + Local residents or stakeholders are often motivated by fears: of health consequences for themselves and their children, and financial consequences from declining property values and the inability to sell and move to another phase in their life. The presence of even one high result can feed those fears.
* When planning a specific project, it is time to get very specific about **what area/volume and concentration qualifies as a hot spot.**
  + Unless I know what I am looking for, I cannot design a sampling strategy to find it!!

**Appendix L**

**Why Recommend at Least 5 Increments per Composite?**

1. Determine the **theoretical** chances of a composite “catching” a hot spot

To perform this exercise, we will make a few assumptions

* Assume that within a hot spot the chance that any single increment will exceed the hot spot criterion is 50%
  + The assumed probability of exceedance could be increased or decreased from 50%, and this will change the outcome somewhat.
  + Recall the [Becker Study example](#Table_4_5_5), where “false negative results” ranged between 3 to 75 % for individual arrays, with an average of 34% for all arrays (a total area of 4 sq.ft.). So 50% is a reasonable value to use.
  + Last, but not least, 50% is easiest to use for a teaching example.
* Assume for this exercise that if a composite contains even one increment above the criterion, the composite result will be exceed the criterion. This is not an unreasonable assumption for a true hot spot area. Refer again to the [Becker Study](#Black_asterisk), where 1 high result (at 8100 ppm) pulled the average over 1000 ppm, despite the fact that 6 of the results were less than 177 ppm.
  + However, it must be noted that if the single high triplet had an average concentration of 1850 ppm or less, that 13-sq.in. area would have had an average concentration of 400 ppm or less.

If 50% of increments taken from a hot spot are false negative results, the probability that a composite taken from the hot spot will give a “true positive result” is determined by how many increments are taken. The procedure to determine the theoretical chances for a correct (true positive result) is as follows:

Note that whether an increment’s individual concentration is above or below the criterion is given by > or <. Different colors of the > or < symbols represents different increments.

* A single increment (i.e., a discrete sample) has a 50% chance of exceeding the criterion because the only two possible outcomes are. There is 1 > out of these 2 possibilities.
* A 2-increment composite has a 75% chance of exceeding the criterion because the possible combinations are as follows: and each combination is equally likely.
  + This means that 3 out of the 4 (75%) possible combinations have at least 1 > sign (which makes the composite exceed the criterion).
* A 3-increment composite has a 87.5% chance of exceeding the criterion because because the possible combinations are as follows:
  + This means that 7 out of 8 possible combinations have at least 1 > sign (7/8 = 87.5%)
* A 4-sample composite has a 93.75% chance of exceeding:  etc. has 15 out of 16 combinations with at least 1 > sign.

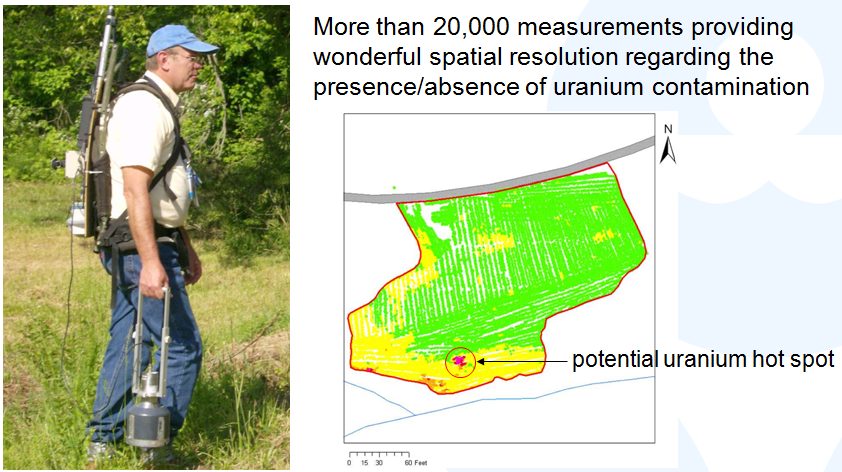
**Appendix L, continued**

* ***5-sample composite: ~97%*** theoretical possibility of exceeding the criterion ( etc. for 31 out of 32 combinations).
* If you want to calculate the theoretical possibilities for other numbers of increments (e.g., 6) in the 50% false negatives scenario, you just raise 2 to the power equal to the number of increments (26 = 64). There is only 1 combination for which all signs are <. So subtract 1 (e.g., 64 – 1 = 63). Divide that number by the total to get your percentage (63/64 = 0.984 = 98%).

2) Estimate **Actual** Hot Spot Detection Performance

The Department of Energy performed an experiment at a uranium (U) contaminated site. A 5 meter by 5 meter (25 m2) area had been identified as a hot spot, using a hot spot criterion of an average of 90 ppm U over a 25-m2 sampling unit.

The U hotspot was initially identified by a gamma walk-over survey, which uses a radiation-responsive sensor to detect and measure gamma radiation coming from the first 4 or so inches of soil when held about 4 inches above the ground surface. The lines visible in the map below show the 1-meter wide paths walked by the technician.



An area of high readings were identified as a likely hot spot of uranium contamination. Later, the same area was also identified as a hot spot by a composite sampling design that covered the entire red-outlined area of the map above.

After positive identification of the hot spot area, an experiment (described below) was performed to evaluate actual hot spot detection performance.

**Appendix L, continued**

* 10 discrete samples were collected from within the hotspot footprint.
* Results ranged from 4 to 649 ppm, with an average of 174 ppm (well above hotspot criterion of 90 ppm)
* 5 of the samples were > 90 ppm & 5 samples were < 90 ppm (= 50% hotspot detection rate)
* Then the question was asked: If the 10 samples were considered increments & randomly combined into composites using the actual concentration values, what is the probability of hotspot detection?
* The simulation was performed on the actual data using Monte Carlo analysis. The following hot spot detection rates were found:
  + 1 increment (1 of the 10 randomly selected) had a 50% detection rate
  + 2 increments (2 selected & “composited” by mathematical averaging) had a 66% detection rate
  + 3 increments (3 selected & “composited”) had a 74% detection rate
  + 4 increments (4 selected & “composited”) had a 78% detection rate
  + 5 increments (5 selected & “composited”) had an 85% detection rate.
* Note that the real world detection rate for this hot spot (85%) was less than the theoretical detection rate (97%).
* Until more experimental information becomes available, reasonably good hot spot detection rate should be achieved if no fewer than 5 increments per composite are used when hot spot detection performance is unknown.
* Naturally, using more than 5 increments provides greater confidence in hot spot detection.
* Site-specific hot spot detection rates can be determined using the DOE procedure described above. A known hot spot must be available from which to take at least 10 samples to run the statistical analysis.
  + A site-specific hot spot detection rate can then be used to set the spacing of increments using the VSP hot spot module as shown in [Appendix C](#Appendix_C).
  + If composites are being used to determine concentration trends and cleanup boundaries or perform hot spot searches, the site-specific hot spot detection rate can also be used to determine the number of increments to use per composite.