U.S. DEPARTMENT OF ENERGY

PNNL-18253

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Validation of Statistical Sampling Algorithms in Visual Sample Plan (VSP): Summary Report

LL Nuffer
LH Sego
JE Wilson
NL Hassig
BA Pulsipher
B Matzke

February 2009

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

This document was printed on recycled paper.
(9/2003)

PNNL-18253

Validation of Statistical Sampling Algorithms in Visual Sample Plan (VSP): Summary Report

February 2009

LL Nuffer BA Pulsipher
NL Hassig LH Sego
JE Wilson B Matzke

Prepared for
U.S. Department of Homeland Security
Science and Technology Directorate

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Executive Summary

The U.S. Department of Homeland Security, Office of Technology Development (OTD) contracted with a set of U.S. Department of Energy national laboratories, including Pacific Northwest National Laboratory (PNNL), to write a *Remediation Guidance for Major Airports After a Chemical Attack*. The report identifies key activities and issues that should be considered by a typical major airport following an incident involving release of a toxic chemical agent. Four experimental tasks were identified that would require further research in order to supplement the *Remediation Guidance*. One of the tasks, Task 4, *OTD Chemical Remediation Statistical Sampling Design Validation*, dealt with statistical sampling algorithm validation. This report documents the results of the sampling design validation conducted for Task 4.

In 2005, the Government Accountability Office (GAO) performed a review of the past U.S. responses to Anthrax terrorist cases. Part of the motivation for this PNNL report was a major GAO finding that there was a lack of validated sampling strategies in the U.S. response to Anthrax cases. The report (GAO 2005)[1] recommended that probability-based methods be used for sampling design in order to address confidence in the results, particularly when all sample results showed no remaining contamination. The GAO also expressed a desire that the methods be validated, which is the main purpose of this PNNL report. The objective of this study was to validate probability-based statistical sampling designs and the algorithms pertinent to within-building sampling that allow an investigator to prescribe or evaluate confidence levels of conclusions based on data collected as guided by the statistical sampling designs. Specifically, the designs found in the Visual Sample Plan (VSP) software were evaluated. VSP was used to calculate the number of samples and the sample location for a variety of sampling plans applied to an actual release site.

Most of the sampling designs validated are probability based, meaning samples are located randomly (or on a randomly placed grid) and the number of samples is calculated such that *if* the amount and spatial extent of contamination exceeds levels of concern, at least one of the samples would be taken from a contaminated area, at least X% of the time. Hence, "validation" of the statistical sampling algorithms is defined herein to mean ensuring that the "X%" (confidence) is actually met.

The validation effort focused on four VSP sampling designs based on the following sampling objectives that were deemed pertinent for sampling within a building after a chemical or biological attack.

- Upper Tolerance Limit Based Sampling – Statement that X% confident that at least Y% of surface area is below some quantitative contaminant limit where only random samples are obtained.

- Compliance Sampling – Statement that X% confident that at least Y% of surface area contains no detectable contamination where only random samples are obtained.

[1] GAO. 2005. *Anthrax Detection: Agencies Need to Validate Sampling Activities in Order to Increase Confidence in Negative Results.* GAO-05-251, General Accountability Office.

- Combined Judgment and Random Sampling – Statement that X% confident that at least Y% of surface area contains no detectable contamination where both random and judgmental samples are obtained.

- Hotspot Sampling – Statement that at least X% confident that any contaminated area greater than a given size and shape is sampled.

Validation was accomplished by first creating a "ground truth" building and data set. The ground truth building and contaminant distribution was based on data from an actual building where a simulant had been released (Coronado Club in Albuquerque, NM). Contaminant estimates were derived for each 0.3 m x 0.3 m grid within the building using geostatistical modeling methods. Contaminant action levels were then varied to produce different ground-truth scenarios to make parts of the building more or less "contaminated," thereby changing the Y% parameters mentioned above.

There are two types of decision errors that may be of concern that have pertinence relative to validating any sampling design. The first is concluding that an area is sufficiently uncontaminated when in fact it is not sufficiently uncontaminated. The second is concluding that an area is contaminated when it is in fact not contaminated. The health, political, cost, and credibility consequences of the first type of decision error are the primary concern that the GAO review outlined. Although the second type of decision error may be costly and require unnecessary actions, the health consequences are usually negligible. Thus, we have chosen to focus our validation efforts on ensuring that the VSP derived number and placement of samples for all sampling design methods is adequate to confidently conclude that an area is contaminated if in fact it is contaminated, i.e., adequately protecting against erroneously concluding that an area is uncontaminated.

The method for validating each VSP design was to apply each design to the same simulated site (in some cases, different areas of the site), taking the number of samples suggested by VSP to meet the design parameters, and using the decision rules in VSP to conclude whether or not the total decision unit is contaminated.

For example, our primary sampling goal may be to state that we are 95% (X%) confident that at least 99% (Y%) of the surface area within some airport terminal is uncontaminated if none of the samples detect contamination (Compliance Sampling, C=0). This same objective can be translated to state that we want to be at least 95% (X%) confident of detecting contamination (having at least one contaminated sample) if 1% (100-Y%) or more of the area is contaminated. The number of samples required is the same for either of these translations. Therefore, we can validate the performance of each VSP method/parameter combination by setting our ground-truth building data to have approximately 100-Y% of the grid cells to be at or above the action level, and determining the proportion of times (achieved confidence) that at least one sample is found that contains contamination. We expect this proportion to be at least X% (goal confidence).

Continuing the example above, to validate we met the 95% confidence requirement, we repeat the process of taking samples and using the sample results, over 10,000 trials, and for each trial decide whether to call the site contaminated or not. If in 9,500 or more of those trials we identified the area as contaminated (given that 1% or more of the area in the ground truth was indeed contaminated), we said we validated the 95% confidence level. That is, our achieved confidence is equal to or greater than our goal confidence.

A similar set of realistic goals and simulation trials was used to test the other designs. Numerous simulations were performed for a variety of X% parameter values while varying the ground truth Y% parameter or the hotspot size and shape. The "achieved confidence" was compared to the "goal confidence." **The results from the simulations did indeed validate the selected VSP sampling designs.** This means that the algorithms within VSP that calculate sample size, sample location, and the conclusions from statistical tests provided the information we expected and achieved the goal confidence levels (to within acceptable tolerances).

A few of the comparisons between achieved and goal confidences derived from some of the simulation runs are shown below to illustrate the type of results obtained. In Figure E.1, the graphs show some results for the Upper Tolerance Limit based designs, with Non-Parametric Design results on the left and Parametric Design results on the right. For various goal confidences (90%, 95%, 99%) and percent clean ground-truths (90%, 95%, 99%), the achieved confidence based on the 10,000 VSP trials is graphed. All results show that the achieved confidence exceeds or meets the goal confidence. In Figure E.2, graphs are shown for some of the compliance sampling design simulations. Again, all results show that the achieved confidence exceeds or is very close to the goal confidence. In cases where the achieved confidence is slightly lower than the goal confidences, such occurrences can be explained by our inability to precisely control the underlying ground-truth data set or inherent simulation variations.

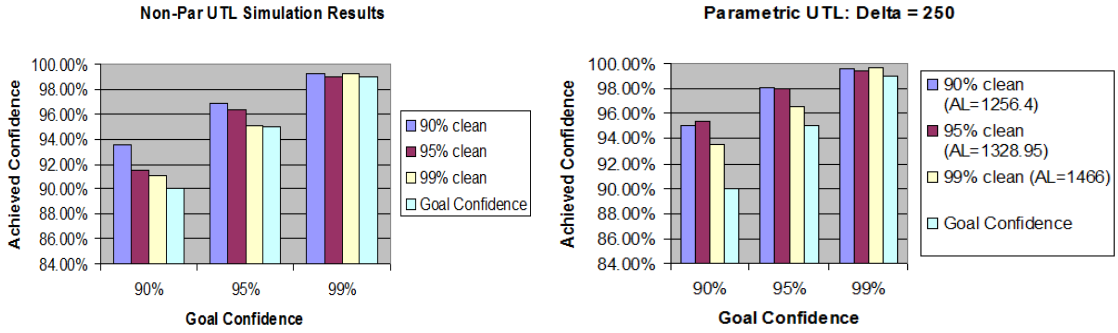


Figure E.1. Comparison Between Achieved and Goal Confidences for UTL Based Designs

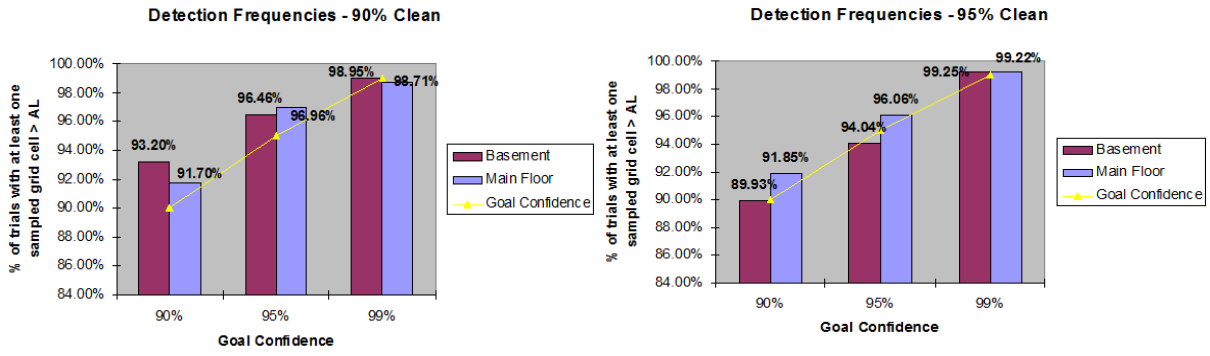


Figure E.2. Comparison Between Achieved and Goal Confidences for Compliance Sampling

Other results for these and other VSP designs are included in the body and appendices of this report. The methods and computer code used for the simulations, the ground-truth data sets, and the procedures for making decisions are available for validation of sampling designs other than those used in this

validation effort, and for sampling designs not currently in VSP. Thus, alternative sampling designs can be tested and validated in follow-on efforts using the output and software products created in this project.

Contents

Executive Summary iii
1.0 Background 1.1
2.0 3-Tiered Approach to Validation for VSP Software 2.1
2.1 Goal of Validation 2.3
2.2 Metrics for Validation 2.3
3.0 Ground Truth Used in Simulations 3.1
4.0 Sampling Plans Validated 4.1
4.1 Hot Spot Sampling Design 4.1
4.2 Parametric UTL Sampling Design 4.2
4.3 Non-Parametric UTL Sampling Design 4.2
4.4 Compliance Sampling 4.2
4.5 Combined Judgment and Random Sampling 4.3
5.0 Methods of Validation 5.1
5.1 Method of Validation for Hot Spot, UTL, and Compliance Sampling 5.1
5.2 Method of Validation for CJR 5.2
6.0 Simulations used for Hot Spot, UTL, and Compliance Sampling 6.1
7.0 Results 7.1
7.1 Compliance Sampling 7.1
7.1.1 Ground Truth Matches Design Acceptable Percent Clean 7.1
7.1.2 Ground Truth Does not Match Design Acceptable Percent Clean 7.3
7.2 UTL 7.5
7.2.1 Non-Parametric Case 7.5
7.2.2 Parametric Case 7.7
7.3 Hot Spot 7.8
7.4 CJR 7.10
8.0 Conclusions 8.1
9.0 References 9.1
Appendix A – Hot Spot Sampling Validation: Details on Sampling Design Parameters and Results of Simulation A.1
Appendix B – UTL Sampling Validation: Details on Sampling Design Parameters and Results of Simulation B.1
Appendix C – Compliance Sampling Validation: Details on Sampling Design Parameters and Results of Simulation C.1
Appendix D – Combined Judgment and Random (CJR) Sampling Validation: Detailed Description of the CJR Sampling Design and Simulation Results D.1

Figures

E.1 Comparison Between Achieved and Goal Confidences for UTL Based Designs v

E.2 Comparison Between Achieved and Goal Confidences for Compliance Sampling v

2.1 3-Tiered Approach to Validated Sampling Strategy 2.1

3.1 Color-Coded Map of the Main Floor of the Coronado Club, Based on Kriged 0.3-m x 0.3-m Grid Cells 3.2

3.2 Color-Coded Map of the Basement of the Coronado Club, Based on Kriged .3m x .3m Grid Cells 3.3

3.3 Basement of Coronado Club with Nine Hot Spots Labeled 3.4

3.4 Color-Coded Map of Concentration Values of Hot Spot 8 in Room 24 in the Basement of the Coronado Club 3.6

3.5 Hot Spot 8 in Room 24 Coded to Red/Green to Show Grid Cells with Concentrations Values $<275\ \mu g/m^2$ and $\geq 275\ \mu g/m^2$ 3.7

7.1 Compliance Sampling – Detection Frequencies 90% Clean 7.1

7.2 Compliance Sampling – Detection Frequencies 95% Clean 7.2

7.3 Compliance Sampling – Detection Frequencies 99% Clean 7.3

7.4 Mismatch Compliance Sampling – Detection Frequencies Requiring 90% Clean 7.4

7.5 Mismatch Compliance Sampling – Detection Frequencies Requiring 95% Clean 7.4

7.6 Mismatch Compliance Sampling – Detection Frequencies Requiring 99% Clean 7.5

7.7 Non-Parametric UTL Test with Original Data, Floor Only, Basement Coronado Club 7.6

7.8 Parametric UTL Test with Normalized Data, Floor Only, Basement Coronado Club 7.7

7.9 Hot Spot 8 Samples in Room 24 ..... 7.9

7.10 Goal versus Achieved Confidence for the Cases That Required Random Sampling 7.11

Tables

3.1 Listing of all Nine Hot Spots in Basement of Coronado Club 3.5

3.2 Summary Table of Concentration Values in the Ground Truth[(a)] 3.8

3.3 Matrix of Sampling Designs Applied to Ground-Truth Data Sets 3.9

7.1 Simulation Results for Hot Spot 8. ..... 7.9

7.2 Description of Input Parameter Values Used to Generate the 64,561 Cases in the CJR Validation 7.10

8.1 Summary of the Sampling Designs, Sampling Goals, and Location of Contamination in the Coronado Club 8.3

1.0 Background

The U.S. Department of Homeland Security, Office of Technology Development (OTD) contracted with a set of U.S. Department of Energy national laboratories, including the Pacific Northwest National Laboratory (PNNL), to write a *Remediation Guidance for Major Airports After a Chemical Attack*. The report identifies key activities and issues that must be considered by a typical major airport following an incident involving release of a toxic chemical agent. In writing this document, four experimental tasks were identified that would require further research in order to supplement the *Remediation Guidance*. One of the tasks, Task 4, *OTD Chemical Remediation Statistical Sampling Design Validation*, dealt with statistical sampling algorithm validation.

Sampling is a critical aspect of restoration. Some possible sampling strategies require a very large number of samples to both characterize the contamination in a facility and verify that a facility has been remediated following decontamination. One way to decrease the time to restore a facility following a chemical agent release is to reduce the number of required samples. The use of statistical sampling methods has the potential to do this. However, these methods have not been validated for the interior restoration problem. The objective of this study is to validate potential statistical sampling algorithms against data from actual release sites. Probability-based statistical sampling algorithms that allow the user to prescribe or evaluate confidence levels are of particular interest.

Statistical theory is the basis for the formulas for calculating sample size, and the rationale for locating samples in the designs validated in this report. All the designs validated are probability based, meaning most, if not all samples are located randomly so and the number of samples is calculated such that *if* contamination is present at the level of concern, at least one of the samples would hit the contamination, at least X% of the time. Hence, "validation" of the statistical sampling algorithms means ensuring that the "X%" is actually met. This is done by creating a "ground truth" – where a 2-dimensional grid of contamination values represents a site with a known amount and pattern of contamination. A sample design is overlaid on the ground-truth grid data, and random locations on the grid are "sampled." These values are used in a statistical test, the conclusion from which, states whether contamination has been found above the target level, and with what confidence we can conclude this. For this validation effort, only certain designs, and only certain target contamination levels and levels of confidence were tested. This is described below.

This report describes only PNNL's effort to validate the sampling designs and algorithms in the software Visual Sample Plan.[1] In the future, other reports may be written by Sandia National Laboratories and Lawrence Livermore National Laboratory to discuss the validation they did that was part of the overall Task 4.

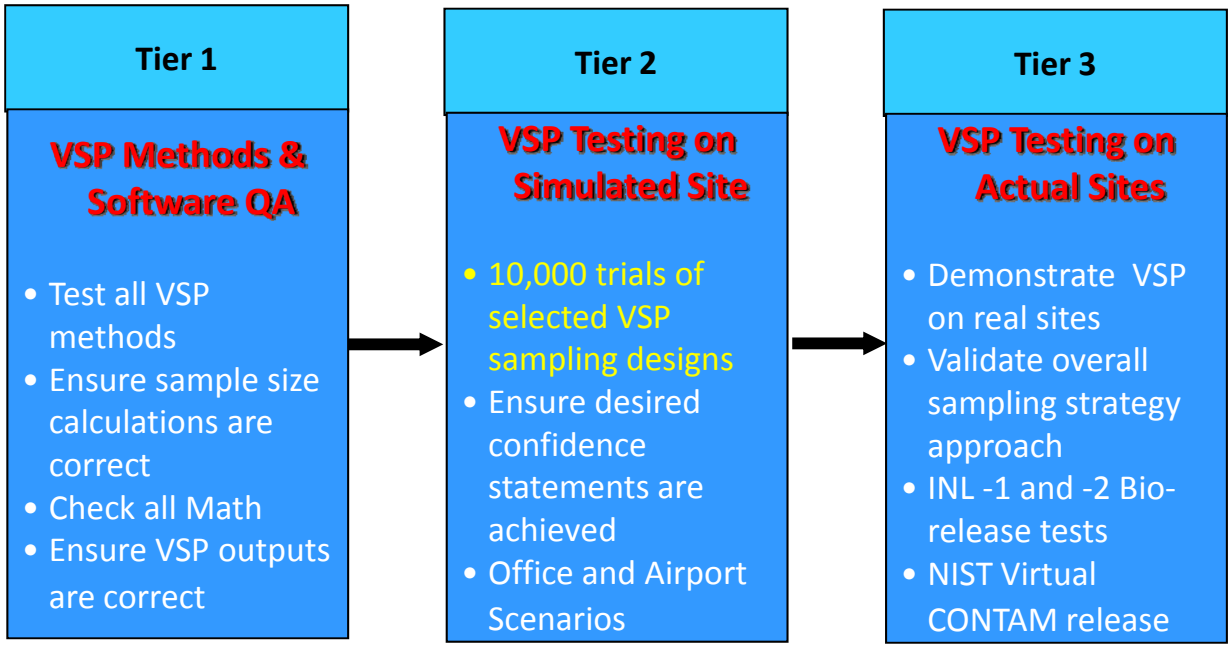
[1] Visual Sample Plan (VSP) is a software tool for selecting the right number and location of environmental samples so that the results of statistical tests performed on the data collected via the sampling plan have the required confidence for decision making. VSP User's Guide, Matzke et al. (2007).

2.0 3-Tiered Approach to Validation for VSP Software

Validation is part of a 3-tiered approach to overall quality control (QC) of sampling designs taken by PNNL in support of the VSP software. The three tiers of QC and validation are

1. Verify that VSP sample size and placement algorithms are based on sound, documented statistical theory and that the calculations performed in VSP are correct by testing against calculations performed independently
2. Validate using ground-truth data and by running simulations of sampling against the ground truth that designed confidence levels are in fact achieved, and
3. Demonstrate that the sampling designs and statistical tests perform as expected at actual sites.

The three tiers are shown in Figure 2.1. The simulations using ground-truth data described in this report comprise the bulk of tier 2.



(Notes: CONTAM is PNNL-designed software for modeling contaminant release. NIST = National Institute of Science and Technology, INL = Idaho National Laboratory)

Figure 2.1. 3-Tiered Approach to Validated Sampling Strategy

The tier 2 validation discussed in this report was a simulation study, in which a variety of potential sampling design methods are run against one or more simulated "ground truths" derived from real data from one or more actual release sites.

Part of the motivation for the tier 2 validation came from a report from the General Accountability Office (GAO). The report (GAO 2005)[1] recommended that probability-based methods be used for sampling design in order to address confidence in the results. The GAO also expressed a desire that the methods be validated, which is the main purpose of this work.

The other two tiers of the approach taken to validation are:

- **Tier 1** – Quality assurance (QA) reports and QA testing of VSP – The algorithms and methods for sampling design and analysis have gone through extensive QA/QC testing. An internal, living QA document that describes the extensive testing of the codes and algorithms in VSP is being maintained by PNNL. Hand calculations or calculations performed on independent software (not VSP) have been performed to verify correctness of the VSP calculations. Parameter ranges were varied to explore the full range of possibilities. A formal VSP QA[2] plan has been developed and implemented which documents the QA testing that should be performed with each new VSP module. In all tests performed to date, the VSP methods have performed correctly.
- **Tier 3** – The third tier for validation of VSP is demonstration of the tools within VSP at actual sites. Since VSP has been distributed for over 10 years and is widely used across the United States and abroad, there are many demonstrations of the effectiveness of VSP in numerous applications. Particularly, the within building modules are being used at many Department of Energy sites and other sites around the world for decontamination and decommissioning projects. In 2007, DHS, EPA, and DoD embarked on a joint demonstration of a within-building bio release response and restoration using a facility at the Idaho National Laboratory (INL). VSP was used to determine the number and placement of samples required for characterization and clearance phases.

The PNNL report (Amidan et al. 2007) describes the test events and numbers of samples comprising an experimental design developed to assess sampling strategies and methods for detecting contamination in a building and releasing the building for use after decontamination. INL identified Building PBF-632 as a test bed facility for evaluating protocols for response to potential contamination by biological agents. The contamination, sampling, decontamination, and re-sampling occurred as specified by the experimental design. This study is referred to as the INL Sample Collection Operational Test.[3]

Two objectives were developed to guide the construction of the experimental design for the INL Sample Collection Operational Test. The first objective was to assess the relative abilities of judgmental and probabilistic sampling strategies to detect contamination (or the extent of contamination) in individual rooms or on a whole floor of the INL PBF-632 building. The hot spot sampling design is suited for this phase of characterization. The second objective was to assess the use of traditional probabilistic sampling strategies (acceptance sampling and upper tolerance limit sampling designs) and a

[1] GAO. 2005. *Anthrax Detection: Agencies Need to Validate Sampling Activities in Order to Increase Confidence in Negative Results.* GAO-05-251, General Accountability Office.
[2] Pulsipher BA, J Wilson, and L Nuffer. 2007. *Quality Assurance Plan for Visual Sample Plan (VSP) Development.* Internal PNNL document, Pacific Northwest National Laboratory. Richland, Washington.
[3] Evaluation Report (Draft), September 2007: Indoor Field Evaluation of Sample Collection Methods and Strategies at Idaho National Laboratory, May 2008. INL Draft Report 041608. Note this report is not available for distribution.

Bayesian sampling strategy (combined judgment and random [CJR] sampling design) to make clearance statements of the form "we have X% confidence that at least Y% of a room (or floor of the building) does not contain detectable contamination."

Judgmental and probabilistic samples were collected according to the pre-specified sampling plan. Judgmental samples were selected based on professional judgment and prior information. Probabilistic samples were selected with a random aspect and in sufficient numbers to provide desired confidence for detecting contamination or clearing.

2.1 Goal of Validation

The goal of this tier-2 validation exercise, as part of the Task 4 described above, is to demonstrate, with simulated data created from an actual contaminated site, that using a statistical sampling plan, the confidence levels of the plan are indeed achieved if the VSP-recommended number of samples, and sampling locations, are followed. For example, if we use a sampling design that purports to find hot spots of a certain size, say 95% of the time, then in 95% of our simulation trials, we should find hot spots of that size or larger. In the terminology of this paper, we say we are comparing the achieved confidence from the simulated trials to the goal confidence from the VSP sampling plan design. The achieved confidence is the percent of time, in say 10,000 simulated trials, we indeed find contamination when it is there at levels of concern specified in the design.

The items to be validated are the number of samples VSP tells you to take, the placement of samples (random vs. judgment locations), the algorithms used in the VSP test to conclude contaminated vs. not contaminated, and the level of confidence assigned to the conclusion. VSP has formulas and/or algorithms for these items. Hence, we are validating the algorithms in VSP so that when you take the number of samples VSP tells you to take, you can indeed draw the conclusions VSP tells you to conclude, at the confidence levels you told VSP you wanted.

2.2 Metrics for Validation

The metrics for the validation performed by PNNL are the following:

- *Areas of Contamination of Concern*. Did we find contamination when it was present? If we specify we want to find hot spots of a certain minimum size, did we find them? Alternatively, if the goal is stated as a X%/Y% tolerance limit, can we find areas that have Y% or more of the surface area contaminated?
- *Goal Confidence*. Are we X% confident that we can find the areas of concern?

The above two metrics assume we use VSP to calculate the number of samples taken, the placement of samples, and apply the appropriate decision rules.

3.0 Ground Truth Used in Simulations

The ground truth we used in the simulations was a set of data that was a combination of actual sample results from a site where a biological contaminant stimulant was released, and supplemented kriged data to fill out the data set to represent a fine-mesh grid of contamination. The final data set of ground-truth contamination had 37,168 floor, wall, and ceiling 0.3-m x 0.3-m grid cells in the basement, and 38,531 floor, wall, and ceiling grid cells on the main floor of a building. Both floors were divided into rooms; 16 rooms on the main floor, and 21 rooms in the basement. Of the total of 37 rooms, only 34 showed any level of contamination, so only 34 rooms were considered in our simulations. The actual data showed how the walls of the rooms acted as barriers, or not, in some cases.

The following is the background on how the final data set of ground truth was created.

In February 2005, a joint exercise involving Sandia National Laboratories (SNL) and the National Institute for Occupational Safety and Health (NIOSH) was conducted in Albuquerque, New Mexico. See Griffith et al. (2006) for complete details of the exercise. The exercise was held at an SNL facility, the Coronado Club (CC), a now-closed social club for Sandia employees located on Kirtland Air Force Base.

In this exercise, a fluorescent-tagged tracer aerosol (Visolite) was used as a bioaerosol simulant. The median particle diameter of the tracer aerosol was on the order of a micrometer, which is roughly comparable to a bacterial spore. There were two basement release locations for the stimulant. Air conditioning systems were run during the release. The aerosol quickly spread from the basement to the main floor. Two floors of the building (main floor, basement) were sampled and the analytical results were used to generate a detailed contamination distribution map for the facility. Mostly surface (floor) samples were collected at sample locations based on expert judgment. Approximately 600 samples were collected and quantitatively analyzed, mostly wipes (30 cm x 30 cm), but some swab (5 cm x 5 cm) and vacuum samples (100 cm x 100 cm) were also taken. The 600 actual sample locations were not collected at regular-spaced locations.

Since the samples were taken in the most likely places for surface deposition, the actual data needed to be supplemented, so a complete picture of the building could be constructed. This was done using geostatistical kriging estimation methods, where the demonstrated spatial correlation found between actual samples was extended to unsampled areas. Estimated grid data, what we call our "ground truth," was generated for 30-cm x 30-cm (1-ft x 1-ft) grid cells, covering the entire floor space of the facility, both the main floor and basement. There were approximately 9,000 data points generated using kriging estimation methods. Sean McKenna, Sandia National Laboratories, created the kriged data set and provided it to the PNNL project staff who input it into VSP.[1] Figures 3.1 and 3.2 show the two floors of contamination that was the floor ground truth used in the simulations.

[1] Kriged data set provided to PNNL by Sean McKenna, Geohydrology Department, Sandia National Laboratories, PO Box 5800 MS 0735, Albuquerque NM 87185-0735. CClubGroundTruth.xls, email Sean McKenna to Nancy Hassig, 9/21/06.

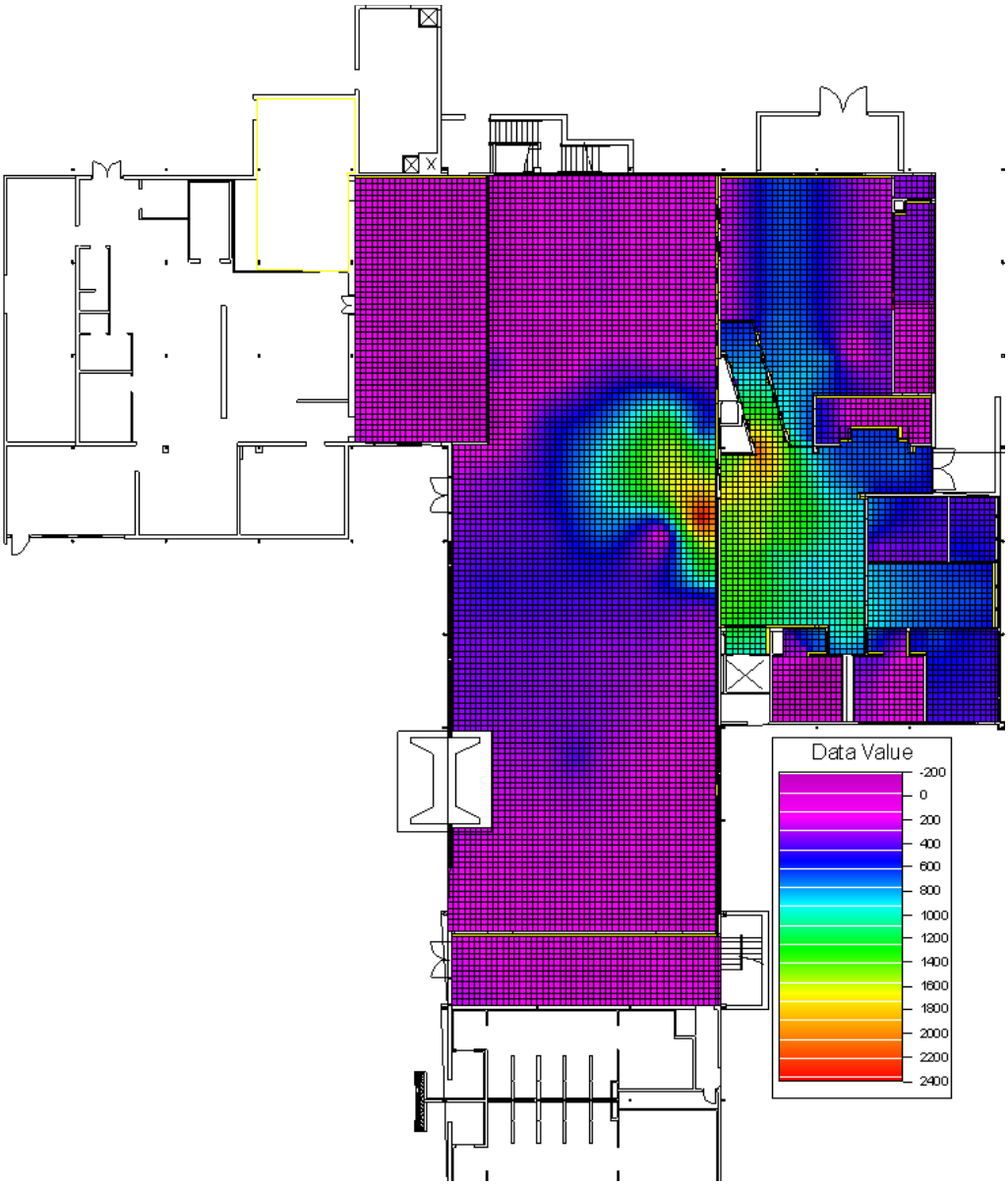


Figure 3.1. Color-Coded Map of the Main Floor of the Coronado Club, Based on Kriged 0.3-m x 0.3-m Grid Cells. Red areas represent highest concentrations. Maximum concentration grid cell value = 2,349 µg/m^2.

Pacific Northwest National Laboratory (PNNL) VSP programmers took this kriged data and imported it into VSP. VSP grid cells of the same size were generated, 0.3 m x 0.3 m (1 ft x 1 ft), and assigned values according to the original kriged data. Figure 3.1 shows the floor surface of the main floor of the CC with ground-truth grid cells colored according to a log scale of the measurement values. The rooms that are not colored did not have data values assigned to them. The dense mesh grid seen in Figure 3.1 shows the 0.3 m x 0.3 m cells. Figure 3.2 shows the floor of the basement of the CC with ground-truth grid cells colored according to a log scale of the measurement values. Most of the contamination was in the basement since the release of the stimulant contaminant was in the basement. The basement was divided into rooms. There were 34 rooms in all in the basement. Basement rooms 15 and 25 did not have data values assigned to them.

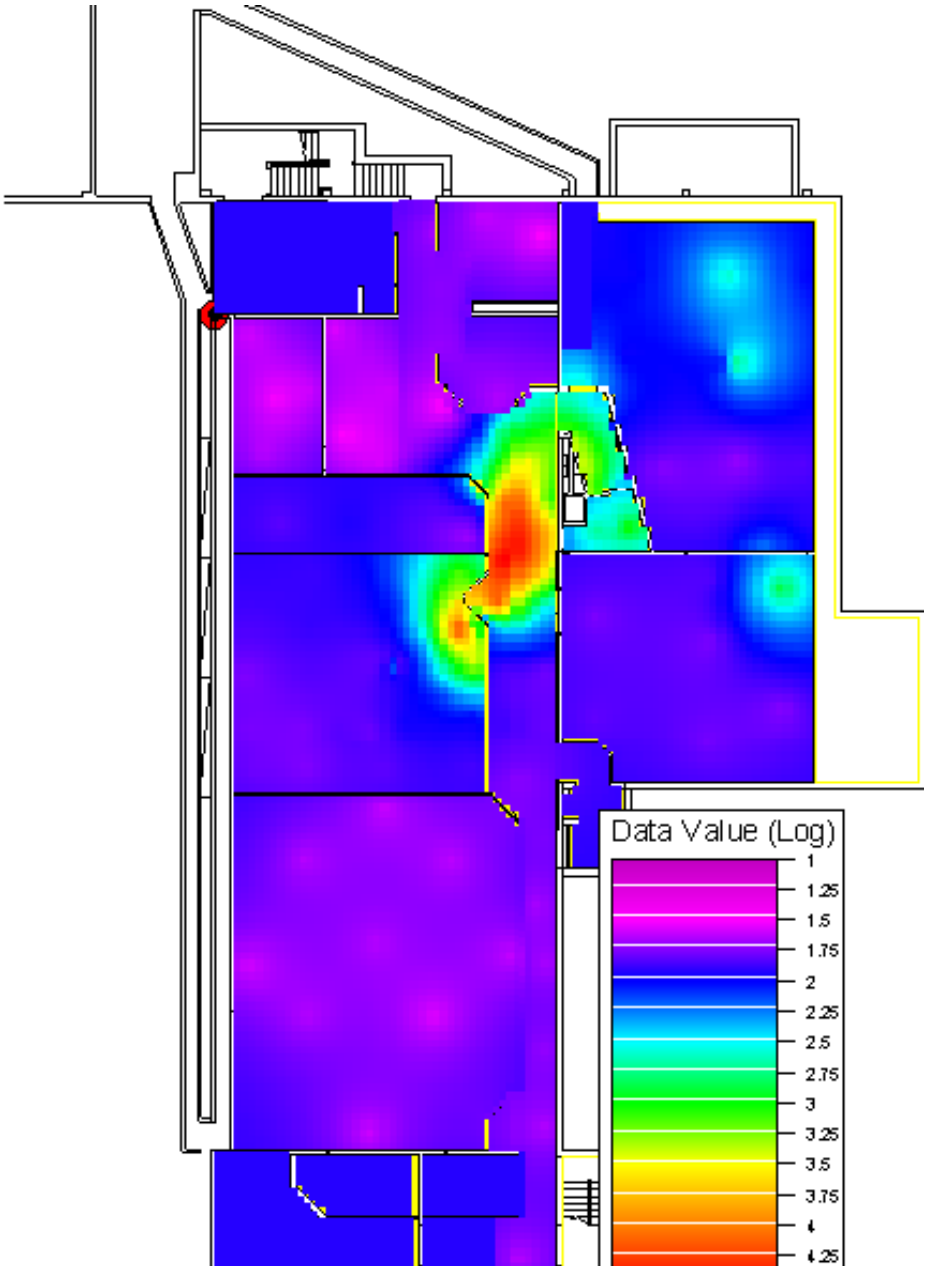


Figure 3.2. Color-Coded Map of the Basement of the Coronado Club, Based on Kriged .3m x .3m Grid Cells. Data converted to log scale so gradients of contamination visible. Maximum concentration grid cell value = 31,315 µg/m^2.

Nine "hot spots" were defined by setting the action level to 275 µg/m^2. They are labeled HS-1 through HS-9. Hot spots were defined as contiguous grid cells with concentrations ≥275 µg/m^2. All the hot spots were in the basement where the release occurred. There was wide variation in the concentration values in the basement grid cells. The release occurred in the hall and was spread by the HVAC equipment. Figure 3.3 shows the location of the nine hot spots in the basement. Since the CC basement had separate rooms, these hot spots were often naturally contained within rooms, and spread up the walls and onto the ceiling. Some of the hot spots were large, over 30 ft in diameter.

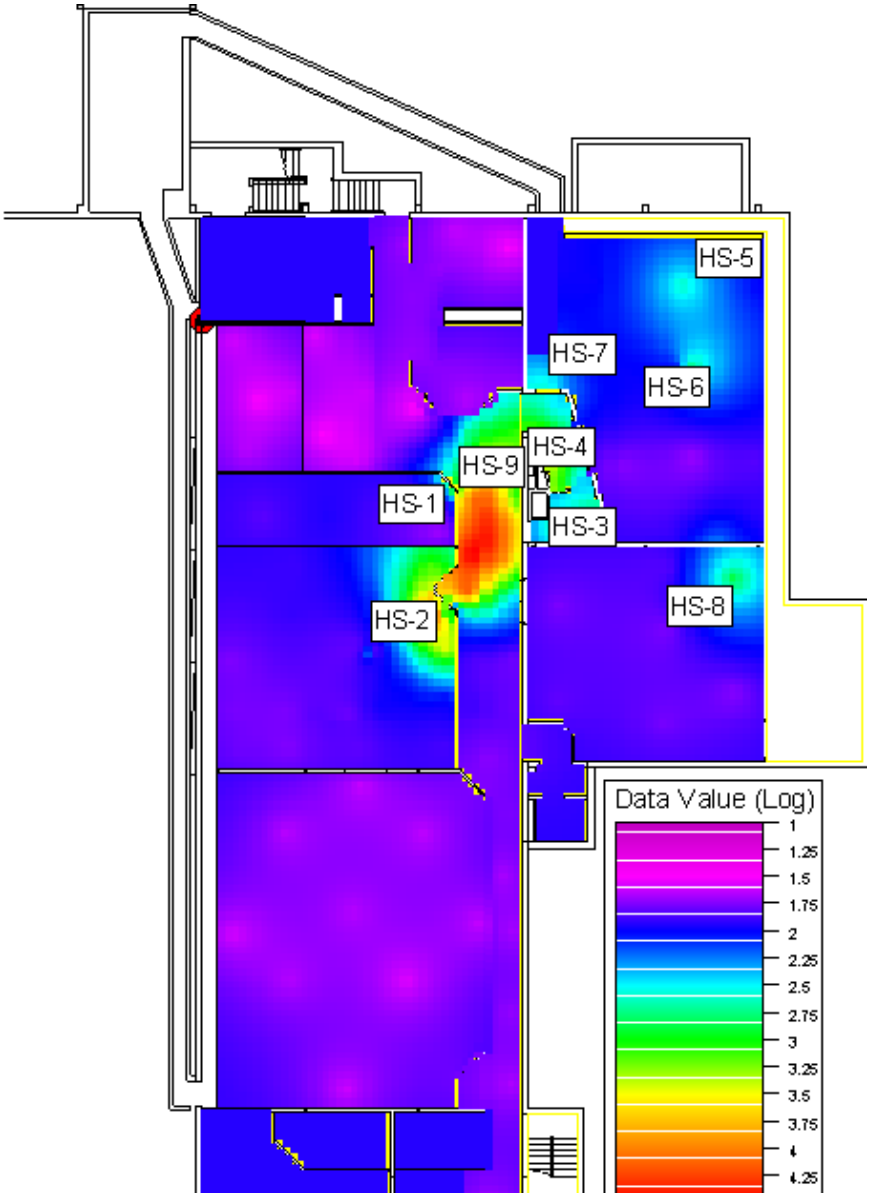


Figure 3.3. Basement of Coronado Club with Nine Hot Spots Labeled. Data converted to log scale so gradients of contamination are visible.

Table 3.1 is a list of the nine hot spots in the basement of the Coronado Club. Shown are the room number where the hot spot is located, the number of 0.3 m x 0.3 m VSP-defined cells in the hot spot, the total surface area covered by the hot spot, and the length of the major and minor axes of the hot spot. This table gives a sense for how much area each hot spot covered.

To give an idea of how hot spots appear in a 3-dimensional map-view, Figure 3.4 is a color-coded picture of Hot Spot 8. The VSP view, splayed room view, shows how it spreads over the floor, up the walls, and to the ceiling. Hot Spot 8 was in the room labeled Room 24. Figure 3.5 is a floor-only view of Hot Spot 8, showing cells with concentrations <275 µg/m^2 (color coded to green), and cells with concentration >275 µg/m^2 (color coded to red) the action level for defining a hot spot.

Table 3.1. Listing of all Nine Hot Spots in Basement of Coronado Club

Name	Room	Description	Total Number of Cells	Surface Area	Approximate Semi-Major/ Semi-Minor Axis
HS-1	Room 8	Mostly wall cells; floor cells appear to be a result of "bleed-through" from hot spot in adjacent room	24	2.23 m^2	1.3 m/0.3 m
HS-2	Room 9	Large fairly contiguous hot spot extending onto floor, some walls, and ceiling	485	45.06 m^2	3.9 m/3.75 m
HS-3	Room 17	Covering most of floor, spreads onto walls	152	14.12 m^2	1.6 m/1.6 m
HS-4	Room 18	Irregularly shaped covering almost entire floor, lots on walls, and some on ceiling	351	32.61 m^2	2.4 m/1.5 m
HS-5	Room 23	Northernmost of the two round spots	20	1.86 m^2	0.75 m/0.6 m
HS-6	Room 23	Southernmost of the two round spots	18	1.67 m^2	0.75 m/0.45 m
HS-7	Room 23	Smaller spot on the wall, result of "bleed-through" from release hot spot in hallway	28	2.60 m^2	1.0 m/0.9 m
HS-8	Room 24	Hot spot being designed for	41	3.81 m^2	1.28m/1.28 m
HS-9	Room 35	Large main release hot spot, irregularly shaped but contiguous enough to call a single hot spot	1007	93.55 m^2	4.5 m/4.5 m

The hot spots are on the floors, walls, and ceiling to create a more realistic, 3-dimenstional ground truth. The PNNL VSP team created this 3-dimensional component of the ground truth. The team started with the data set that had approximately 9,000 data points from the main floor and basement, floors only. Actual samples taken at the CC were floor surface samples. The formula used to spread the contamination up the walls and to the ceiling was: for each foot above the floor, on the wall that had a junction with the floor, decrease the contamination by 10% per foot. The limit was reached after 5 ft. Hence, the ceiling directly above the wall that had a junction with the floor had 50% of the floor contamination. Contamination was not extended along the ceiling; it was recorded just for the grid cells that joined the wall. The result was a total of 37,168 VSP floor, wall, and ceiling grid cells in the basement, and 38,531 floor, wall, and ceiling VSP grid cells on the main floor.

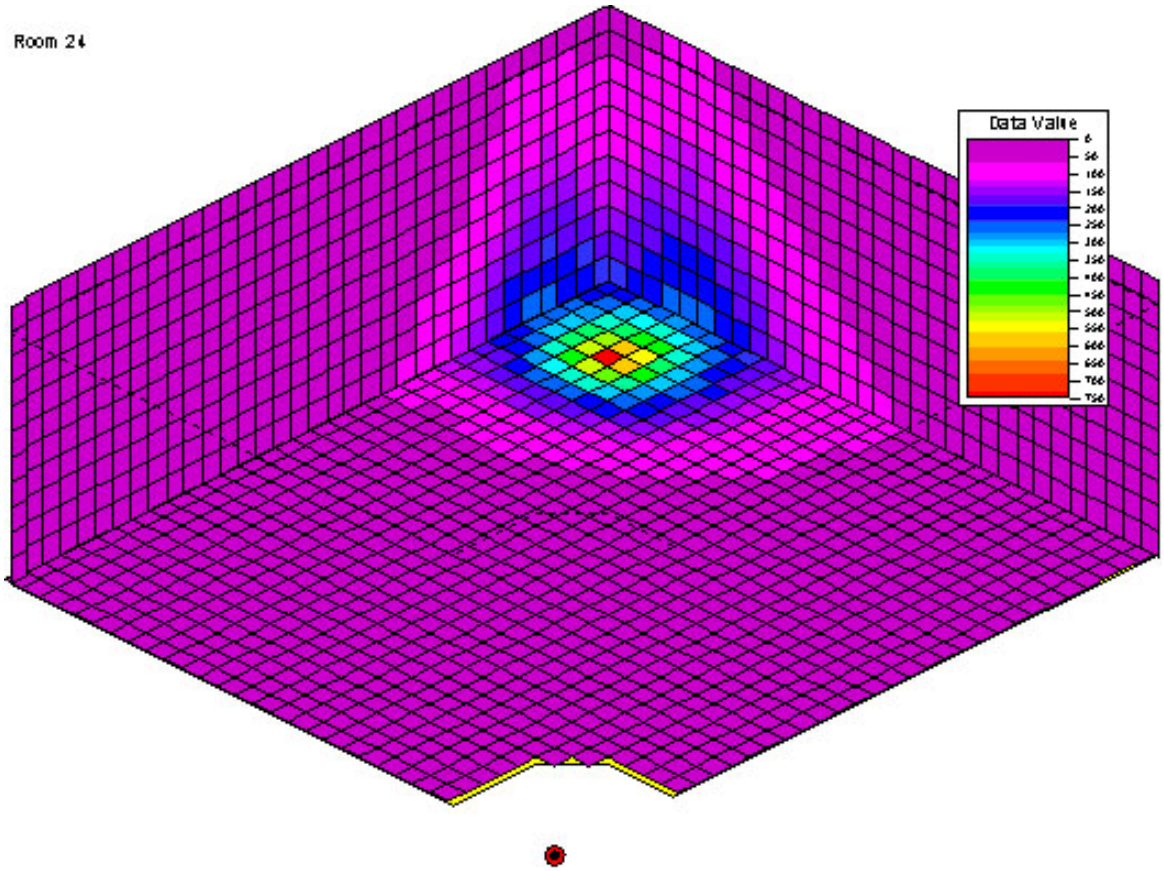


Figure 3.4. Color-Coded Map of Concentration Values of Hot Spot 8 in Room 24 in the Basement of the Coronado Club. Red dot with black dot inside shows location of VSP orientation/ reference point.

Table 3.2 is a summary table of all the grid cells in the basement and main floor of CC. We generated four different data sets from the ground-truth data: the main floor of the CC with floor, wall, and ceiling grid cells identified (38,531 cells total); the basement of the CC with floor, wall, and ceiling grid cells identified (37,168 cells total); the basement with floor only grid cells identified (8,761 cells); and the basement floor grid cells but with their concentration values normalized through a normal distribution transformation (8,761 cells). The basement floor-only data set was created to have a simplified data set and one that could be directly tied to the original sampling results from the CC. Note that not every sampling design was validated on each data set – mostly to simplify the validation project and because the effort would not provide additional information to what was already learned. Table 3.3 shows the match of sampling designs to data sets. For example, the Compliance Sampling design was only validated on the CC main floor – floor, wall, and ceiling – data set, and the CC basement – floor, wall, and ceiling – data set. The assignment of data set to sampling design was driven by the assumptions of the sampling design (e.g., the parametric upper tolerance limit [UTL] design requires a data set that is normally distributed so it was only applied to the normally transformed data set). The hot spot design was only applied to the basement data since that is where all the hot spots were found in the actual contamination of the CC. The hot spots were made to cover the floor, walls, and ceilings. The non-parametric UTL design was applied to three of the four data sets to determine if results were different under the different data sets (i.e., floors only vs. floors, walls, and ceilings). Future work could include a more complex experimental design approach.

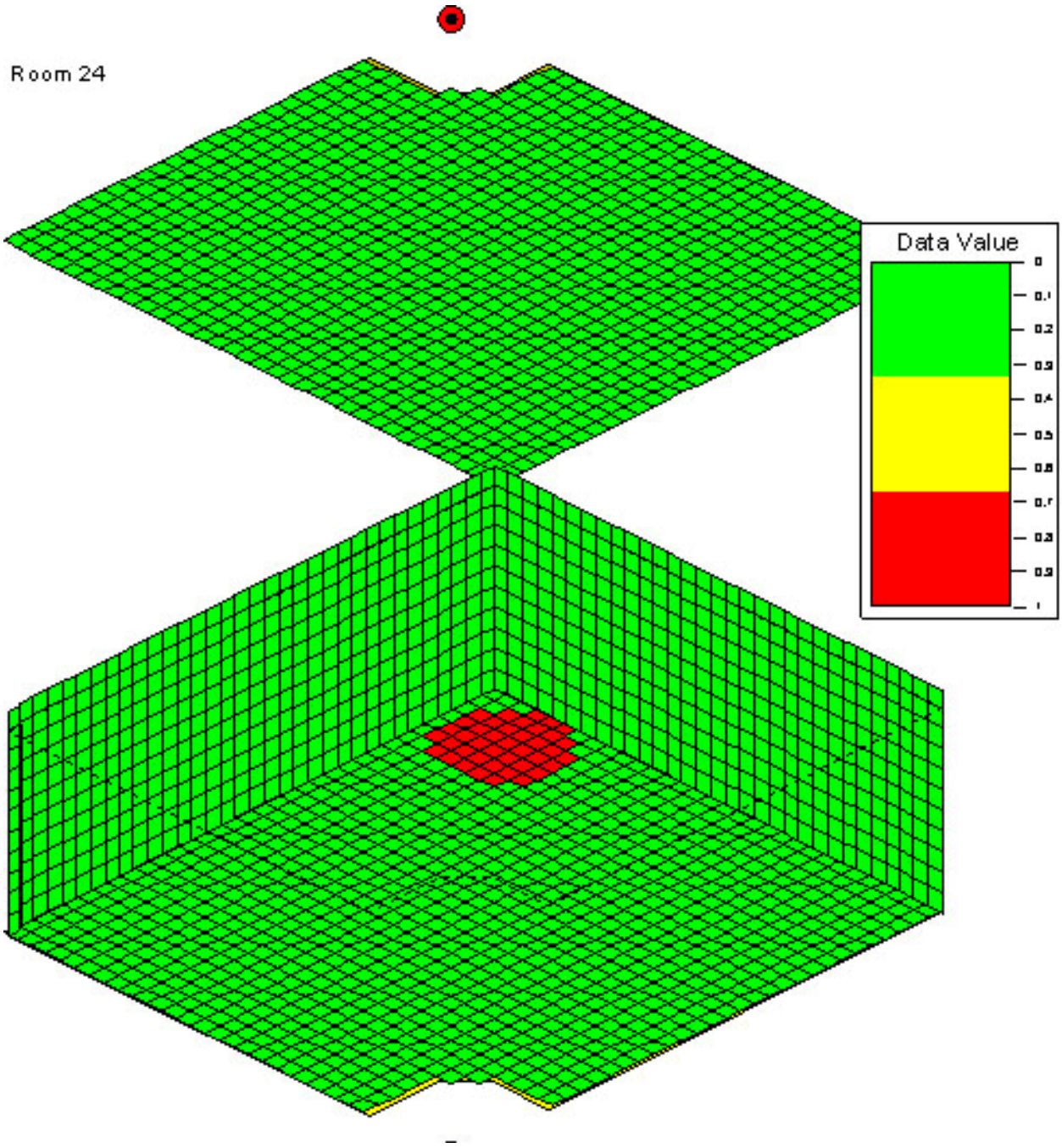


Figure 3.5. Hot Spot 8 in Room 24 Coded to Red/Green to Show Grid Cells with Concentrations Values <275 µg/m^2 (green) and ≥275 µg/m^2 (red)

Table 3.2. Summary Table of Concentration Values in the Ground Truth[(a)]

Total Number of 0.3-m x 0.3-m Grid Cells (f,w,c)[(b)]	Maximum Concentration Value (µg/m^2)	99th Percentile Concentration Value (µg/m^2)	95th Percentile Concentration Value (µg/m^2)	90th Percentile Concentration Value (µg/m^2)	No. Grid Units with Concentration Value > 99th Percentile	Percent of Grids > 99th Percentile (i.e., % of grids contaminated)	No. Grid Units with Concentration Value > 95th Percentile	Percent of Grids > 95th Percentile (i.e., % of grids contaminated)	No. Grid Units with Concentration Value > 90th Percentile	Percent of Grids > 90th Percentile (i.e., % of grids contaminated)
Coronado Club Main Floor – floor, wall, ceiling; 16 rooms										
38,531	2,349.29	1,445	886	646.6	385	0.999195	1,926	4.998573	3,853	9.999740
Coronado Club Basement – floor, wall, ceiling; 21 rooms										
37,168	31,315.6	2,751	337.3	125.76	371	0.998170	1,858	4.998924	3,716	9.997848
Coronado Club Basement – floor only, 21 rooms										
8,761	31,315.6	7,636	515	192	87	0.993037	438	4.999429	876	9.998859
Coronado Club Basement – floor only, transformed to follow normal distribution; 21 rooms										
8,761	1,863.58	1,466	1,328.95	1,256.40	87	0.993037	438	4.999429	876	9.998859

(a) This table shows the range of values for the ground-truth grid cells.

(b) f = Floor; w = wall, and c = ceiling.

Table 3.3. Matrix of Sampling Designs Applied to Ground-Truth Data Sets

	Sampling Designs					
		Compliance Sampling				
Ground Truth Data Sets	Hot Spot	Matched	Mismatched	Non-Par UTL	Par UTL	CJR
cclub_mainfloor[(a)]		X				NA
cclub_basement[(b)]	X	X	X	X		NA
cclub_basement_floor[(c)]				X		NA
cclub_basement_floor_norm[(d)]				X	X	NA

(a) Coronado Club main floor: floor, walls, and ceiling.
(b) Coronado Club basement: floor, walls, ceiling.
(c) Coronado Club basement: floor grid cells only.
(d) Coronado Club basement: floor grid cells only, data transformed to be normally distributed.
CJR = Combined judgment and random.
NA = Non-applicable. The CJR validation method did not use ground truth data sets.
UTL = Upper tolerance limit.

Table 3.2 shows the number of 0.3-m x 0.3-m grid cells in the ground truth we generated, the range of values in the ground-truth data sets, the concentration values for 99th, 95th, and 90th percentiles of the data sets; the number of grid cells greater than the 99th, 95th, and 90th percentiles; and the percent of all grid cells those upper-percentile grid cells represent. If the action level is set at one of those upper-percentiles, the grid cells with concentrations greater than the action level are hence the "contaminated grids" in the data set. The values in the table show that it was impossible to achieve exactly 10%, 5%, or 1% of the ground-truth cells to be defined as contaminated based on where the action level was set. Thus, some simulations used a ground truth of, say, 5.0017% of the cells being greater than the action level, i.e., "contaminated."

The grid cell concentrations in the ground truth show a wide range of values. The basement exhibited more variation in cell-to-cell concentrations than did the main floor. While the data lent itself to a realistic set of hot spots, this "spottiness" resulted in a set of concentration values that, as a data set, did not follow a normal distribution. One of the sampling designs we validated, the parametric UTL, required normal data in order to use the sample size and UTL formulas in VSP. We transformed one set of data, the basement floor data, so that it followed a normal distribution. In the Results section (Section 6.0), we discuss the results of the validation when we applied the parametric UTL test to both the original data, and then the normalized data (see Figure 6.8).

4.0 Sampling Plans Validated

The sampling plans that were selected for validation in this exercise were picked for several reasons:

- They are suited for situations where the goal of sampling is to verify remediation has been successful. They are used to confidently state that no contamination exists in a certain area of a building/structure and can be released for re-occupancy.
- The sampling plans do not consider "false positives" – falsely deciding a clean unit is contaminated. VSP has some designs that do, but they were not considered for this validation project.
- The plans deal with situations where the sampling results come back with "no hits," or "no contamination found." The plans address what can be said with statistical confidence when "no hits" are reported.
- The plans are suited for surface contamination, typically within a building or structure – the focus of the *Remediation Guidance for Major Airports after a Chemical Attack.*
- The plans deal with comparing individual measurements to a threshold – when decisions will be made on individual values and/or percentiles, not the average of all sampling results.
- All of the designs, except the UTL designs, take into consideration the total number of grids in the decision unit when calculating sample size. The decision unit was either the entire basement or the entire main floor. Room-by-room decisions were not considered in this validation.
- The only design that takes into account spatial correlation is the hot spot design; the others assume independence between population units.

The hot spot and compliance design use presence/absence measurements (i.e., sample measurement greater than or less than a threshold or action level). The UTL designs use quantitative measurements to calculate either the maximum, or the UTL, which is then compared to an action level.

All of the designs are implemented in Visual Sample Plan (VSP) version 5.3.1, and have extensive Help for explaining the components of the sampling plan and associated decision rules in the software.

4.1 Hot Spot Sampling Design

The goal of the hot spot designs is to ensure that hot spots of a certain size or larger *do not* exist or go undetected in an area. The design consists of laying out a regular, 2-dimensional grid over the area of concern (i.e., the decision unit), and samples are taken at each node of the grid. If any of the nodes show a hit (i.e., are found to contain contamination above an action level), it *cannot be concluded* that a hot spot of the specified size and shape *does not exist*. The problem is a geometry problem – how far apart can the grids be in order to have a say 95% chance that at least one node falls on the hot spot if in fact it exists? If you lay down the grid 1,000 times, each time selecting a different random location for the starting grid node, how many times does at least one node fall on the hot spot? That is the confidence number. The user can specify the level of confidence required, up to 100%, depending on the cost/risk trade-offs for the problem.

The method used in VSP to achieve these sampling objectives is provided in Gilbert (1987). The VSP software User's Guide (Matzke et al. 2007) provides guidance for how to use VSP to set up the sampling design to detect hot spots. Additional information is provided in the summary report of the design that VSP automatically generates.

4.2 Parametric UTL Sampling Design

One-sided UTLs can be used to statistically test whether a specified area or room in a building is contaminated at concentrations greater than a fixed action level. The statistical meaning, use, and computation of tolerance limits are discussed in Hahn and Meeker (1991) and Helsel (2005, Chapter 6).

VSP computes the number of sample measurements (n) needed to compute a one-sided UTL to statistically test if the true Pth percentile of a normally distributed population exceeds a fixed action level. A discussion of this use of tolerance limits is given in Millard and Neerchal (2001, page 339). Once the sampling results are obtained, the user inputs the n measurements into VSP using the Data Analysis tab of the dialog box. VSP computes the one-sided UTL as well as the mean, standard deviation, median, and the estimated Pth percentile of the n data values. VSP also computes the Shapiro-Wilk W test (Gilbert 1987, pp. 158-160) and a histogram and box plot of the data to assess if the data are normally distributed. One of the VSP outputs is the conclusion that can be drawn from the tests – i.e., whether the site can be declared to be contaminated or not.

In statistical terminology, a one-sided UTL on the true Pth percentile of a population of measurements is identical to a one-sided UTL on the true Pth percentile of the population (Hahn and Meeker 1991, page 61). The true Pth percentile is the value above which 100(1-P)% of the population lies and below which 100P% of the population lies.

VSP calculates the one-sided UTL on the true Pth percentile of a population using the n sample measurements such that at least 100P% of the total population is less than the computed UTL, with X% confidence. For example, if P = 0.90 and X = 0.95, then at least 90% of the population is less than the computed value with 95% confidence.

4.3 Non-Parametric UTL Sampling Design

For the non-parametric design, VSP determines the required number of samples such that the UTL will be the largest of the n measurements in the sample. The rest of the design for the non-parametric case is similar to the parametric case. The non-parametric UTL design should be used for populations where the underlying distribution of the population units has an unknown distribution. If the total population of measurements is not known to be normally distributed, the non-parametric design should be used. If the VSP user selects the parametric design, but VSP determines (using the Shapiro-Wilk W test) that the data cannot be declared normal, VSP advises the user, and suggests using the non-parametric test.

4.4 Compliance Sampling

Compliance sampling with no sample exceedances allowed (C=0) is the sampling design used when the goal is to achieve a high confidence that very few if any grids in the area of concern (decision unit) contain contamination. C=0 refers to the requirement that no measurements in the sample can contain

contamination above the action level in this design in order to conclude the decision unit is not contaminated. There are compliance sampling designs where C can be greater than 0. These designs are in VSP but were not considered for this validation project.

If there is a strong reason to believe that remedial actions have succeeded in removing essentially all contamination from all surfaces of rooms in a building, this may be a useful sampling design. Compliance sampling (Schilling 1982, Chapter 17, pages 474-482) is a statistically based method that can be used to determine the number of samples that should be collected from room surfaces and measured to establish with X% confidence that at least Y% of the surface areas of interest in the room is not contaminated.

Compliance sampling requires that the total room surfaces for which a decision is needed be divided into non-overlapping, equal-size grid units of specified size. The number of all possible grid units in the population is denoted by N. The VSP user specifies values for N, Y, and X. VSP computes the number (n) of the N units ($n < N$) that must be sampled and measured. If the measurements of one of more of the n grid units equals or exceeds a specified action level or is otherwise determined to be unacceptable, then one *cannot* state that there is X% confidence that at least Y% of the surface area of interest in the room is not contaminated. This result could trigger a reassessment of whether additional remedial action is needed. Note that only one contaminated grid unit among the n units sampled reduces confidence to less than X%, and hence could trigger a potentially expensive response action. Compliance sampling is not practical unless none or very few of the N grid units of interest in the room(s) are indeed contaminated. The sample size equation VSP uses for n is from Bowen and Bennett (1988, page 887, Equation 17.8).

4.5 Combined Judgment and Random Sampling

Combined judgmental and random (CJR) sampling (Sego et al. 2007) is a statistically based method that can be used to determine the number of randomly located samples that should be taken in addition to a predetermined number of judgmental (targeted) samples to establish with X% confidence that at least Y% of the decision area does not contain detectable contamination. In addition to combining the results from judgmental and randomly placed samples, the methodology also employs a Bayesian approach to incorporate prior belief regarding the likelihood of detecting contamination with judgmental samples. Recent improvements to the CJR methodology not discussed in Sego et al. 2007 are given in Appendix D.

The CJR method requires that all surfaces in the decision area be divided into non-overlapping, equal-size grid cells of specified size that correspond to the sampling methodology (e.g., grid size of 10 cm x 10 cm). The design is especially suited for use in decision areas where contamination is deemed unlikely, either because there is no likely pathway linking the source to the area, because the area has been decontaminated, or because the initial judgment samples obtained from most likely contaminated areas come back clean. In any case, the objective is to demonstrate, with high probability, that a high percentage of the decision area does not contain detectable contamination, given that none of the samples reveal contamination. If at any time during the sampling process, one of the samples indicates the presence of contamination, the decision area is declared to be contaminated and no further samples for the CJR design need be taken. If this occurs, it may be desirable to implement a hot spot or geospatial sampling plan to characterize the extent of the contamination.

We presume that judgment samples are taken in areas that are more likely to be contaminated than areas available for random sampling. Consequently, if none of the judgment samples reveal detectable contamination, that information, along with prior belief in the cleanliness of the area, is leveraged to reduce the number of random samples required to achieve the desired level of confidence that the decision area does not contain detectable contamination. Consequently, the CJR methodology requires fewer samples than compliance sampling method to achieve the same level of confidence—however, these fewer samples come at the "cost" of the assumptions regarding the prior distribution and the weight of the judgmental samples.

5.0 Methods of Validation

5.1 Method of Validation for Hot Spot, UTL, and Compliance Sampling

There are two types of decision errors that may be of concern that have pertinence relative to validating any sampling design. The first is concluding that an area is sufficiently uncontaminated when in fact it is not sufficiently uncontaminated. The second is concluding that an area is contaminated when it is in fact not contaminated. The health, political, cost, and credibility consequences of the first type of decision error are the primary concern that the GAO review outlined. Although the second type of decision error may be costly and require unnecessary actions, the health consequences are usually negligible. Thus, we have chosen to focus our validation efforts on ensuring that the VSP derived number and placement of samples for all sampling design methods is adequate to confidently conclude that an area is contaminated if in fact it is contaminated, i.e., adequately protecting against erroneously concluding that an area is uncontaminated.

The method for validating the hot spot, UTL, and compliance sampling designs was to apply each design to the same simulated site (in some cases, different areas of the site), taking the number of samples suggested by VSP to meet the design parameters, and using the decision rules in VSP to conclude whether or not the total decision unit is contaminated.

For example, our primary sampling goal may be to state that we are 95% (X%) confident that at least 99% (Y%) of the surface area within some airport terminal is uncontaminated if none of the samples detect contamination (Compliance Sampling, C=0). This same objective can be translated to state that we want to be at least 95% (X%) confident of detecting contamination (having at least one contaminated sample) if 1% (100-Y%) or more of the area is contaminated. The number of samples required is the same for either of these translations. Therefore, we can validate the performance of each VSP method/ parameter combination by setting our ground-truth building data to have approximately 100-Y% of the grid cells to be at or above the action level, and determining the proportion of times (achieved confidence) that at least one sample is found that contains contamination. We expect this proportion to be at least X% (goal confidence).

Continuing the example above, to validate we met the 95% confidence requirement, we repeat the process of taking samples and using the sample results, decide whether to call the site contaminated or not. We do this for 10,000 trials, each trial taking the samples in a different set of n random locations. If in 9,500 or more of those trials we identified the area as contaminated (given that 1% or more of the area in the ground truth was indeed contaminated), we said we validated the 95% confidence level. That is, our achieved confidence is equal to or greater than our goal confidence.

A similar set of realistic goals and simulation trials was used to test the other designs.

Details of the validation include:

- We used an actual indoor contamination scenario. We supplemented actual contamination results with kriged data to obtain a finer mesh grid of contamination. We called this our ground-truth data set. How we created the ground truth was explained in an earlier section.

- Our decision unit was either the main floor or the basement. For various designs, we considered only the floor, or a combination of the floor, walls, and ceiling. For the hot spot design, while we talked about the hot spots being contained within individual rooms on the main floor or basement, rooms were not the decision unit.
- For compliance sampling and UTL sampling designs, we set the action level for saying whether the site was contaminated or not at three different values so we could test scenarios where 90% or more of the site was clean, 95% or more of the site was clean, and 99% or more of the site was clean.
  - We tested three different confidence levels: 90% confidence required, 95% confidence required, and 99% confidence required.
  - We matched each level of contamination, with each level of confidence, creating a 3 x 3 matrix of nine difference designs. We put these nine design parameter sets into VSP to get nine different sample sizes, n. We did this for each sampling design.
- For the hot spot design, we had only one value for the action level, AL = 275 $\mu g/m^2$. At this value, there were "hot spots" of varying size. However, for different scenarios, we varied the definition of the minimum size of the hot spot we said was of concern to us. We called this the hot spot of concern and called it our "design hot spot." The size of the hot spot of concern determined the grid spacing for taking samples. So when we set our grid spacing to look for hot spots of concern that were, say 8 ft or larger in diameter, we would find those 8-ft hot spots, and also find smaller and larger hot spots. As expected, we did not consistently find the smaller hot spots – hence, those smaller hot spots had lower levels of achieved confidence.
- For most tests, we made ground truth match the design parameters. That is, when our X%/Y% goal was to find areas that had 5% or more of the surface area contaminated, we set the action level to a number so that approximately 5% of the units in the ground truth were above the action level. For a few test cases, we changed ground truth by changing the AL was used to define contamination. We wanted to test how the designs performed when ground truth did not match design parameters – when ground truth was either cleaner and dirtier. We labeled these tests as "mismatch." Similarly on the hot spot problem, we designed for finding hot spots of size, say 8 ft in diameter or greater. But we also recorded how we did at finding hot spots of less than 8 ft in diameter.

5.2 Method of Validation for CJR

The CJR validation did not involve repeatedly exercising the sampling design against a known ground truth, as the other sampling design validations did. This is because the Bayesian approach used by the CJR method is not amenable to validations performed on the basis of a single ground truth. Under the Bayesian paradigm, investigators quantify their belief regarding the "state of nature" of the sampling area in terms of a prior distribution. After data are collected, the state of nature is updated and reflected in the form of a posterior distribution. The CJR approach uses this posterior distribution to calculate the probability that a high percentage of the sampling area does not contain detectable contamination.

As with the other sampling designs, the validation of the CJR approach sought to determine whether the random sample size required by the CJR module in VSP resulted in a confidence statement that achieved the goal confidence. Consequently, a large simulation study was conducted in accordance to the assumptions which underlie the CJR model. (Technical details of this study are provided in Appendix D). The study involved using 64,561 cases which covered an extensive range of the six input parameters: the

total number of grid cells, the number of judgmental samples, the a priori probability that a judgmental sample would detect contamination, the risk ratio between judgmental and randomly placed samples, the goal confidence (X%), and the fraction of the decision area that is clean (Y%). By testing this large number of cases, our intent was to validate the CJR methodology under practically all of the conditions in which it may be reasonably used.

In addition to the validation of the CJR confidence, Appendix D also describes a sensitivity study to examine the performance of the CJR method when some of the model assumptions are violated.

Calculating the sample size for the CJR design is computationally intensive. It involves the use of non-trivial numerical integrations and optimization routines that can be difficult to execute correctly and are prone to producing incorrect results without special modifications. Consequently, the systematic effort to ensure that VSP calculates the random sample size correctly is also discussed in Appendix D.

6.0 Simulations used for Hot Spot, UTL, and Compliance Sampling

The simulations involved taking samples from the grid cells in the ground truth and making a decision on whether the area was contaminated or not. For different validation exercises, we could set the action level for what was defined as contaminated to be anything, so we could make the ground truth any level of contamination (e.g., it could be 10% contaminated, 1% contaminated, etc.). We repeatedly sampled the ground truth, each time taking a set of n sample from a set of different random locations. Each time we took a new set of n samples, it was called a "trial." The number of samples we took at a single trial was the number n suggested by VSP, calculated from the formula for the sampling design of interest. For a single trial, we would take either n randomly located samples, or lay down a grid (with defined spacing) in a random location and take a sample at each of n nodes. We recorded the concentration value for the grids in the sample. For each set of samples, we recorded whether we got a "hit" or not. The criterion for a "hit" was defined by the design we were validating. We repeated this for 10,000 trials.

- For the hot spot validation, we noted all hot spots that were encountered by the sampling grid, but we used only the "target" or "design" hot spot for calculating the grid spacing.
- For the UTL and compliance designs, we considered a floor (basement or main floor) to be one decision unit and recorded whether the criterion was met for deciding the decision unit was contaminated or not.

After 10,000 trials, we counted the proportion of trials where we made the right decision – that is, we correctly declared there to be contamination or not. Remember, we controlled ground truth, so we knew whether we were correct or not. We called this percent the *achieved confidence.* We compared the *achieved confidence* to the *goal confidence* (i.e., the X% confidence the user input to VSP as one of the design parameters). If the *achieved confidence* was equal to or greater than the *goal confidence*, the VSP module was deemed "validated." What we were validating was whether VSP directed us to take enough samples, and where to take them, so that at least *some* of the samples would encounter contamination if it existed in X% of the trials. Random samples are just that – sometimes they fall on contaminated areas, other times they miss. But by taking random samples in different locations 10,000 different times, some number of those trials will encounter the contamination if it exists.

The results tables below show the *achieved confidence* vs. the *goal confidence* for each of the sampling designs validated.

For one of the hot spot designs, we tested the sensitivity of the achieved confidence results to the number of simulation trials. We started with just over 600 trials, and tried 2,000, 5,000, and 10,000 trials. The results did not change very much from the 600 trials (achieved confidence of 96.06%) vs. the full 10,000 trials (achieved confidence of 95.78%). The conclusion is that the number of 10,000 is somewhat of an arbitrary number. The details of the number of trials vs. achieved confidence are in a table in Appendix A.

7.0 Results

In almost all of our simulations, our achieved confidence matched or exceeded our goal confidence. This is our metric for validation, so in most cases, we did indeed validate the sampling plans tested. For the few cases that did not match, the achieve confidence was just slightly under the goal confidence by less than a fraction of a percent. The results are presented in a somewhat different format for each of the sampling plans validated. We present here only summary charts of the results. Details are contained in the appendixes.

7.1 Compliance Sampling

Compliance sampling was used for the sampling goal of making sure that Y% or more of the surface area in the decision unit is not contaminated. We tested three different values for the level of confidence: 90%, 95%, and 99%, and three different values for the percent of the surface area that should be uncontaminated: 90%, 95%, and 99%. The results from this nine-way comparison are shown in Figures 7.1 through 7.3.

7.1.1 Ground Truth Matches Design Acceptable Percent Clean

Figure 7.1 shows the results when the action level was set at the value where approximately 90% of the floor, wall, and ceiling grid cells in the ground truth basement and in the main floor were less than the action level. We look at three cases: where the percent acceptable clean we input to VSP as our design criteria is 90% of the grids clean and the goal confidence is set at 90%, 95%, and 99%. The maroon bars are for the basement grid cells; the blue bars are the main floor grid cells. The height of the bars is the percent of the 10,000 trials where at least one of the sampled grid cells was greater than the action level, indicating we would conclude the site was contaminated. Remember, the decision rule for compliance sampling is that none of the *n* samples can have contamination greater than the action level for the decision unit to be declared uncontaminated.

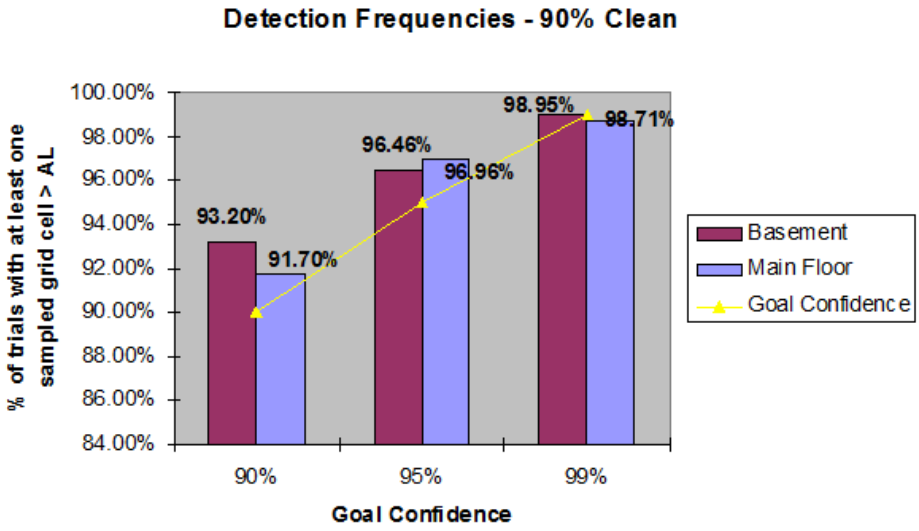


Figure 7.1. Compliance Sampling – Detection Frequencies 90% Clean

Three levels for the goal confidence are shown. The first two sets of bars show that when the goal confidence was 90%, in 93.20% of the trials we found contamination in at least one of the basement grid cell samples, and in 91.70% of the trials we found contamination in at least one of the main floor grid cell samples when the action level was set at the 90th percentile, i.e., 90% of the ground-truth grid cells were clean. The second two sets of bars show that when the goal confidence was 95%, in 96.46% of the trials we found contamination in at least one of the basement grid cell samples and in 96.96% of the trials we found contamination in at least one of the main floor grid cell samples. The third set of bars show that when the goal confidence was 99%, in 98.95% of the trials we found contamination in at least one of the basement grid cell samples, and in 98.71% of the trials we found contamination in at least one of the main floor grid cell samples. The yellow line through the bar graphs shows 90%, 95%, and 99% goal confidence. In all the cases except the 99% goal confidence, the height of the bars was above the yellow goal confidence marks. With rounding, all cases met the goal confidence.

Figure 7.2 is a similar set of six cases, but for these trials we set the action level such that approximately 95% of the floor, wall, and ceiling grid cells in the ground-truth basement, and in the main floor were less than the action level. For the 90% and 95% goal confidence for the basement, we see that our achieved confidence was just slightly under the design goal confidence (89.93% vs. 90% and 94.04% vs. 95%).

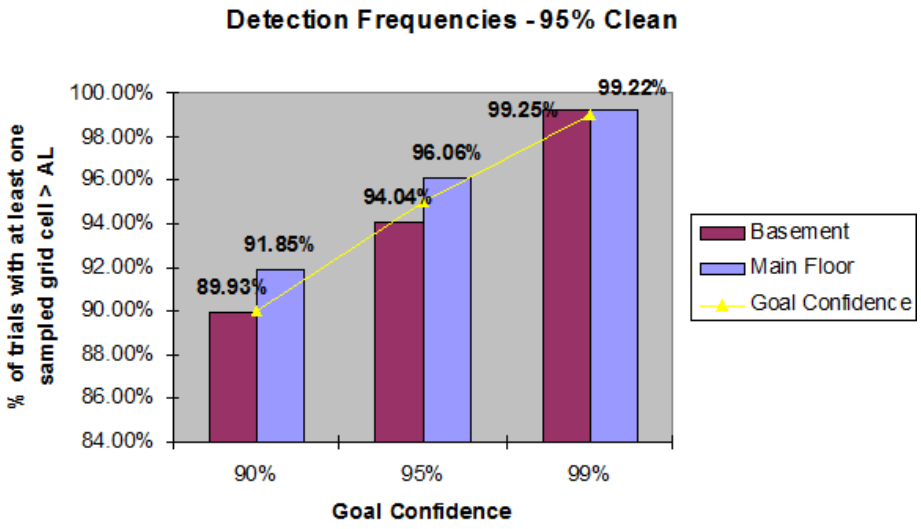


Figure 7.2. Compliance Sampling – Detection Frequencies 95% Clean

Figure 7.3 is a similar set of six cases, but for these trials we set the action level such that approximately 99% of the floor, wall, and ceiling grid cells in the ground truth basement, and in the main floor were less than the action level. For the 90% goal confidence for the basement and main floor, we see that our achieved confidence was just slightly under the design goal confidence (89.92% basement and 89.63% main floor vs. 90%). For the 95% goal confidence for the basement, we were slightly under (94.86% vs. 95%). For the 99% goal confidence for the main floor, we were slightly under (98.78% vs. 99%).

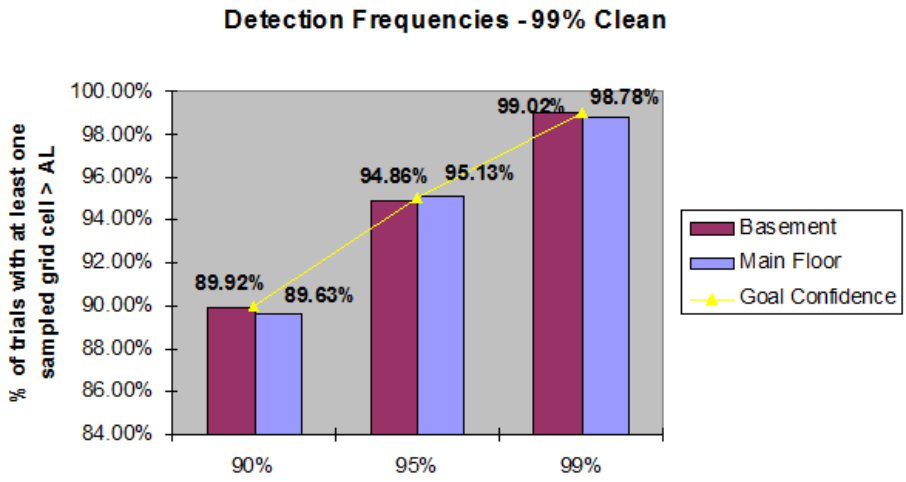


Figure 7.3. Compliance Sampling – Detection Frequencies 99% Clean

7.1.2 Ground Truth Does not Match Design Acceptable Percent Clean

In Figures 7.4 through 7.6, we set the action level to a value such that the percent of the grid cells in the ground truth did NOT match the design percent acceptable clean. We did this to see how robust our results were when ground truth was different from the levels of percent clean we input to VSP for calculating the number of samples to take.

In Figure 7.4, for the maroon bars, we set the action level to a value such that 85% of the grid cells in the ground truth are less than the action level, i.e., uncontaminated. The data set used for these validation trials is the basement of CC, floor, walls, and ceiling. For the trials reported under the maroon bars, approximately 85% of the grid cells in ground truth were defined as uncontaminated. However, we told VSP our design percent clean was 90% so the sample number calculated by VSP was based on 90% clean. Ground truth is thus dirtier than the design parameter. We see in the first maroon bar, when our goal confidence was 90%, in 98.33% of the trials we found at least one of the sampled grid cells to be greater than the action level. This is what we would expect because we are detecting contaminated cells at a greater frequency than we would if in fact 90% of the ground truth was clean. The difference between achieved confidence and goal confidence percent increases in the 95% and 99% goal confidence cases because we are taking more samples and are more likely to get at least one contaminated grid cell in our sample, hence concluding the site is contaminated.

The blue bars show a different situation. Here, we set the action level for ground truth to be greater than the action level we would have selected to have 90% of the grid cells clean. We set the action level so that 95% of the grid cells are clean in the ground truth; we will not encounter as many contaminated cells. In the first blue bar, we see that in 71.04% of the trials we find at least one of the sampled grids to be greater than the action level. This is less than our goal confidence of 90%, but that is what we would expect because our ground truth is cleaner than we expected and we are not encountering contaminated grid cells as frequently. Similar results hold when the goal confidence is 95% and 99% – we under-achieve the goal confidence when the site is cleaner than we say is acceptable. We are not finding contaminated grids in our sample as often because there are indeed fewer of them.

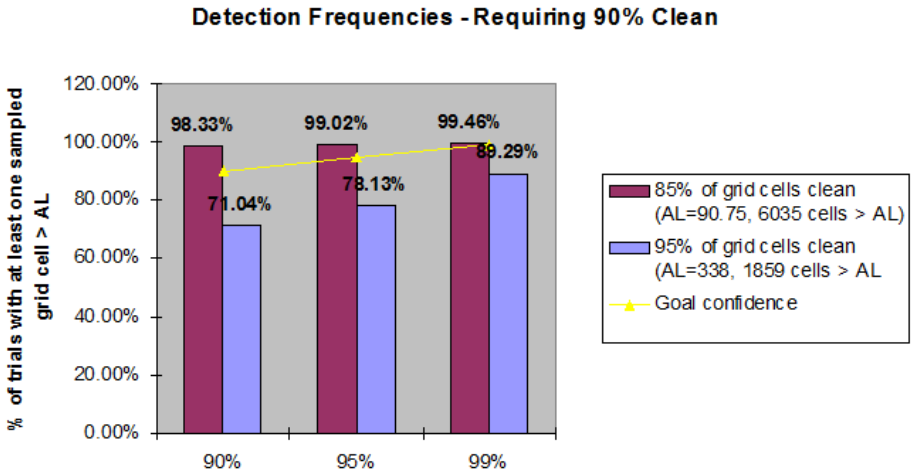


Figure 7.4. Mismatch Compliance Sampling – Detection Frequencies Requiring 90% Clean

Figure 7.5 is a similar set of mismatched cases, where we set ground truth to be dirtier than what we say is acceptable (maroon bars), and cleaner than what we say is acceptable (blue bars). The results are for the case where we say our design percent clean is 95%. We have two difference cases for ground truth – the maroon bars are when the action level is set such that 90% of the grid cells are clean; the blue bars are for when 99.9% of the grid cells are clean. We see the discrepancy between the achieved confidence and the goal confidence to be even larger. For example, when the goal confidence is 90% (first set of bars), we get at least one grid in the sample to be greater than the action level 98.92% of the time when ground truth is dirtier than we expect. But in only 4.68% of the trials do we get one grid in the sample to be greater than the action level when ground truth has 99.9% of the grids clean. Ground truth has very few contaminated grid cells.

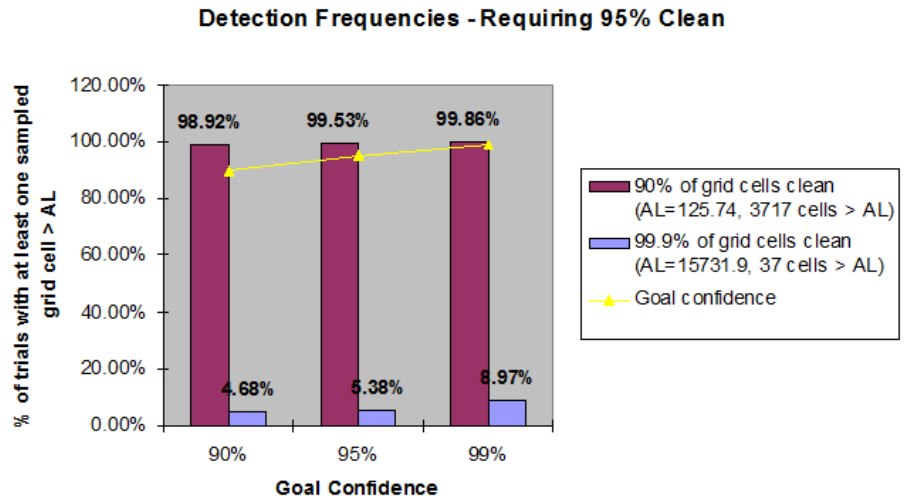


Figure 7.5. Mismatch Compliance Sampling – Detection Frequencies Requiring 95% Clean

Similar results are shown in Figure 7.6 when the acceptable percent clean we give to VSP to be 99% clean.

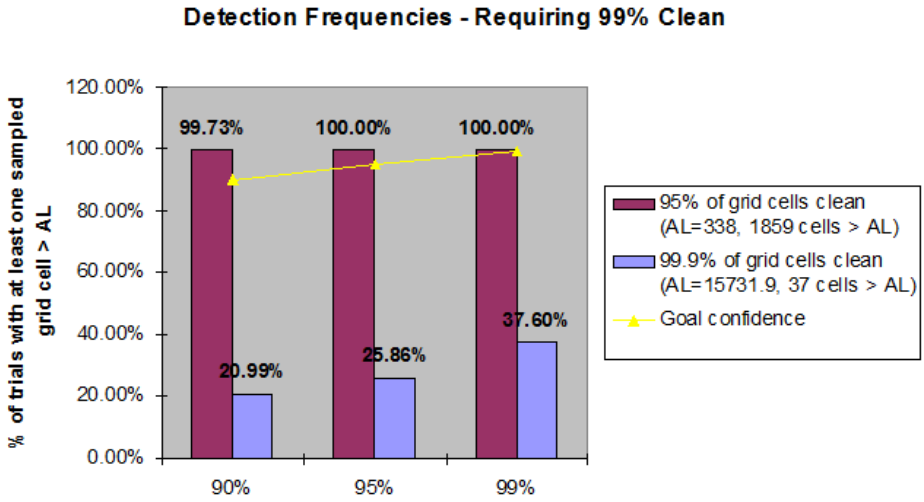


Figure 7.6. Mismatch Compliance Sampling – Detection Frequencies Requiring 99% Clean

One caution that we learn from these Mismatch cases is that the Compliance Sampling (C=0) designs are very conservative and require many samples. (See the tables in Appendix C for sample sizes for each of the studies presented here.) As such, encountering a contaminated grid cell in the sample becomes more likely. This could result in cleanup of a site even though it meets the criteria for "acceptable contamination" – because a contaminated cell is encountered in the sampling. If there is a concern about cleaning up a clean site, sampling designs other then Compliance Sampling (C=0) should be considered.

7.2 UTL

UTL sampling plans were used for the sampling goal of making sure that the Pth percentile of the population is less than the action level of concern, i.e., P% of the units in the population are clean. We tested three different values for the level of confidence: 90%, 95%, and 99%, and three different values for the P% of the population units to be clean: 90%, 95%, 99%. There are two types of UTL sampling plans: one for populations to be sampled where there is no known underlying distribution for the concentration values of the grid cells, the non-parametric UTL; and one for populations where there is a known normal distribution of the grid cells, the parametric UTL. More variability can be expected in populations that do not follow a normal distribution. Hence, the non-parametric UTL designs typically require larger sample sizes than their parametric counter-cases.

7.2.1 Non-Parametric Case

The grid cells in our ground truth, taken from actual contamination data at the CC, comprised a set of data whose distribution was not normal. We stated earlier that the contamination at the CC was "spotty" – with most of the units being uncontaminated, and a few hot spots of high contamination. This does not lend itself to the typical bell-shaped curve expected in population units that are distributed normally. When the data from which samples are being drawn are non-normal, the non-parametric UTL design must be used.

The non-parametric results from the nine-way comparison are shown in Figure 7.7. The data are from the basement (floor only) of the CC.

Figure 7.7 shows the results when the data set used was for the floors only of the basement, non-transformed. The blue bars are when the design Pth percentile was set at 90% clean, the maroon bars are when the design Pth percentile was set at 95% clean, and the cream bars are when the design Pth percentile was set at 99% clean. The light turquoise bar gives the design goal confidence shown here just for reference (for Compliance Sampling, we showed this as a yellow line). Three values for the goal confidence were tested: 90%, 95%, and 99%. The height of the bars, labeled Achieved Confidence, is the % of times the UTL (maximum of the sample for the non-parametric case) was greater than or equal to the AL so the site was considered contaminated. The lower percentiles % clean show lower number of trials where the maximum of the samples was greater than the action level. This would be expected because there are fewer "contaminated" grids in the ground truth, fewer samples are being taken, and hence maximum of the samples will not be as likely to be greater than the AL (which is set at that % clean percentile).

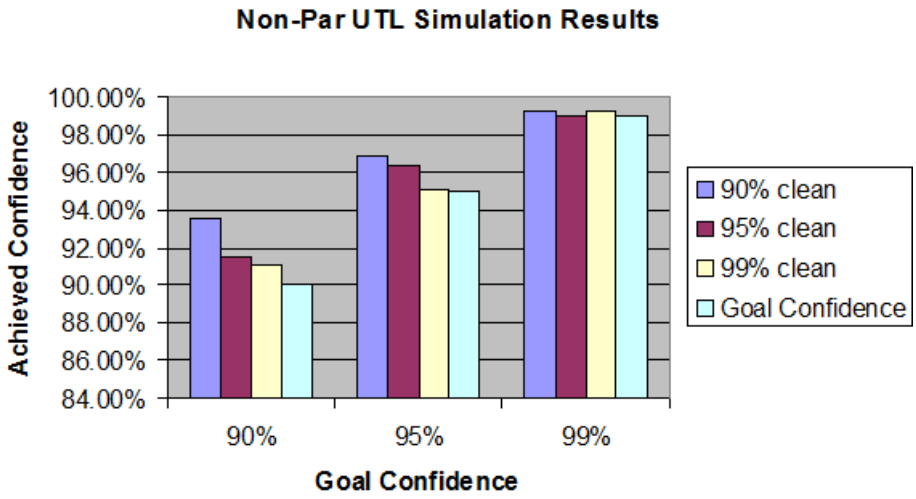


Figure 7.7. Non-Parametric UTL Test with Original Data, Floor Only, Basement Coronado Club

It can be seen that in all three goal confidence cases, the purple, maroon, and cream bars were greater than the reference light turquoise bar. This shows that the achieved confidence was greater than the goal confidence for all three % clean cases. By our definition of "validated," all three cases were validated.

The achieved confidence (the height of the blue, maroon, and cream bars) was calculated by counting the number of times in the 10,000 trials in which the maximum of the n samples taken was greater than the action level. This is the decision rule for the non-parametric UTL test. In Figure 7.1, for the case where the design Pth percentile was 90%, with 99% goal confidence, in 9,925 of the 10,000 trials the maximum of the n samples was greater than the action level (shown in Table 3.2). We conclude we found contamination in 99.25% of the trials. The details of these simulation results for the UTL can be found in Appendix B.

7.2.2 Parametric Case

One option for dealing with non-normal data is to transform the data using a normal transformation algorithm so that the resulting data set is now normal. This was done for the results shown in Figure 7.8. The reason to transform the data is to be able to use the parametric UTL design and test. The parametric test requires fewer samples than the non-parametric test when the assumptions can be met.

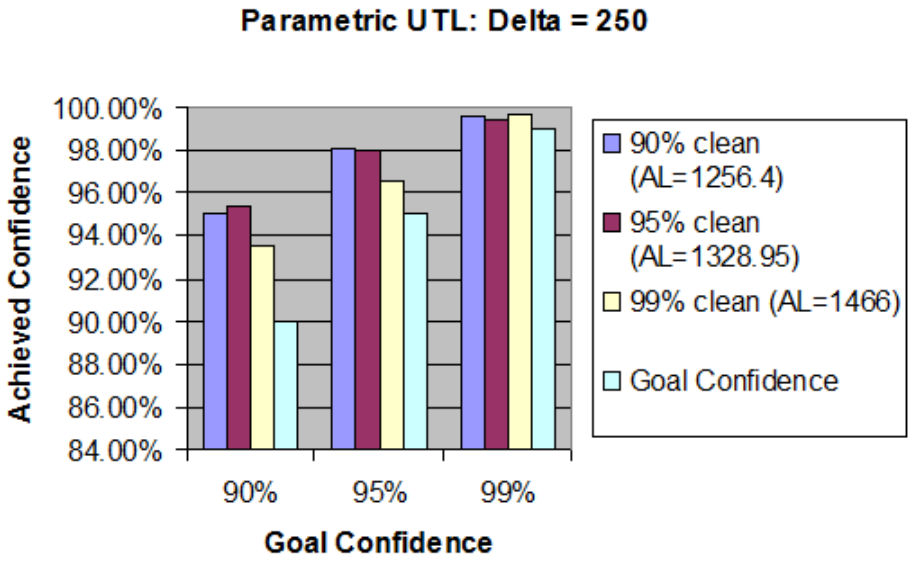


Figure 7.8. Parametric UTL Test with Normalized Data, Floor Only, Basement Coronado Club

Once the data set, used here were the floor samples in the basement of the CC, is normal, we set the action level such that 90%, 95%, or 99% of the grid cells in the population are indeed "clean." In other words, we made ground truth to be 90%, 95%, or 99% clean by selecting the action levels shown in Table 3.2.

The results in Figure 7.8 show that for all goal confidences, the blue, maroon, and cream bars representing Pth percentile to be 90%, 95%, and 99% clean, were all taller than the light turquoise bar, the design goal confidence. Hence, in all nine cases, the achieved confidence exceeded the goal confidence, and we can conclude the designs were validated.

For the parametric UTL case, the decision rule is to conclude a site is contaminated if the calculated UTL on the Pth percentile is greater than the action level. The achieved confidence (the height of the blue, maroon, and cream bars) was calculated by counting the number of times in 10,000 trials where the UTL calculated from the transformed n sample values was greater than or equal to the action level. For example, in the 90% clean, 99% confidence case, this happened in 9,955 out of 10,000 trials. Hence, in 99.55% of the trials we concluded there was contamination because the UTL was greater than the action level. The Delta value, Delta = 250 $\mu g/m^2$, was one of the design parameters input to VSP to calculate sample size. It is discussed in the VSP documentation.

7.3 Hot Spot

In the hot spot case, we were successful in finding a hot spot when one of the nodes of the grid we laid down over the population landed on a hot spot of a size that would cause concern. The algorithms in VSP calculated the grid spacing required to meet the design parameters for the hot spot problem. For example, VSP calculates the grid spacing required to find approximately circular hot spots of 8 ft diameter or greater, and find them 95% of the time (user inputs the design parameters of 8 ft and 95% confidence). The results of the simulation are the number of times, in the 10,000 trials, where at least one grid node lands on a hot spot of concern, which in this example was 8 ft in diameter or more.

The results for the hot spot sampling design validation are presented in a different format from the results of the X%/Y% designs of compliance sampling and UTLs. The ground truth used in the simulations had nine hot spots of various sizes. Some of the hot spots were circular in shape (the semi-major axis is approximately equal to the semi-minor axis); some were not. The algorithm used by VSP assumes an elliptical hot spot. To be conservative, for irregular-shape hot spots, we found the largest ellipse that would totally enclose the irregular hot spot, and input the semi-major and semi-minor axes of that largest enclosing ellipse to VSP as the size of the hot spot of concern. Think of the enclosing ellipse as the ellipse that would result if a rubber band was stretched around the irregular-shaped hot spot so that the hot spot is totally contained within the ellipse. The semi-major and semi-minor axis of that ellipse was input to VSP and shown in Table 3.1. (Note: diameter = twice the radius for circle, and when the semi-minor and semi-major axes are approximately equal, that length of the radius defines the bounding/enclosing circle.) For one set of 10,000 trials, we defined one hot spot as the "design" hot spot – that is, the hot spot size of concern. VSP calculated the grid spacing for this design for sampling to detect this hot spot. For validation results reporting, for each trial we noted ALL hot spots where a node in the design grid fell on a hot spot. We recorded the number of times in 10,000 trials each hot spot was hit. The criteria for validation was that the design hot spot had to have been hit in X% of the trials.

In the hot spot validation, we tested/simulated several hot spot designs. For each design, there was a specific hot spot that was the "design hot spot." Here in the main report we discuss the results when Hot Spot 8 was the design hotspot. In Appendix A, we discuss the results when Hot Spot 2 and Hot Spot 4 were the design hot spot.

In Table 7.1, we see that, for example, Hot Spot 8 was hit 95.78% of the time when we used the grid spacing defined by VSP to find bounding circular hot spots of approximate diameter of 8 ft or greater, 95% of the time. We see that even though the "design hot spot" was Hot Spot 8, with the grid spacing suggested by VSP, we found the other eight hot spots with various success rates. We would only expect to have success rates of 95% or better for hot spots larger than Hot Spot 8 (those hot spots being HS 2, 3, 4, and 9). This is indeed what is shown in Table 7.1. Since the achieved confidence of 95.78% is greater than the goal confidence of 95%, we said the simulation results indeed validated the algorithms in VSP for the hot spot sampling design.

Table 7.1. Simulation Results for Hot Spot 8. Percent of trials in which Hot Spot 8 was "hit" with one of grid sample locations.

	Percentage of Trials Where at Least One Sample was Placed on the Hot Spot								
Number of Trials	HS-1	HS-2	HS-3	HS-4	HS-5	HS-6	HS-7	HS-8	HS-9
10,000	57.92%	99.97%	99.95%	99.95%	58.55%	49.22%	89.42%	95.78%	99.91%

Targeted hot spot is labeled in **blue** and should be detected ~95% of the time.
Hot spots larger than targeted hot spot are labeled in **red** and should be detected >95% of the time.
Hot spots smaller than targeted hot spot are labeled in **green** and will likely be detected <95% of the time.

In Table 7.1, we list the number of trials as 10,000. While we normally ran simulations of 10,000 trials, for Hot Spot 8 we tried different trial numbers to determine if our arbitrarily chosen 10,000 affected the results. We found that we achieved a success rate of 96.06% after only 600 trials. For the Hot Spot 8 design, VSP calculated a square grid size of 2.47 m was required, resulting in 603 samples taken in basement of the CC, which was the decision unit for this test. The details are discussed in Appendix A.

Figure 7.9 shows the actual sample locations in one of the rooms, Room 24 where Hot Spot 8 was located from one of the 10,000 simulation trials. Due to limitations of the VSP display modules, we cannot show Hot Spot 8 concurrent with the sample locations. Room 24 with Hot Spot 8 can be seen in Figure 3.4.

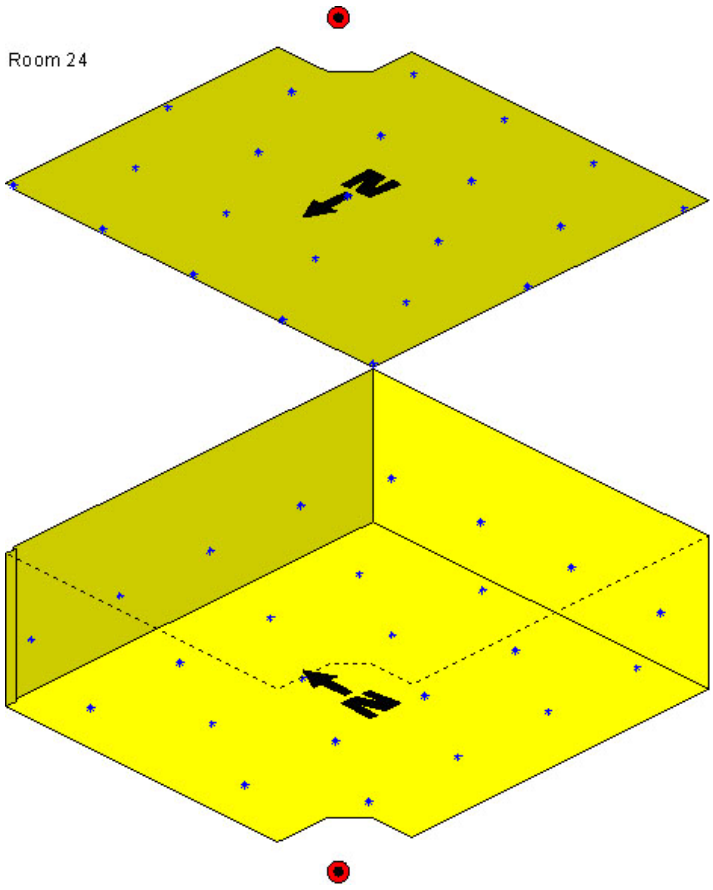


Figure 7.9. Hot Spot 8 Samples in Room 24

Results for the other hot spots are in Appendix A.

7.4 CJR

The CJR design requires the values of six input parameters in order to calculate the number of random samples required to achieve the goal confidence. Table 7.2 shows the range of values that were considered for each input parameter. In all, the various values of these input parameters resulted in 64,561 unique cases. For each case we took 50,000 draws from the posterior predictive distribution of the number of unsampled cells that may contain detectable contamination, assuming that neither the judgmental nor the randomly placed samples detect the presence of contamination. The results of these 50,000 draws were then used to estimate the confidence with a high level of precision.

Table 7.2. Description of Input Parameter Values Used to Generate the 64,561 Cases in the CJR Validation

Parameter Description	Lowest Value	Highest Value	Number of Distinct Values
Number of judgmental samples	1	70	69
Total number of grid cells in the decision area	1	1,000,367	12,353
A priori probability that a judgment sample contains detectable contamination	0.0001	0.50	4,865
Risk ratio between judgmental and randomly placed samples	1	6.034	4,557
The goal confidence (X%)	80%	99.99%	2,072
The percentage of the decision area that is free of detectable contamination (Y%)	80%	100%	2,091

The results of the CJR validation are shown in Figure 7.10**.** Only the cases which required[1] random sampling were plotted. The yellow line indicates where the goal confidence matches the achieved confidence. At this point, we digress briefly to distinguish between the achieved (simulated) confidence and the goal confidence. When investigators implement a sampling design, they must specify their goal level of confidence that they wish the sampling design to achieve. This goal confidence is used (along with other parameters) to determine the required number of samples. Ideally, the number of required samples is the smallest sample size that satisfies the goal confidence level. Consequently, the achieved confidence should always be at least as large as the goal confidence (after accounting for simulation error).

[1] In approximately 35% of the cases, the parameter inputs for the CJR model resulted in no random samples being required because the goal confidence was already achieved by only taking judgmental samples. This can occur when the number of judgmental samples is large, the percentage of the decision area that needs to be clean is relatively low, the goal confidence is low, and/or the prior likelihood of contamination is low. When this does occur, VSP warns the user that the chosen sampling design may rely too heavily on the model assumptions.

In almost all of the cases shown in Figure 7.10, the achieved confidence is no more than 2% higher than the goal confidence (above the yellow line), because calculated sample sizes are rounded up to the nearest whole integer. In approximately 1.4% of the cases that did not require random sampling, the achieved confidence was significantly[2] larger than the goal confidence because a larger than usual random sample size was required by VSP to ensure that the number of required random samples be a non-decreasing function of the total number of grid cells in the decision area, i.e., larger decision areas should require more samples than smaller decision areas (see Section D.1.3 for details). The small proportion of cases whose achieved confidences lie slightly below the yellow line is readily explained by simulation error, since the achieved confidence was calculated using a Monte Carlo simulation. This is thoroughly substantiated in Section D.2.1. It is interesting to note that the cluster of points in Figure 7.10 is somewhat thicker for lower levels of the goal confidence than for higher confidences. This could be a topic of future research.

Having examined the nuances of Figure 7.10, we can state conclusively that these results clearly validate CJR methodology in VSP. Specifically, the CJR module in VSP correctly determines random sample sizes which achieve the goal confidence when neither the judgmental nor any of the random samples detect the presence of contamination.

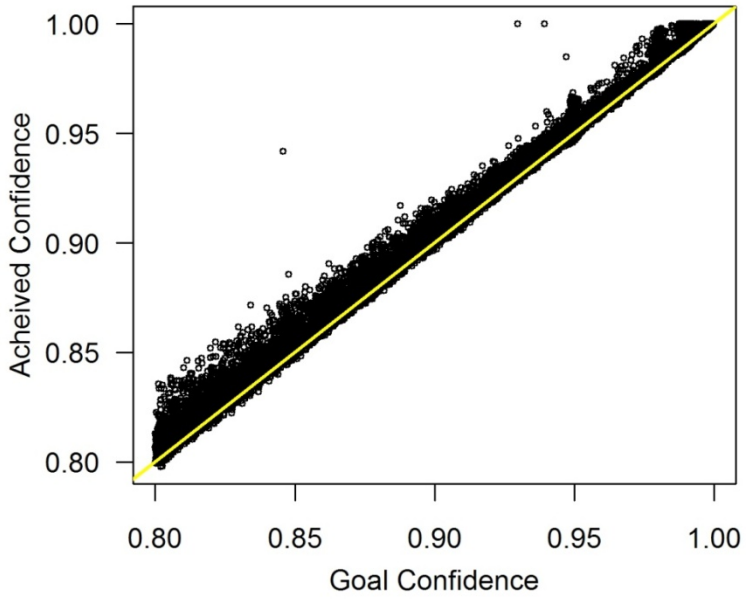


Figure 7.10. Goal versus Achieved (simulated) Confidence for the Cases That Required Random Sampling

As noted, Appendix D describes additional studies regarding the performance of the CJR methodology. The following points summarize the key findings of those studies:

- To explore a broad range of possible values for the CJR input parameters, the confidence was calculated for an extensive number of cases (64,561).

[2] Not explainable by simulation error.

- Simulated values of the confidence were demonstrated to be statistically equivalent to calculated values of the confidence, thus corroborating the algorithms used to calculate the confidence exactly.

- The calculated confidence (based on the sample size indicated by the CJR module in VSP) always meets or exceeds the nominal confidence level.

- Even when the assumptions regarding the a priori probability of contamination and/or the risk ratio between judgmental and randomly placed samples is/are completely wrong, the loss in confidence was no more than 10% in more than half the cases we considered.

- The complex algorithms used to calculate the confidence and determine the required number of random samples have been correctly implemented in VSP and they agree with extensive tests against independently written code.

8.0 Conclusions

The goal of this validation exercise is to demonstrate (with simulated data created from an actual contaminated site) that the desired confidence levels of statistical sampling plans are indeed achieved—provided that investigators adhere to the number and location of samples recommended by VSP. For example, if we use a sampling design that purports to find hot spots of a certain size, say 95% of the time, then in 95% of our simulation trials, we should find hot spots of that size or larger. In the terminology of this report, we say we are comparing the achieved confidence from the simulated trials to the goal confidence from the VSP sampling plan design. The achieved confidence is the percent of time, in say 10,000 simulated trials, we indeed find contamination when it is present at levels of concern specified in the design. In the few cases where the goal confidence was not met, it was not met by less than a percentage point, and most often by less than a fraction of a percentage point. These small aberrations can be attributed to simulation error. The exceptions were the irregular hot spots discussed in Appendix A. However, it should be noted that irregularly shaped hotspots somewhat violate the mathematical assumptions of the hotspot design and, consequently, diminished performance would be expected.

The ground truth we used in the simulations was a set of data consisting of actual sample results from a site where a biological contaminant stimulant was released combined with kriged data in order to represent a fine-mesh grid of contamination. The final data set of ground-truth contamination had 37,168 floor, wall, and ceiling grid cells of size 0.3 m x 0.3 m in the basement, and 38,531 floor, wall, and ceiling 0.3 m x 0.3 m grid cells on the main floor of a building. Both floors were divided into rooms; 16 rooms on the main floor and 21 rooms in the basement. Of the total of 37 rooms, only 34 showed any level of contamination, so only 34 rooms were considered in our simulations. The actual data showed how the walls of the rooms acted as barriers, but in some cases the contamination crossed wall boundaries.

The sampling designs validated in this project are based on probability models, meaning samples are located randomly and the number of samples is calculated such that *if* contamination is present at the level of concern, at least one or more of the samples would encounter the contamination at least X% of the time. Hence, "validation" of the statistical sampling algorithms means ensuring that the "X%" is actually met during the simulation trials. This is done by creating a "ground truth," and a sample design is overlain on the ground-truth grid data. Random locations on the grid are "sampled." The contamination values associated with the sampled grid cells are used in a statistical test (an algorithm contained within VSP), the conclusion from which states whether contamination has been found above the target level and with what confidence this can be concluded. We used VSP to determine the number of samples to take and where to take the samples. The number of times, out of 10,000 simulation trials, where contamination (at the level of concern) was found was called the "achieved confidence." For each simulation trial, a different random set of grid cells in the ground truth were sampled. The "achieved confidence" is compared to the "goal confidence" (input to VSP by the user), and if the achieved confidence is equal to or greater than the goal confidence, the designs in VSP is declared to be *validated*.

Four sampling designs were chosen, abbreviated here as: hot spot, upper tolerance limit, compliance sampling, and combined judgmental and random.

In almost all of our simulations, our achieved confidence matched or exceeded our goal confidence. This is our metric for validation, so in most cases, we did indeed validate the sampling plans tested. For

the few cases that did not match, the achieved confidence was just slightly under the goal confidence, by fractions of a percentage. These small aberrations can be attributed to simulation error. Two of the hot spots did not achieve the goal confidences, but this was due to the allocation of samples to floors, walls, and ceilings when the hot spot crosses the boundaries and the hot spots are of irregular shape. In such situations, the part of the hot spot on the surface of concern should be the size of a hot spot input to VSP. The results are presented in a somewhat different format for each of the sampling plans validated. We present in the main report only summary charts of the results. Details are contained in the appendixes.

The results from the simulations did indeed validate the selected VSP sampling designs. This means that the algorithms within VSP that calculate sample size, sample location, and the conclusions from statistical tests provided the information we expected and achieved the goal confidence levels (to within acceptable tolerances).

During validation effort we discovered a number areas that may prove valuable for future study:

- Cost vs. risk trade-offs were not considered. Type II errors were not considered, i.e., we did not consider situations when sampling designs and analytical testing methods erroneously declare contamination was found when in fact none exists at or above the level of concern. The focus of the validation project was ensuring contamination was found if it existed above the level of concern (i.e., action level). We did however change the ground truth from being contaminated to not being contaminated and evaluated the performance of the sampling designs under these altered conditions. The designs were conservative and found even slightly contaminated areas, but not at the level of the goal confidence.
- "Real life" conditions almost never match the idealized assumptions required by probabilistic sampling designs. Investigations into how these designs perform when the assumptions are not met would provide knowledge of how to best create sampling designs that would be robust under a variety of real life conditions.
- The methods and computer code used for the simulations, the ground-truth data sets, and the procedures for making decisions are available for validation of sampling designs other than those used in this validation effort. These tools could also be used for validating sampling designs not currently available in VSP. Thus, alternative sampling designs can be tested and validated in follow-on efforts using the output and software products created in this project.

We conclude with Table 8.1, which provides a summary of the sampling designs that were validated.

Table 8.1. Summary of the Sampling Designs, Sampling Goals, and Location of Contamination in the Coronado Club

Sampling Design	Description of Design	Goal of Design	Validation Method	Ground Truth[(a)]	Results	Display for Results	Comments
Upper tolerance level (UTL): par, non-par	Calculate an upper confidence limit on a percentile. Compare it to a threshold.	X%/Y%[(b)]	Take n samples in random locations from area where X% of units are clean. Use test to draw conclusion. Repeat for 10,000 trials. Y% of trials should conclude area clean.	For non-parametric designs: basement of Coronado Club: floor, wall, ceiling (f,w,c) For parametric designs: Basement of Coronado Club: f (normalized).	Design confidence matched achieved confidence in all cases.	Bar charts of percent contaminated by confidence, for different levels of ground truth. Excel sheets with bar graphs	Assumes an infinite population. Two designs: parametric, non-parametric. Transformed data (floors in basement only) to match normal before ran parametric design.
Hot spot	Sample at nodes of a grid and if get a "hit" (concentration > threshold) declare that a hot spot was found	Find circular or elliptical hot spots of a minimum size	Take n grid samples (with a random start) from an area where a hot spot of a certain size exists. If any node has a hit, conclude hot was correctly found. Repeat for 10,000 trials. Y% of trials should find the hot spot. Design reports contain table called design summary for each hot spot we designed a sampling plan for. Of all eight hot spots, only created designs and performed simulations for Hot Spots 2, 4, and 8	Basement of Coronado Club: f,w,c. Considered hot spot to spread across f,w,c. 35 separate rooms considered. Considered performance in detecting hot spots whose shape and size were not explicitly accounted for in the design	Design confidence matched achieved confidence in most cases.	Tables of all the hot spots, desired goal confidence, grid spacing for samples, size of hot spot, and percent of time design sample node fell on hot spot (achieved confidence). Report has screen captures from VSP: color by log scale, and red/green hot spot. Design plan has placement of samples from VSP.	Designed for a hot spot of specific size, but using that design, looked at all hot spots. Found hot spots of design size or greater with designed confidence or greater. Found hot spots smaller than design size with smaller-than designed for confidence. .

Table 8.1. (contd)

Sampling Design	Description of Design	Goal of Design	Validation Method	Ground Truth[a]	Results	Display for Results	Comments
Compliance sampling (CS): matched, mismatched	Compare randomly located individual measurements to a threshold	X%/Y%[b]	Take n samples in random locations from area where X% of units are clean. Use test to draw conclusion. Repeat for 10,000 trials. Y% of trials should conclude area clean.	Main floor and basement of Coronado Club: f, w, c. Changed ground truth from design to look at false positives	Design confidence matched achieved confidence in most cases.	Bar charts of design % cont. Tables of desired goal confidence, ground truth, samples taken, achieve goal confidence. Different ground truths in mismatch. Excel sheets with bar graphs.	Assumes a finite population
Combined judgment and random (CJR)	Uses user's beliefs about likelihood of contamination and the risk ratio between judgmental and randomly placed samples to determine number of random samples	X%/Y%[c]	Using the Bayesian posterior predictive distribution, simulate the number of contaminated cells in the unsampled portion of the decision area given 1) the values of the design parameters and 2) neither judgmental nor random samples detected the presence of contamination.	Not applicable.	Goal confidence was achieved in all cases. Sample number required for CJR was not more than would have been required under CS	Plot of the achieved (simulated) confidence versus the goal confidence.	

(a) Contamination at the Coronado Club was spread across main floor and basement. Basement had distinct hot spots that were distributed within and across the 35 rooms. The main floor has smears of contamination on the ceiling.
(b) To say we are Y% confident that X% or more of the units in the population are uncontaminated.
(c) To say we are Y% confident that X% or more of the decision area does not contain detectable contamination.

9.0 References

Amidan BG, GF Piepel, and BD Matzke. 2007. *Experimental Design for the INL Sample Collection Operational Test*. PNNL-17129, Pacific Northwest National Laboratory, Richland, Washington.

Bowen MW and CA Bennett. 1988. *Statistical Methods for Nuclear Material Management*. NUREG/CR-4604, U.S. Nuclear Regulatory Commission, Washington, D.C.

Gilbert RO. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Wiley & Sons, New York.

Griffith RO, JL Ramsey, PD Finley, BJ Melton, JE Brockmann, DA Lucero, SA McKenna, CE Peyton, RG Knowlton, W Einfeld, P Ho, GS Brown, and MD Tucker. 2006. *Joint Sandia/NIOSH Exercise on Aerosol Contamination Using the BROOM Tool*. SAND2006-3784, Sandia National Laboratories, Albuquerque, New Mexico.

Hahn GJ and WQ Meeker. 1991. *Statistical Intervals*. Wiley & Sons, Inc, New York.

Helsel DR. 2005. *Nondetects and Data Analysis, Statistics for Censored Environmental Data*. Wiley & Sons, New York.

Matzke BD, NL Hassig, JE Wilson, RO Gilbert, BA Pulsipher, LL Nuffer, ST Dowson, J Hathaway, CJ Murray, and LH Sego. 2007. *Visual Sample Plan Version 5.0 User's Guide*. PNNL-16939, Pacific Northwest National Laboratory, Richland, Washington. See: http://vsp.pnl.gov.

Millard SP and NK Neerchal. 2001. *Environmental Statistics with S-Plus*. CRC Press, New York.

Schilling EG. 1982. *Acceptance Sampling in Quality Control*. Marcel Dekker, Inc, New York.

Sego LH, KK Anderson, BD Matzke, WK Sieber, S Shulman, J Bennett, M Gillen, JE Wilson, and BA Pulsipher. 2007. *An Environmental Sampling Model for Combining Judgment and Randomly Placed Samples*. PNNL-16636, Pacific Northwest National Laboratory, Richland, Washington.

Appendix A

Hot Spot Sampling Validation: Details on Sampling Design Parameters and Results of Simulation

Appendix A

Hot Spot Sampling Validation: Details on Sampling Design Parameters and Results of Simulation

This appendix gives the details on the sampling design used in the simulations and the detailed validation results for the hot spot (HS-8), the hot spot discussed in the main report. It also gives the design and results details for two other hot spots, HS-4 and HS-2. The results for HS-2 and HS-4 are not as "well-behaved" as the results for HS-8 because they are large relative to the size of the room; are on floors (f), ceilings (c), and walls (w); and Visual Sample Plan (VSP) targets each grid surface independently, i.e., there is no continuous grid layout for all three surfaces.

The conclusion to be drawn for HS-2 and HS-4 is that you should pick the surface (floor, wall, ceiling) where the hot spot component on that surface is of the most concern, determine the size of the hot spot on that surface, and then input that size and shape into VSP, and then use that grid spacing for all surfaces. That way you will be assured of getting the achieved confidence, for the surface of concern, and the size of hot spot of concern.

A.1 Simulation Results for Hot Spot Sampling Designs – Hot Spot 8

A.1.1 Design Summary

Design Parameters	
Map	Coronado Club basement
Design Type	Hot Spot Detection with no false negative
Ground Truth Source	Kriged estimates of surface concentration values calculated from 130 yellow Visolite wipe samples obtained in the joint Sandia/NIOSH Coronado Club exercise (Griffith et al. 2006)
Targeted Hot Spot	HS-8: Round hot spot in Room 24
Action Level	275 $\mu g/m^2$
Required Probability of Detection (1-β)	95%
Decision Unit	Floors, walls, and ceilings of all rooms (excluding Rooms 15 and 25, which have no ground-truth data)
Sample Type	Point samples
Sample Placement	Systematic placement with triangular grid pattern
Hot Spot Shape	1.0 (circular)
Hot Spot Radius	1.28 meters
Calculated by VSP	
Grid Size	8.11 feet (2.47 meters)
Grid Area	57.0107 ft^2 (5.296 m^2)
Required Samples	603

A.1.2 Design Details

For this initial validation testing, the Hot Spot Detection design with no false negatives was used on the basement ground-truth data for the Coronado Club. The ground-truth values being used for this scenario are the estimated surface concentrations obtained by performing ordinary kriging using 130 collected yellow Visolite wipe samples. Figure A.1 shows the floor of the Coronado Club basement with the ground-truth grid cells colored according to a log scale of the measurement values. The hot spot targeted by this design is the somewhat round hot spot in the corner of Room 24, shown in detail in Figure A.2, which we label HS-8. See Chapter 4 of the main report for a description and labeling of hot spots.

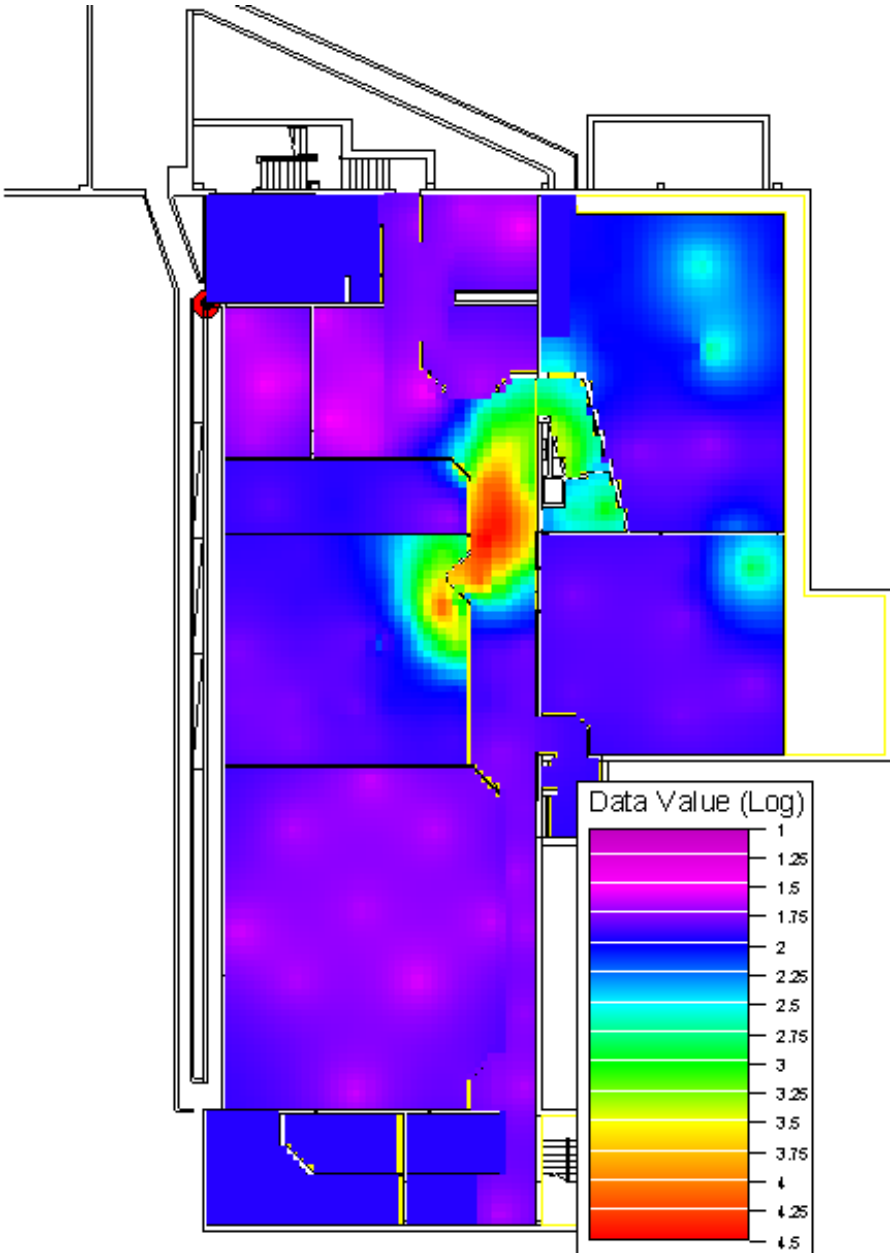


Figure A.1. Coronado Club Basement Floor. Grid cells colored according to log scale of measurement values.

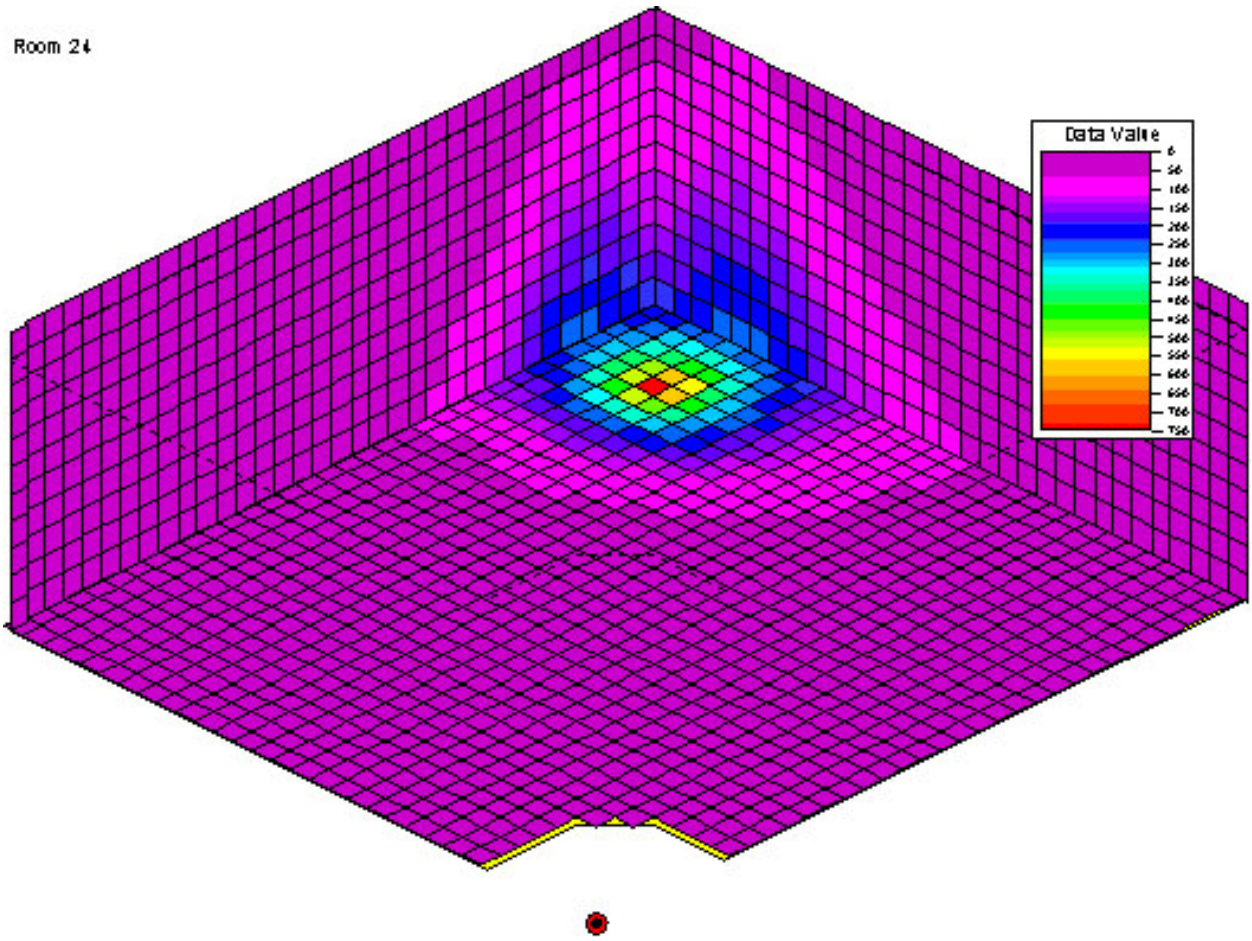


Figure A.2. Hot Spot HS-8 in Room 24

Figure A.3 also shows hot spot HS-8, but with the grid cells colored relative to the action level instead of colored according to their measurement values. The grid cells colored red in this figure have measurement values above the action level of 275 μg/m^2. The estimated smallest ellipse that bounds these action level exceeding grid cells is a circle with a radius of 1.28 meters.

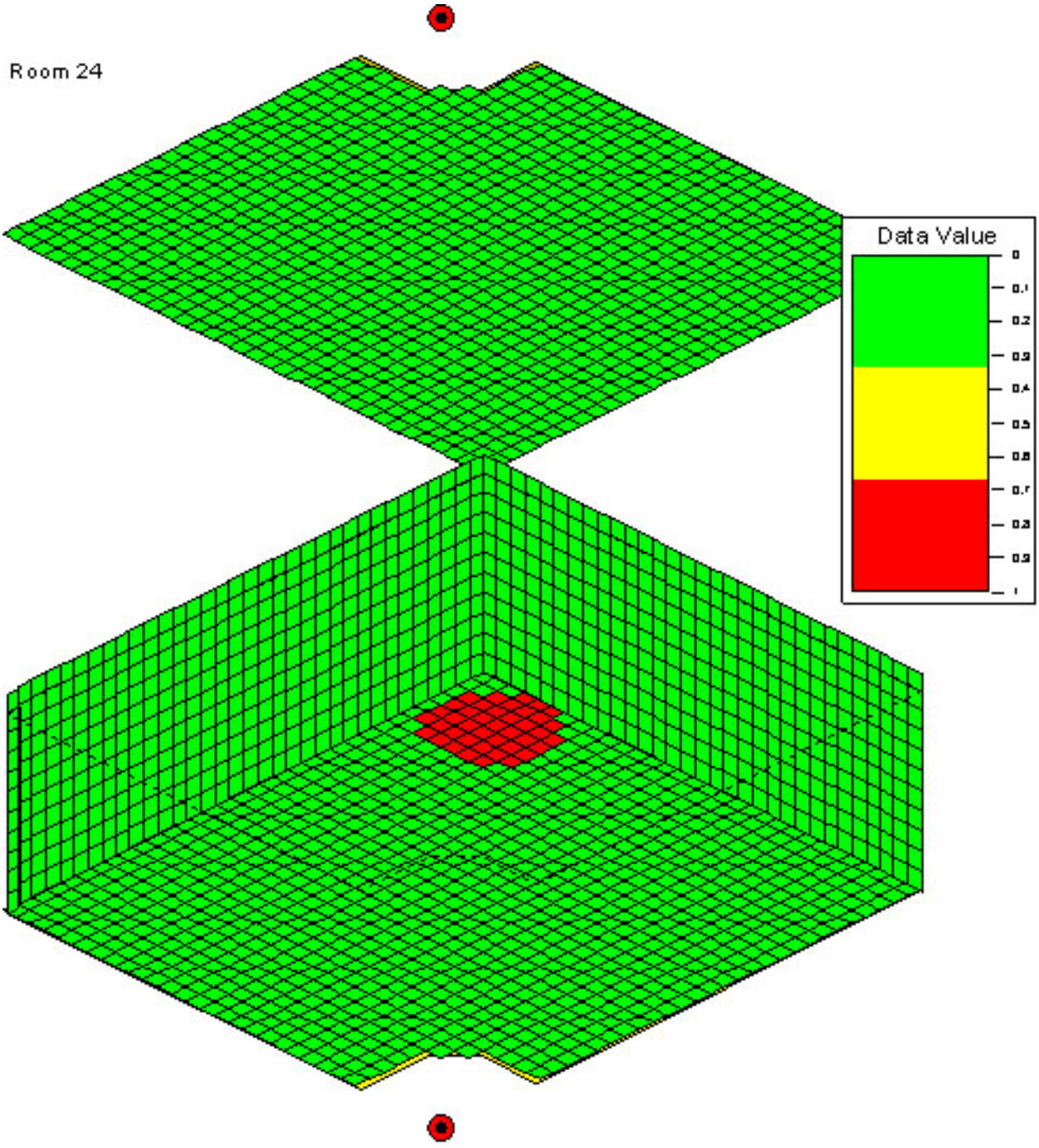


Figure A.3. HS-8 in Room 24 with Cells Exceeding Action Level of 275 μg/m^2 (colored red)

All rooms and surfaces for which we have ground truth or simulated data are included in the decision unit. Point samples were placed systematically in a triangular grid pattern, at the node points. As calculated by VSP, approximately 603 samples spaced 2.47 meters apart, were required to have a 95% probability of locating the hot spot. Figure A.4 shows an example of this grid spacing in Room 24, the room containing the targeted hot spot HS-8. The actual number of samples placed will vary because of grid edge effects. The red dot shows the orientation of the pixel numbering in VSP.

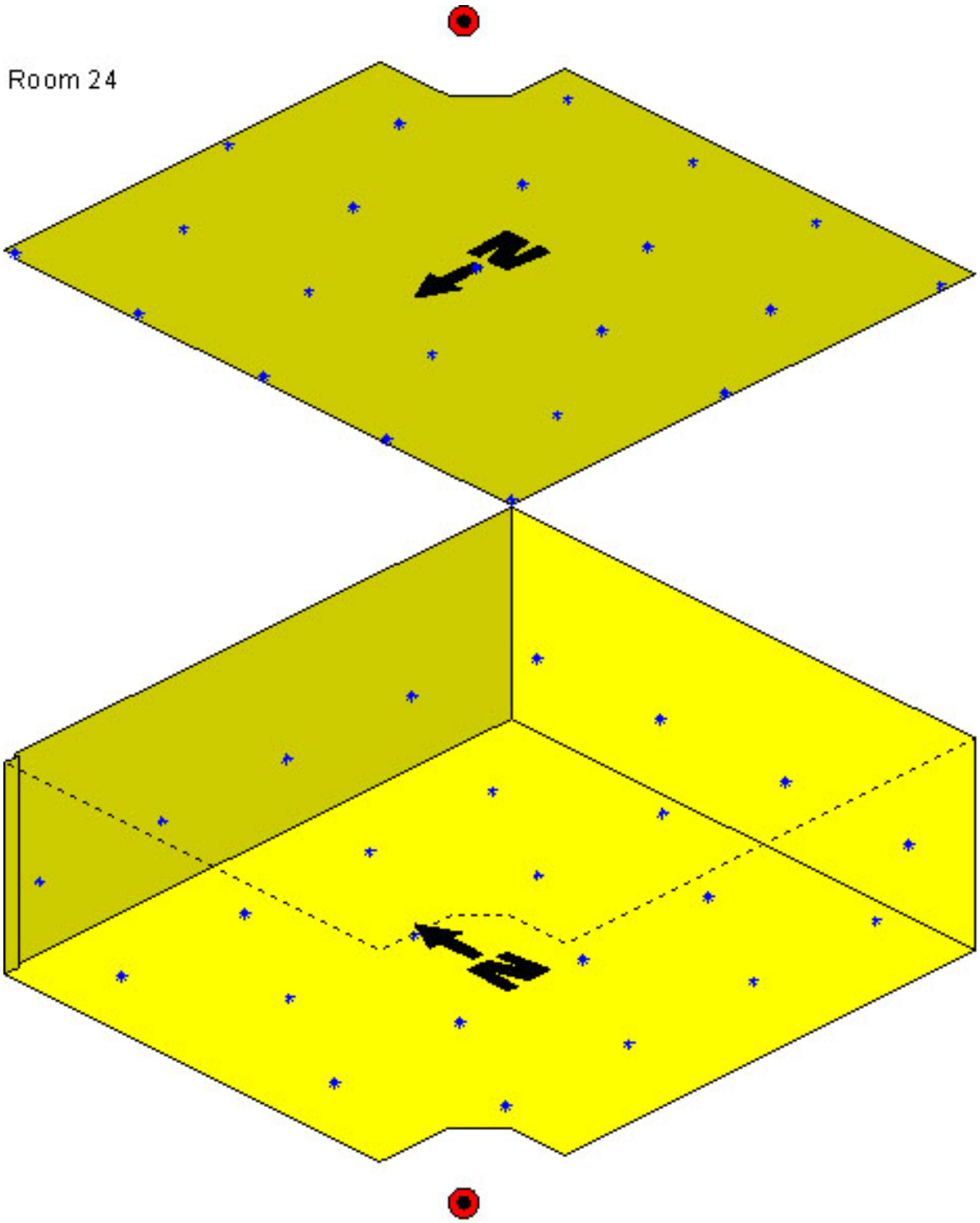


Figure A.4. Samples Placed in Room 24 According to the Grid Spacing Required for this Design

A.1.3 Validation Results

A.1.3.1 Results Summary

A total of 17,755 trials were performed. These sets of trials used an ellipse with a radius of 1.28 meters for the targeted hot spot size. The hot spot HS-8 being targeted by this design was detected in 95.78% of the trials. All hot spots larger than HS-8 were detected in more than 95% of the trials.

A.1.3.2 Trials Performed

For HS-8 only, several sets of 10,000 trials were performed for a total of 17,755 trials. This was done to see if the hit frequencies changed with increased number of trials. They did not. For each trial, samples were systematically randomly placed on the Coronado Club basement map according to the grid pattern and spacing calculated for the design requirements described in the Design Summary, A.1.1. Each time a sample location was placed within a grid cell with a measurement value that exceeded the action level, it was recorded as a hit for the hot spot containing that grid cell. These tallies of hits were used to calculate the frequency that each hot spot was hit.

A.1.3.3 Hit Frequencies

Table A.1 shows the hit percentages for each of the sets of trials performed, along with the total hit frequencies calculated over all 17,755 trials.

Table A.1. Hot Spot Hit Frequencies (HS-8)

Date	No. of Trials	Percentage of Trials Where at Least One Sample was Placed on the Hot Spot								
		HS-1	HS-2	HS-3	HS-4	HS-5	HS-6	HS-7	HS-8	HS-9
4-Dec	634	59.78%	100.00%	100.00%	100.00%	55.99%	50.16%	91.01%	96.06%	100.00%
5-Dec	2,121	58.04%	100.00%	100.00%	100.00%	58.56%	49.22%	89.20%	95.43%	100.00%
5-Dec	5,000	57.62%	99.96%	99.96%	99.96%	58.58%	49.66%	89.72%	95.88%	99.96%
5-Dec	10,000	57.92%	99.97%	99.95%	99.95%	58.55%	49.22%	89.42%	95.78%	99.91%
Total	**17,755**	**57.92%**	**99.97%**	**99.96%**	**99.96%**	**58.47%**	**49.38%**	**89.54%**	**95.78%**	**99.94%**

Targeted hot spot is labeled in **blue** and should be detected ~95% of the time.
Hot spots larger than targeted hot spot are labeled in **red** and should be detected >95% of the time.
Hot spots smaller than targeted hot spot are labeled in **green** and will likely be detected <95% of the time.

A.1.3.4 Results Analysis

These results are what would be expected given the requirements of the design. The targeted hot spot HS-8 was detected at least 95% of the time, while hot spots that were larger than HS-8 were always detected more than 95% of the time. Hot spots which are smaller than the targeted hot spot and therefore were less likely be detected using the specified sampling scheme were detected much less frequently. This validation verifies that the required probability of detection is being met.

A.2 Simulation Results for Hot Spot Sampling Designs – Hot Spot 4

A.2.1 Design Summary

Design Parameters	
Design Name	AV-HS-070120
Map	Coronado Club basement
Design Type	Hot Spot Detection with no false negative
Ground Truth Source	Kriged estimates of surface concentration values calculated from 130 yellow Visolite wipe samples obtained in the joint Sandia/NIOSH Coronado Club exercise (Griffith et al. 2006)
Targeted Hot Spot	HS-4: Irregularly shaped hot spot in Room 18 covering almost entire floor, lots on walls and some on ceiling.
Action Level	275 μg/m^2
Required Probability of Detection (1-β)	95%
Decision Unit	Floors, walls, and ceilings of all rooms (excluding Rooms 15 and 25, which have no ground-truth data)
Sample Type	Point samples
Sample Placement	Systematic placement with triangular grid pattern
Hot Spot Shape	0.809524
Hot Spot Radius	4.2 meters
Calculated by VSP	
Grid Size	23.3212 feet (7.108 meters)
Grid Area	471.011 ft^2 (43.758 m^2)
Required Samples	73

A.2.2 Design Details

Figure A.5 also shows hot spot HS-4 with the grid cells colored relative to the action level instead of colored according to their measurement values. The grid cells colored red in this figure have measurement values above the action level of 275 μg/m^2. The estimated smallest ellipse that bounds these action level exceeding grid cells is a circle with radius 4.2 meters.

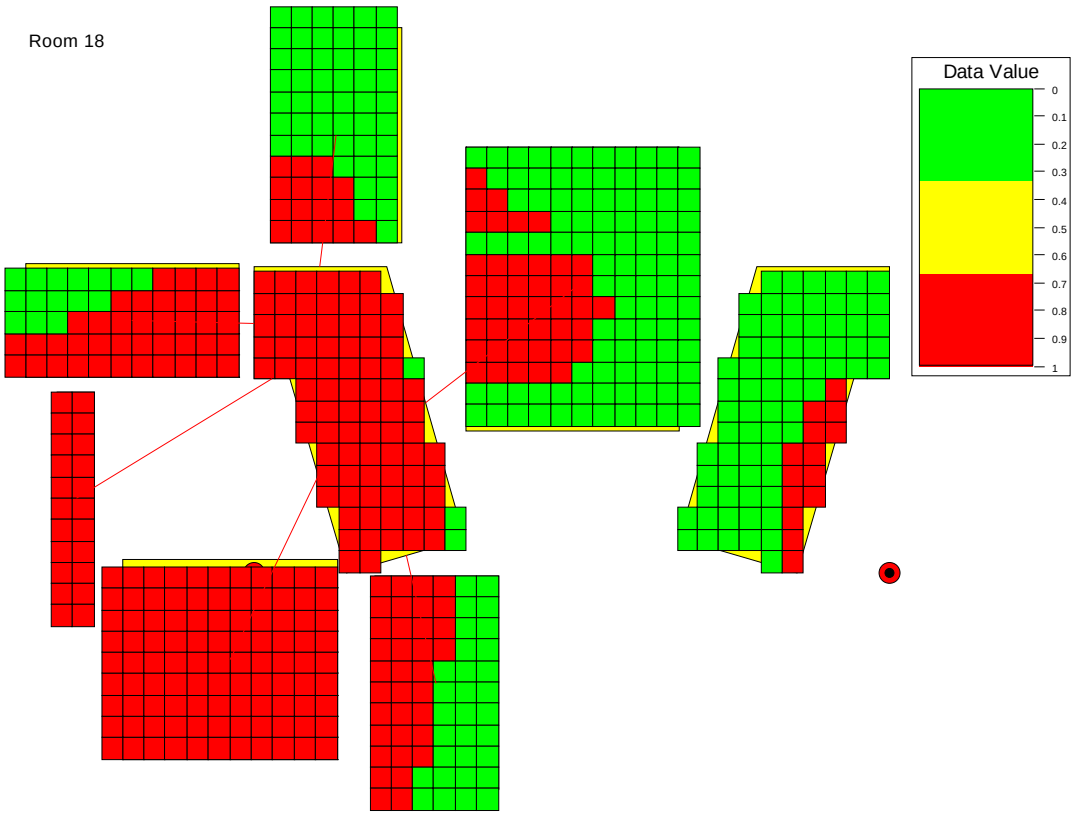


Figure A.5. HS-4 in Room 18 with Cells Exceeding Action Level of 275 µg/m^2 (colored red)

All rooms and surfaces for which we have ground truth or simulated data were included in the decision unit. Point samples were placed systematically in a triangular grid pattern. As calculated by VSP, approximately 73 samples spaced 7.108 meters apart were required to have a 95% probability of locating the hot spot.

A.2.3 Validation Results

A.2.3.1 Results Summary

A total of 10,000 trials were performed. These sets of trials used an ellipse with a radius of 4.2 meters for the targeted hot spot size. The hot spot HS-4 being targeted by this design was detected in 53.91% of the trials. No hot spot was detected more than 90% of the time, including all hot spots larger than HS-4. Hot spot HS-9, the largest hot spot and approximately 2.87 times larger than HS-4, was detected in 89.30% of the trials. These low percentages of detection suggest that the design was insufficient.

A.2.3.2 Trials Performed

A total of 10,000 trials were performed. For each trial, samples were systematically randomly placed on the Coronado Club basement map according to the grid pattern and spacing calculated for the design requirements described in AV-HS-070120-Plan.doc. Each time a sample location was placed within a

grid cell with a measurement value that exceeded the action level, it was recorded as a hit for the hot spot containing that grid cell. These tallies of hits were used to calculate the frequency that each hot spot was hit.

A.2.3.3 Hit Frequencies

Table A.2 shows the hit percentages for each of the sets of trials performed, along with the total hit frequencies calculated over all 10,000 trials.

Table A.2. Hot Spot Hit Frequencies (HS-4)

Date	No. of Trials	Percentage of Trials Where at Least One Sample was Placed on the Hot Spot								
		HS-1	**HS-2**	**HS-3**	**HS-4**	**HS-5**	**HS-6**	**HS-7**	**HS-8**	**HS-9**
20-Jan	10000	6.85%	72.75%	50.84%	53.91%	8.39%	6.03%	14.24%	18.59%	89.30%

Targeted hot spot is labeled in **blue** and should be detected ~95% of the time.
Hot spots larger than targeted hot spot are labeled in **red** and should be detected >95% of the time.
Hot spots smaller than targeted hot spot are labeled in **green** and will likely be detected <95% of the time.

A.2.3.4 Results Analysis

The frequency of detection for HS-4 should be expected to be close to 95%; instead it was detected in 53.91% of the trials. None of the hot spots were detected more than 90% of the time. The hot spot HS-9 was much larger than HS-4 in 89.30% of trials, when according to the requirements of our design, it should have been expected more than 95% of the time.

It is likely that these low percentages of hits occurred because the design was set up to detect the largest ellipse that would fit entirely around the grid cells of hot spot HS-4 when all surfaces were laid flat. Because HS-4 is a large hot spot in relation to the room size and lays on multiple surfaces, it is difficult to identify a two-dimensional shape to represent the hot spot in the design. This issue will be looked at further to design a methodology that will perform robustly in this scenario.

A.3 Simulation Results for Hot Spot Sampling Designs – Hot Spot 2

A.3.1 Design Summary

Design Parameters	
Design Name	AV-HS-070121
Map	Coronado Club basement
Design Type	Hot Spot Detection with no false negative
Ground Truth Source	Kriged estimates of surface concentration values calculated from 130 yellow Visolite wipe samples obtained in the joint Sandia/NIOSH Coronado Club exercise (Griffith et al. 2006)
Targeted Hot Spot	HS-2: Large fairly contiguous hot spot in Room 9 extending onto floor, some walls, and ceiling
Action Level	275 $\mu g/m^2$
Required Probability of Detection (1-β)	95%
Decision Unit	Floors, walls, and ceilings of all rooms (excluding Rooms 15 and 25 which have no ground-truth data)
Sample Type	Point samples
Sample Placement	Systematic placement with triangular grid pattern
Hot Spot Shape	0.867
Hot Spot Radius	4.5 meters
Calculated by VSP	
Grid Size	26.1114 feet (7.959 meters)
Grid Area	590.46 ft^2 (54.856 m^2)
Required Samples	59

A.3.2 Design Details

Figure A.6 also shows hot spot HS-2 with the grid cells colored relative to the action level instead of colored according to their measurement values. The grid cells colored red in this figure have measurement values above the action level of 275 $\mu g/m^2$. The estimated smallest ellipse that bounds these action level exceeding grid cells is a circle with radius 4.5 meters.

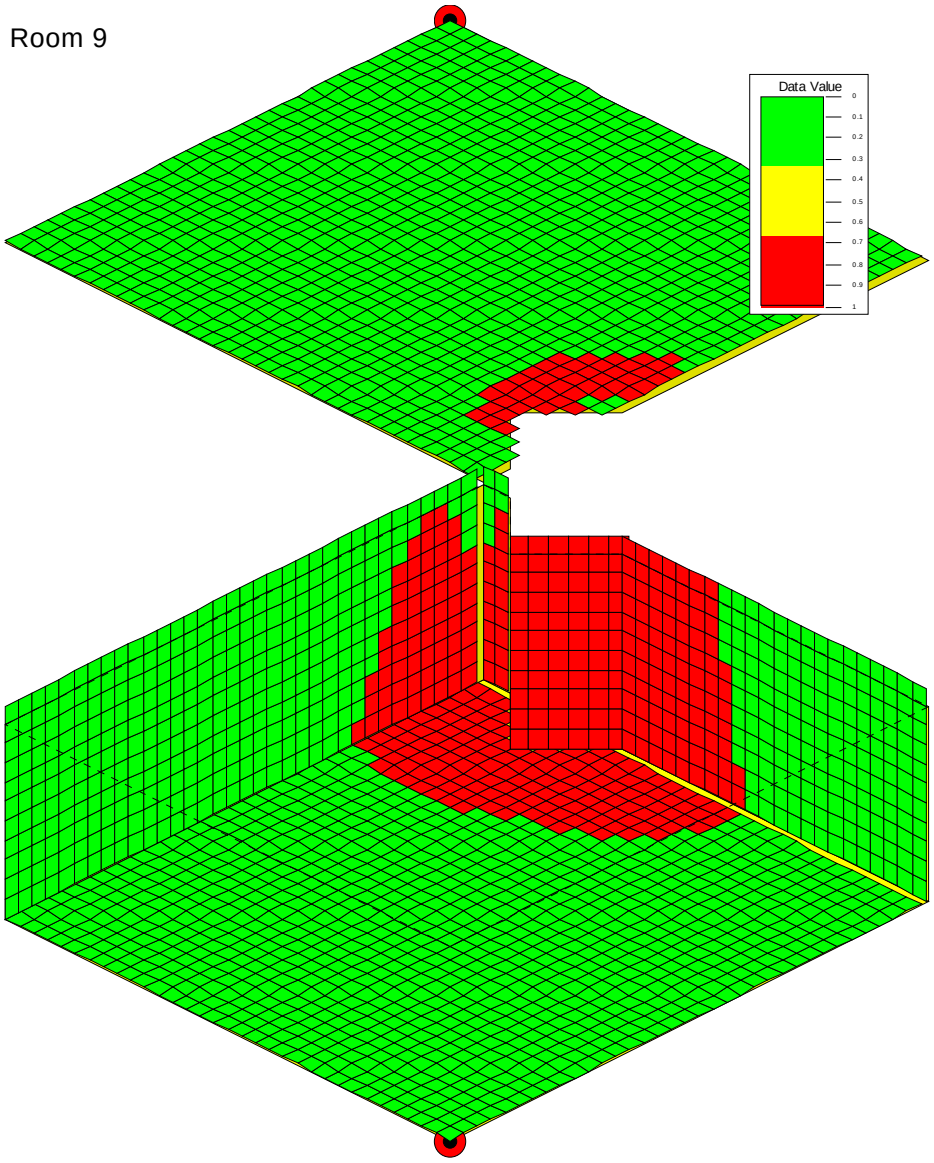


Figure A.6. HS-2 in Room 9 with Cells Exceeding Action Level of 275 µg/m^2 (colored red)

Because our sampling strategy was designed to detect a circular hot spot 4.5 meters in diameter with a 95% probability of detection, we expect that:

1. The targeted hot spot HS-2 will be detected approximately 95% of the time.
2. The hot spots that are larger than the targeted hot spot (HS-9) will be detected at least 95% of the time.
3. The hot spots that are not larger than the hot spot designed for (HS-1, HS-3, HS-4, HS-5, HS-6, HS-7, and HS-8) will be detected less than 95% of the time.

A.3.3 Validation Results

A.3.3.1 Results Summary

A total of 10,000 trials were performed. These sets of trials used an ellipse with a radius of 4.5 meters for the targeted hot spot size. The hot spot HS-2 being targeted by this design was detected in 72.75% of the trials. No hot spot was detected more than 90% of the time. Hot spot HS-9, the largest hot spot and approximately 2.08 times larger than HS-2, was detected in 85.09% of the trials. These low percentages of detection suggest that the design was insufficient.

A.3.3.2 Trials Performed

A total of 10,000 trials were performed. For each trial, samples were systematically randomly placed on the Coronado Club basement map according to the grid pattern and spacing calculated for the design requirements described in the Design Summary, A.2.1. Each time a sample location was placed within a grid cell with a measurement value that exceeded the action level, it was recorded as a hit for the hot spot containing that grid cell. These tallies of hits were used to calculate the frequency that each hot spot was hit.

A.3.3.3 Hit Frequencies

Table A.3 below shows the hit percentages for each of the sets of trials performed, along with the total hit frequencies calculated over all 10,000 trials.

Table A.3. Hot Spot Hit Frequencies (HS-2)

Date	No. of Trials	Percentage of Trials Where at Least One Sample was Placed on the Hot Spot								
		HS-1	**HS-2**	**HS-3**	**HS-4**	**HS-5**	**HS-6**	**HS-7**	**HS-8**	**HS-9**
21-Jan	10,000	5.42%	72.75%	44.78%	47.99%	5.72%	4.91%	12.71%	15.00%	85.09%

Targeted hot spot is labeled in **blue** and should be detected ~95% of the time.
Hot spots larger than targeted hot spot are labeled in **red** and should be detected >95% of the time.
Hot spots smaller than targeted hot spot are labeled in **green** and will likely be detected <95% of the time.

A.3.3.4 Results Analysis

The frequency of detection for HS-2 should be expected to be close to 95%; instead it was detected in 72.75% of the trials. None of the hot spots were detected more than 90% of the time. The hot spot HS-9 was much larger than HS-4 in 85.09% of trials, when according to the requirements of our design, it should have been expected more than 95% of the time.

It is likely that these low percentages of hits occurred because the design was set up to detect the largest ellipse that would fit entirely around the grid cells of hot spot HS-4 when all surfaces were laid flat. Because HS-4 is a large hot spot in relation to the room size and lays on multiple surfaces, it is difficult to identify a two-dimensional shape to represent the hot spot in the design. This issue will be looked at further to design a methodology that will perform robustly in this scenario.

A.4 Conclusions

Large hot spots that cover several surfaces (floor, wall, ceiling) provide the VSP hot spot software code with challenges because VSP treats separate surfaces independently. This accounts for why some of the hot spot problems (e.g., HS-2 and HS-4 which where large, multi-surface hot spots) had simulation results where the *achieved* confidences were less than the *desired* confidences. A possible reason for this is that VSP resets the starting point for the grid layout for sampling at a new random location for each surface independently, thus not allowing for an integrated grid that continues its coverage across floor, wall, and ceiling dividers. The report explained how to handle this situation, but a software fix is required if independently-targeted surfaces is not the goal of the project.

A.5 Reference

Griffith RO, JL Ramsey, PD Finley, BJ Melton, JE Brockmann, DA Lucero, SA McKenna, CE Peyton, RG Knowlton, W Einfeld, P Ho, GS Brown, and MD Tucker. 2006. *Joint Sandia/NIOSH Exercise on Aerosol Contamination Using the BROOM Tool*. SAND2006-3784, Sandia National Laboratories, Albuquerque, New Mexico.

Appendix B

UTL Sampling Validation: Details on Sampling Design Parameters and Results of Simulation

Appendix B

UTL Sampling Validation: Details on Sampling Design Parameters and Results of Simulation

The two Upper Tolerance Limit (UTL) sampling plans considered for this validation project are the parametric, and the non-parametric sampling designs. The main report discusses these two plans. If the population of units in the decision area which is to be sampled, and about which a decision is to be made, has concentration values whose distribution closely resembles a normal distribution, then the sample sizes and sample design under the VSP design called UTL Parametric should be used. But if the distribution of the population units is unknown, or follows something other than a normal distribution, then the VSP design called UTL Non-Parametric should be used.

B.1 Overall Design Summary

Table B.1 shows the general design used in both the parametric and non-parametric designs.

Table B.1. Overall UTL Design Summary

Overall Design Parameters	
Map	Coronado Club basement
Designs Tested	Parametric and Non-Parametric Upper Tolerance Limit
Ground Truth Source	Raw and normally transformed kriged estimates of surface concentration values calculated from 130 yellow Visolite wipe samples obtained in the joint Sandia/NIOSH Coronado Club exercise (Griffith et al. 2006)
Decision Unit	Floors of all rooms (excluding Rooms 15 and 25 which have no ground-truth data)
Sample Type	Grid cell samples 0.3 meter x 0.3 meter square
Sample Placement	Random

B.2 Parametric UTL Sampling Design on Normal Transformed Data

Table B.2 contains the details for the sampling design used on the Normal Transformed Data. Most of this information should match the data in Table 3.2 of the main report. Units for the action levels, width of gray region, and standard deviation are $\mu g/m^2$. Shown are the number of samples taken in the decision unit, which is the Coronado Club basement, floor surface only.

Table B.2. Summary of Design Parameters for Parametric UTL on Normal Transformed Data

Design Parameters			
Design Type	Parametric Upper Tolerance Limit		
Data Set	Floor data transformed to follow a normal distribution		
Total Number of Grid Cells on Map	8,761		
Required Percentages Clean Tested	90%	95%	99%
Action Levels in $\mu g/m^2$	**90% clean**	**95% clean**	**99% clean**
	1,256.4	1,328.94	1,466
Required Confidences Tested	90%	95%	99%
Delta (width of gray region) in $\mu g/m^2$		250	
Estimated Standard Deviation in $\mu g/m^2$		200.02	

Calculated by VSP									
Number of Samples Required	**90% confidence**			**95% confidence**			**99% confidence**		
	90% clean	**95% clean**	**99% clean**	**90% clean**	**95% clean**	**99% clean**	**90% clean**	**95% clean**	**99% clean**
	13	16	24	17	22	32	27	34	49

The results for this simulation are discussed in the main report, Section 7.2.2, and shown in Figure 7.8, Parametric UTL Test with Normalized Data, Floor Only, Basement Coronado Club. As can be seen, all of the achieved confidences equaled or exceeded the goal confidences, indicating that the design was validated.

Table B.3. Simulation Results (achieved confidence) from Parametric UTL Sampling Design. Data: Basement , floor surface, transformed to be Normal

Percent Clean	**Goal Confidence 90%**	**Goal Confidence 95%**	**Goal Confidence 99%**
90% clean (AL=1,256.4)	94.99%	98.09%	99.55%
95% clean (AL=1,328.95)	95.39%	98.02%	99.47%
99% clean (AL=1,466)	93.47%	96.55%	99.62%

B.3 Non-Parametric UTL Sampling Design on Raw Data

Table B.4 contains the details for the sampling design used on the non-normalized data (called "raw data" or "original data"). Most of this information should match the data in Table 3.2 of the main report. Units for the action levels are $\mu g/m^2$. Unlike the parametric case, the non-parametric case does not require "width of gray region" or "standard deviation" because the non-parametric UTL sample size

calculation and test use only number of total grids, required confidence, and action level. The non-parametric UTL is the maximum of the values in the sample. The non-parametric UTL is compared to the action level in the UTL non-parametric test. Shown are the number of samples taken in the decision unit, which is the Coronado Club basement, floor surface only. Note how much larger these sample sizes are than those shown in Table B.2 for the parametric design. This is because the non-parametric test makes no assumptions, and as such, has to be more conservative.

Table B.4. Summary of Design Parameters for Non-Parametric UTL Sampling Design on Raw Data

Design Parameters									
Design Type	Non-Parametric Upper Tolerance Limit								
Data Set	Original floor data								
Total Number of Grid Cells on Map	8,761								
Required Percentages Clean Tested	90%			95%			99%		
Action Levels in µg/m^2	**90% clean**			**95% clean**			**99% clean**		
	192			515			7,636		
Required Confidences Tested	90%			95%			99%		
Calculated by VSP									
Number of Samples Required	**90% confidence**			**95% confidence**			**99% confidence**		
	90% clean	**95% clean**	**99% clean**	**90% clean**	**95% clean**	**99% clean**	**90% clean**	**95% clean**	**99% clean**
	22	45	230	29	59	299	44	90	459

The results for this simulation are discussed in the main report, Section 7.2.1, and shown in Figure 7.7, Non-Parametric UTL Test with Original Data, Floor Only, Basement Coronado Club. As can be seen, most of the achieved confidences equaled or exceeded the goal confidences, indicating that the design was validated. The exception was for the 99% confidence, 95% clean, where the achieved confidence was slightly less than the goal confidence (98.97% vs. 99%)

Table B.5. Simulation Results (achieved confidence) from Non-Parametric UTL Sampling Design. Data: Basement, floor surface, non-transformed.

Percent Clean	**Goal Confidence 90%**	**Goal Confidence 95%**	**Goal Confidence 99%**
90% clean (AL=192)	93.47%	96.91%	99.28%
95% clean (AL=515)	91.49%	96.37%	98.97%
99% clean (AL=7,636)	91.02%	95.10%	99.25%

B.4 Non-Parametric UTL Sampling Design on Normal Transformed Data

Table B.6 contains the details for the non-parametric UTL sampling design used on the normal transformed data. This design is not discussed in the main report because it was done as a comparison effort to see what would happen if a non-parametric design and test were used on parametric data. Shown are the number of samples taken in the decision unit, which is the Coronado Club basement, floor surface only. The sample sizes match those in Table B.4, as should be expected since it is the same non-parametric test, just applied to a different data set.

Table B.6. Summary of Non-Parametric Design Parameters Applied to Normal Transformed Data

Design Parameters									
Design Type	Non-Parametric Upper Tolerance Limit								
Data Set	Floor data transformed to follow a normal distribution								
Total Number of Grid Cells on Map	8,761								
Required Percentages Clean Tested	90%			95%			99%		
Action Levels	**90%**			**95%**			**99%**		
	1,256.4			1,328.94			1,466		
Required Confidences Tested	90%			95%			99%		
Calculated by VSP									
Number of Samples Required	**90% confidence**			**95% confidence**			**99% confidence**		
	90% clean	**95% clean**	**99% clean**	**90% clean**	**95% clean**	**99% clean**	**90% clean**	**95% clean**	**99% clean**
	22	45	230	29	59	299	44	90	459

The results for this simulation, shown in Table B.7, are not discussed in the main report. As can be seen, all of the achieved confidences equaled or exceeded the goal confidences, indicating that the design was validated.

Table B.7. Simulation Results (achieved confidence) from Non-Parametric UTL Sampling Design. Data: Basement (floor), normalized

Percent Clean	**Goal Confidence 90%**	**Goal Confidence 95%**	**Goal Confidence 99%**
90% clean (AL=1,256.4)	93.29%	96.91%	99.21%
95% clean (AL=1,328.95)	91.74%	96.47%	99.02%
99% clean (AL=1,466)	91.02%	95.66%	99.33%

B.5 Non-Parametric UTL Sampling Design on Raw Data, All Surfaces

Table B.8 contains the details for the non-parametric UTL sampling design used on the original, or raw non-transformed data for all surfaces in the basement of the Coronado Club. This design is not discussed in the main report because it was done as a comparison effort to see what would happen if a non-parametric design and test were used on original, or non-transformed floor, wall, and ceiling data. Shown are the number of samples taken in the decision unit, which is the Coronado Club basement, all surfaces. The sample sizes are the same as those in Table B.4, as should be expected since it is the same non-parametric test, just applied to a different data set, and N, the total number of units did not affect the results since in both cases, N was very large.

Table B.8. Summary of Design Parameters for Non-Parametric Design Applied to Coronado Club Basement, Floor, Wall, and Ceiling Raw Data

Design Parameters									
Design Type	Non-Parametric Upper Tolerance Limit								
Data Set	Original Coronado Club Basement floor, wall, and ceiling data								
Total Number of Grid Cells on Map	37,168								
Required Percentages Clean Tested	90%			95%			99%		
Action Levels	**90%**			**95%**			**99%**		
	125.76			337.3			2,751		
Required Confidences Tested	90%			95%			99%		
Calculated by VSP									
Number of Samples Required	**90% confidence**			**95% confidence**			**99% confidence**		
	90% clean	**95% clean**	**99% clean**	**90% clean**	**95% clean**	**99% clean**	**90% clean**	**95% clean**	**99% clean**
	22	45	230	29	59	299	44	90	459

The results for this simulation, shown in Table B.9 are not discussed in the main report. As can be seen, most of the achieved confidences equaled or exceeded the goal confidences, indicating that the design was validated. The few exceptions could be due to the fact that the expanded data set, that included all surfaces, was more variable than just the floor data, but the same sample sizes were used for the floor only case, and the floor, wall, and ceiling case.

Table B.9. Simulation Results (achieved confidence) from Non-Parametric UTL Sampling Design. Data: Basement (floor,wall,ceiling), non-transformed

Percent Clean	**Goal Confidence 90%**	**Goal Confidence 95%**	**Goal Confidence 99%**
90% clean (AL=125.76)	93.05%	96.35%	98.91%
95% clean (AL=337.3)	89.61%	94.16%	99.48%
99% clean (AL=2,751)	90.72%	95.35%	98.95%

B.6 Conclusions

The UTL designs mostly met the validation criteria. The exceptions could be attributed to large ground truth data sets, variable in nature.

B.7 Reference

Griffith RO, JL Ramsey, PD Finley, BJ Melton, JE Brockmann, DA Lucero, SA McKenna, CE Peyton, RG Knowlton, W Einfeld, P Ho, GS Brown, and MD Tucker. 2006. *Joint Sandia/NIOSH Exercise on Aerosol Contamination Using the BROOM Tool*. SAND2006-3784, Sandia National Laboratories, Albuquerque, New Mexico.

Appendix C

Compliance Sampling Validation: Details on Sampling Design Parameters and Results of Simulation

Appendix C

Compliance Sampling Validation: Details on Sampling Design Parameters and Results of Simulation

The major details of the Compliance Sampling designs and results of the simulation are discussed in the main report. This appendix expands on the results presented in the main report and presents a few variations that were simulated during the project but were too complex and lengthy to discuss in the main report.

The Compliance Sampling plans validation had a few variants. However, the overall design for the sampling plans had the characteristics listed in Table C.1

Table C.1. Overall Design Summary

Overall Design Parameters	
Maps	Coronado Club Basement and Main Floor
Designs Tested	Compliance Sampling (high confidence few grids have contamination)
Ground Truth Source	Kriged estimates of surface concentration values calculated from 130 yellow Visolite wipe samples obtained in the joint Sandia/NIOSH Coronado Club (CC) exercise (Griffith et al. 2006)
Decision Unit	Floors, walls and ceilings of all rooms
Sample Type	Grid cell samples 0.3 meter by 0.3 meter square
Sample Placement	Random

The variants of Compliance Sampling that we tested/validated were when the ground truth matched the VSP-input design acceptable percent clean (labeled *Compliance Sampling* in the main report), and when the ground truth did not match the VSP-input design acceptable percent clean (labeled *Mismatched Compliance Sampling*). We did the mismatching to see how robust our achieved confidence results were when we used ground truth data sets that were much cleaner, and much dirtier than "expected." What was expected was a ground truth that was indeed 90% clean when we input design parameters into VSP such as "we want to be able to detect areas that are *90%* clean." In the Matched simulations, we made ground truth to be indeed 90% of the grid cells clean. *Acceptable percent clean* is one of the input parameters VSP uses to calculate the sample size.

Another slight variant for the Compliance Sampling simulations was that the action levels (ALs) used in the Compliance Sampling tests (e.g., a test might be "if any of the sample values are >AL, then declare the site to be contaminated") were slightly different than the ALs used in other sampling designs such as the upper tolerance limit (UTL) and the Hot Spot. The ALs used in the UTL tests were the percentiles of the data sets that make up the ground truth. These percentiles are shown in Table 3.2 of the main report. The ALs used in the Compliance Sampling were different by only 1 $\mu g/m^2$ or less from the actual percentiles. For example, say we wanted ground truth to be 5% contaminated. The 95th percentile was

such that 1,858 grid cells fell into the contaminated category. But this resulted in exactly 4.9989% of the cells to be contaminated. We increased the AL so that 1,859 grids were contaminated, and now we got 5.000161% of the grids to be contaminated.

Table C.2. Comparison of the Matched and Mismatched Compliance Sampling Designs

	Matched	Mismatched
Data Sets Used	CC Main floor (floor, walls, ceiling) and CC Basement (floor, walls, ceiling)	CC Basement (floor, walls, ceiling)
AL for defining clean	ALs shown in Table C.3 for main floor, and Table C.4 for basement	90% and 95% clean ALs shown in Table C.4. 85% clean: (AL = 90.75 µg/m^2, 6,035 grid cells >AL). 99.9% clean: (AL = 15,731.9 µg/m^2, 37 grid cells >AL)
Sample Sizes	Shown in Tables C.3 and C.4	Same as those used in Matched designs, shown in Table C.4

Table C.3, CC Main floor (floor, walls, ceiling), and Table C.4, CC Basement (floor, walls, ceiling) contain the design parameters and the results for the Matched cases. The Mismatched cases used the same sample sizes as the Matched cases. The Matched simulation results are shown in Figures 6.1 – 6.3 of the main report. The Mismatched simulation results are shown in Figures 6.4 – 6.6 of the main report.

Table C.3. Summary of Compliance Sampling Validation Results for the Main Floor Data Set, Matched Case[(a)]

Min % of Clean Grids	Action Level	No. of Grid Units >AL (out of 38,531)	True % >AL	90% Goal Confidence			95% Goal Confidence			99% Goal Confidence		
				n	Trials	Achieved Conf.	n	Trials	Achieved Conf.	n	Trials	Achieved Conf.
90%	646.5	3854	10.0023%	22	10,000	91.70%	29	10,000	96.96%	44	10,000	98.71%
95%	885.7	1927	5.00117%	45	10,000	91.85%	59	10,000	96.06%	90	9,999	99.22%
99%	1444	386	1.0018%	229	9,999	89.63%	297	9,998	95.13%	456	9,997	98.78%

(a) Number shown in red is the cases where the goal confidence was not achieved.
Achieved Confidence = Percentage of trials where at least one of the n grid units selected exceeded the action level.
AL = Action level.

Table C.4. Summary of Compliance Sampling Validation Results for the Basement Data Set, Matched Case

Min % of Clean Grids	Action Level	No. of Grid Units >AL (out of 37,168)	True % >AL	90% Goal Confidence			95% Goal Confidence			99% Goal Confidence		
				n	Trials	Achieved Conf.	n	Trials	Achieved Conf.	n	Trials	Achieved Conf.
90%	125.75	3717	10.00054 %	22	7,000	93.20%	29	10,000	96.46%	44	10,000	98.95%
95%	338	1859	5.00161%	45	9,998	89.93%	59	20,000	94.04%	90	9,999	99.25%
99%	2750	372	1.00086%	229	10,000	89.92%	297	10,000	94.86%	456	10,000	99.02%

(a) Numbers shown in red are the cases where the goal confidence was not achieved.
Achieved Confidence = Percentage of trials where at least one of the n grid units selected exceeded the action level.
AL = Action level.

Conclusions

There were several instances where the Achieved Confidence did not match the Goal Confidence in the Matched cases. We do not have an explanation for this, other than the basement ground truth was very spotty, with a few extremely contaminated areas, but with the majority of the areas uncontaminated. Most of the cases of under-achieved goal confidence were in the basement. One conclusion might be that if it is expected that the decision unit has extreme "spottiness," sample sizes should be supplemented somewhat to ensure the goal confidence is met. The main report contains a discussion of mismatched results.

Reference

Griffith RO, JL Ramsey, PD Finley, BJ Melton, JE Brockmann, DA Lucero, SA McKenna, CE Peyton, RG Knowlton, W Einfeld, P Ho, GS Brown, and MD Tucker. 2006. *Joint Sandia/NIOSH Exercise on Aerosol Contamination Using the BROOM Tool*. SAND2006-3784, Sandia National Laboratories, Albuquerque, New Mexico.

Appendix D

Combined Judgment and Random (CJR) Sampling Validation: Detailed Description of the CJR Sampling Design and Simulation Results

Appendix D

Combined Judgment and Random (CJR) Sampling Validation: Detailed Description of the CJR Sampling Design and Simulation Results

D.1 Description of CJR Sampling Design

The following technical description is provided because recent adjustments to the CJR model are not yet formally documented elsewhere. Additional documentation regarding the background and context of the CJR model can be found in Sego (2007). However, in addition to providing a detailed analysis of the validation results, this Appendix serves as the most recent description of the mathematics used in the CJR model.

D.1.1 Summary of Notation Used in the CJR Model

Symbol	Description	Range
θ	Probability that a high risk cell is contaminated	$0 < \theta < 1$
α	Shape parameter of the Beta distribution	> 0
β	Another shape parameter of the Beta distribution	> 0
$p(\theta)$	Beta density function with shape parameters α and β	≥ 0
r	A high risk cell is r times more likely to be contaminated than a low risk cell	≥ 1
P_J	The *a priori* probability that a single judgment sample will detect contamination	$0 < P_J < 1$
n_1	Number of judgment samples	non-negative integer
n_2	A hypothetical number of random samples	non-negative integer
n_2^*	The number of random samples which satisfies the confidence criteria	non-negative integer
N	Total number of grid cells in the decision area	positive integer
N'	A hypothetical total number of grid cells in the decision area which facilitates computation of the sample size	positive integer
X	Number of judgment samples that indicate contamination	$0,1,\ldots,n_1$
Y	Number of random samples that indicate contamination	$0,1,\ldots,n_2$
Z	Number of unsampled, lower risk grid cells that are contaminated	$0,1,\ldots,N-n_1-n_2$
λ	Desired proportion of decision area that does not contain detectable contamination	$0 < \lambda \leq 1$
C	Desired probability (confidence) that λ x 100 % of the decision area does not contain detectable contamination	$0 < C < 1$

D.1.2 Assumptions of the CJR Model

1. The total number of grid cells, *N*, in the decision area is known and each grid cell is the same size.
2. The size of the grid cell is appropriate for the chosen sampling methodology. If more than one sampling methodology is employed in a decision area, the size of the grid cell is chosen to match the sampling methodology with the smallest sampling area.
3. In the decision area, there are n_1 high risk grid cells which are more likely to contain detectable contamination if it is present. The remaining $N - n_1$ grid cells are low risk cells which are less likely to contain contamination.
4. A high risk cell is *r* times more likely to contain detectable contamination than a low risk cell.
5. All n_1 high risk grid units are sampled with judgment samples.
6. A random sample of size n_2 is taken from the low risk grid cells. The sample locations may be selected using simple random, systematic random, or adaptive fill sampling.
7. The *a priori* probability that a judgment sample location will contain detectable contamination is P_J.
8. The prior probability that a high risk cell contains detectable contamination, θ, follows a Beta distribution with shape parameters $\alpha = 1$ and $\beta = (1 - P_J)/P_J$.
9. Given the value of θ, the number of high risk grid cells containing detectable contamination follows a Binomial distribution with size n_1 and probability θ. Likewise, the number of low risk grid cells containing detectable contamination follows a Binomial distribution of size $N - n_1$ and probability θ / r.
10. The outcome will be presence/absence of contamination, as determined by the loss of detection.
11. The measurement (inspection) method correctly classifies each sample as containing (or not containing) detectable contamination, i.e., a grid unit not containing detectable contamination is never classified as containing detectable contamination (a false positive), and a grid unit containing detectable contamination is never classified as not containing detectable contamination (a false negative).
12. All samples are independent.

D.1.3 Method used in VSP to Compute n_2, the Number of Random Samples

Below we sketch an outline of the methodology used to compute the number of random samples required to achieve *C* x 100% confidence that at least λ x 100% of the selected areas do not contain detectable contamination. Additional details are available in Sego et al. 2007.

We assume $\theta \sim Beta(\alpha = 1, \beta)$, $X \mid \theta \sim Bin(n_1, \theta)$, $Y \mid \theta \sim Bin(n_2, \theta / r)$, and $Z \mid \theta \sim Bin(N - n_1 - n_2, \theta / r)$. We also assume *X*, *Y*, and *Z* are independent.

The first step is to establish a value of β which will specify the prior distribution for θ. We do this by equating the user supplied input for P_J with the expected value of θ:

$$P_J \equiv E(\theta) = \frac{1}{1+\beta} \Rightarrow \beta = \frac{1-P_J}{P_J}. \quad (1)$$

Using Bayes' theorem, we can derive the posterior predictive distribution of Z (the number of unsampled, lower risk grid cells that are contaminated) given that neither the judgment nor any of the random samples detect contamination. This distribution can then be used to determine the number of random samples required to achieve a high probability, C, that at least λ x 100% *of the grid cells* in the decision area do not contain detectable contamination. This is given by finding the smallest value of $n_2 \geq 0$ which satisfies

$$P(Z \leq (1-\lambda)N \mid X=0, Y=0) = \int_0^1 F_Z(\lfloor (1-\lambda)N \rfloor \mid \theta) p(\theta \mid X=0, Y=0) d\theta \geq C \quad (2)$$

where $\lfloor (1-\lambda)N \rfloor$ denotes the floor of (1-λ)N, i.e. the largest integer that is less than or equal to (1-λ)N, F_Z is the binomial cumulative distribution function of Z, and $p(\theta \mid X=0, Y=0)$ is the posterior density of θ given that neither the judgmental nor any of the random samples detect contamination. While the solution to (2) does yield the value n_2 which assures at least (1-λ)N *of the grid cells* in the decision area do not contain detectable contamination, it also results in strong oscillations in n_2 as N increases (Sego 2007). This undesirable behavior occurs because (1-λ)N is typically not a whole number—and rounding down to $\lfloor (1-\lambda)N \rfloor$ has a strong impact on resultant value of n_2. We avoid this problem by using instead a continuous approximation of F_Z, based on the incomplete beta function, which we call F_{Z^*}. Note that for all natural numbers w, $F_Z(w) = F_{Z^*}(w)$ and for all real numbers v, $F_Z(v) \leq F_{Z^*}(v)$. We can now re-write (2) as

$$P(Z \leq (1-\lambda)N \mid X=0, Y=0) \approx \int_0^1 F_{Z^*}((1-\lambda)N \mid \theta) p(\theta \mid X=0, Y=0) d\theta \geq C \quad (3)$$

Note that the solution to (2) is the same as the solution to (3) when (1-λ)N is a whole number. Using (3) slightly changes the inference we can draw when the judgmental and randomly placed samples do not detect contamination. Instead of concluding that we have at least C x 100% confidence that at least λ x 100% of the grid cells are free of detectable contamination, we conclude instead that we have at least C x 100% confidence that λ x 100% of the decision area is free of detectable contamination.

For a relatively small subset of the input parameters N, n_1, P_J, r, C, and λ, it is possible that the required value of n_2 decreases as N increases. This situation appears to be most likely to occur for small

values of P_J and/or large values of *r*. To ensure that n_2 is non-decreasing as *N* increases, we use (3) to identify the largest value of n_2 which occurs using values of the total number of grid cells ranging from n_1/λ up to *N*. Formally, the right hand side of (3) can be written as

$$g_{N'}(n_2) = \frac{1}{k_1 k_2} \int_0^1 \int_{\theta/r}^1 t^{(1-\lambda)N'} (1-t)^{\lambda N' - n_1 - n_2 - 1} (1-\theta)^{\beta + n_1 - 1} (1-\theta/r)^{n_2} \, dt d\theta \tag{4}$$

where

$$k_1 = \int_0^1 t^{(1-\lambda)N'} (1-t)^{\lambda N' - n_1 - n_2 - 1} dt, \quad k_2 = \int_0^1 (1-\theta)^{\beta + n_1 - 1} (1-\theta/r)^{n_2} \, d\theta \tag{5}$$

and N' represents a total number of grid cells. In order to use (4) to identify the required random sample size, n_2^*, which satisfies the confidence criteria, define

$$h_C(N') = \begin{cases} 0 & \text{if } g_{N'}(0) \geq C \\ g_C^{-1}(N') & \text{otherwise} \end{cases} \tag{6}$$

where $g_C^{-1}(N')$ denotes the value of $n_2 > 0$ such that $g_{N'}(n_2) = C$, provided $g_{N'}(0) < C$. Then the required random sample size, n_2^*, is given by

$$n_2^* = \left\lceil \max_{N' \in (n_1/\lambda,\ N)} h_C(N') \right\rceil \tag{7}$$

where the ceiling function, $\lceil x \rceil$, gives the smallest integer that is greater than or equal to *x*.

D.2 Detailed Description of Validation Results

Two approaches were used in the validation of the CJR sampling methodology. In the first approach, we address the question: Is the desired confidence achieved when all of the assumptions of the CJR methodology are satisfied? To answer this question, a large calculation and simulation study was conducted in accordance with the assumptions which underlie the CJR model. The second approach examined the impact of nullifying assumptions 3 through 9 in Section D.1.2 and simply assuming that the unknown number of potentially contaminated grid cells follows a hypergeometric distribution (which is the assumption employed by the compliance sampling methodology). For the second approach, we address the question: Given the number of samples required by the CJR method, what level of confidence can we still have if our assumptions regarding our prior belief and the risk relationship between judgmental and random samples are wrong?

Both approaches were tested using 64,561 cases. These cases covered an extensive range of the six input parameters: the total number of grid cells, *N*; the number of judgmental samples, n_1; the a priori probability that a judgmental sample would detect contamination, P_J; the risk ratio between judgmental

and randomly placed samples, *r*; the desired confidence, C; and the fraction of the decision area that is clean, λ. By testing this large number of cases, our intent was to validate the CJR methodology under practically all of the conditions in which it may be reasonably used. Figures D.1 and D.2 show 1) the distribution of the parameter values for the test cases and 2) their relationship to one another, respectively.

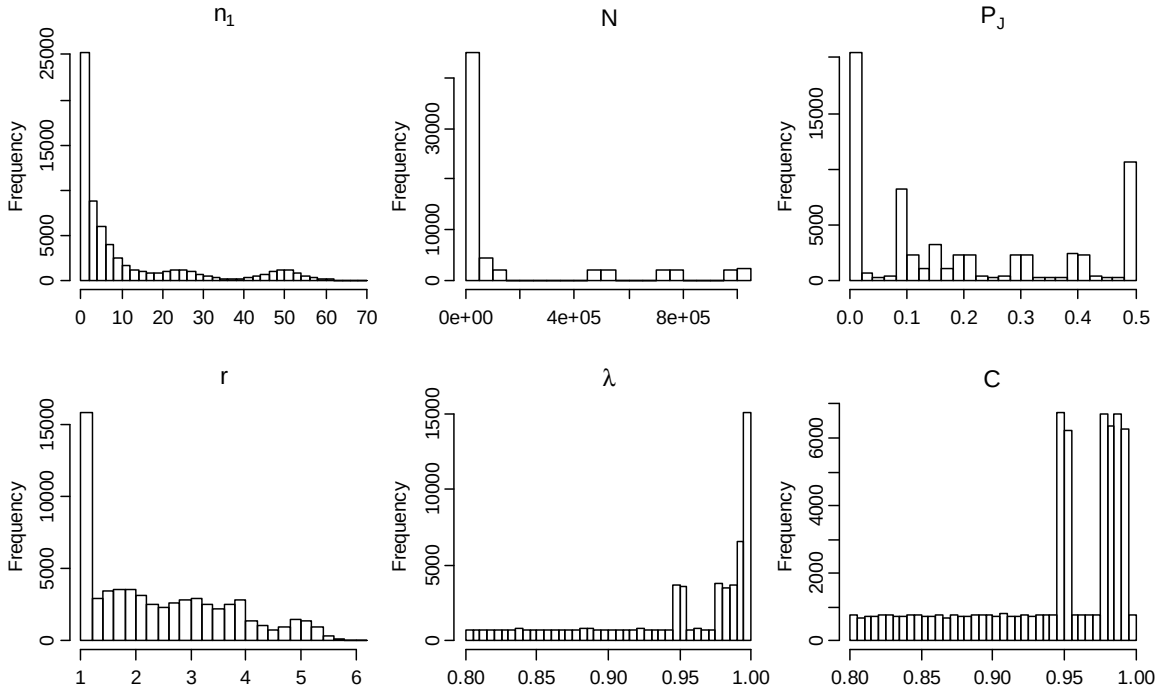


Figure D.1. Histograms which Show the Values of the Input Parameters for the 64,561 Cases

Calculating the sample size for the CJR design is computationally intensive. It involves the use of non-trivial numerical integrations and optimization routines that can be difficult to execute correctly and are prone to producing incorrect results without special modifications. Specifically, these calculations are performed as realizations of Equations (4), (5), (6), and (7) in Section D.1.3. To verify that these calculations are preformed correctly by VSP, the algorithms used to calculate these quantities were developed and implemented independently by Landon Sego and John Wilson[1]. Thus, for each of the 64,561 cases, the sample size was calculated using algorithms programmed and developed by Sego using R, a statistical programming language, and also by using algorithms and programs developed by Wilson using C++ in VSP.

[1] Landon Sego and John Wilson are researchers in the Statistical and Sensor Analytics Group at Pacific Northwest National Laboratory in Richland, WA.

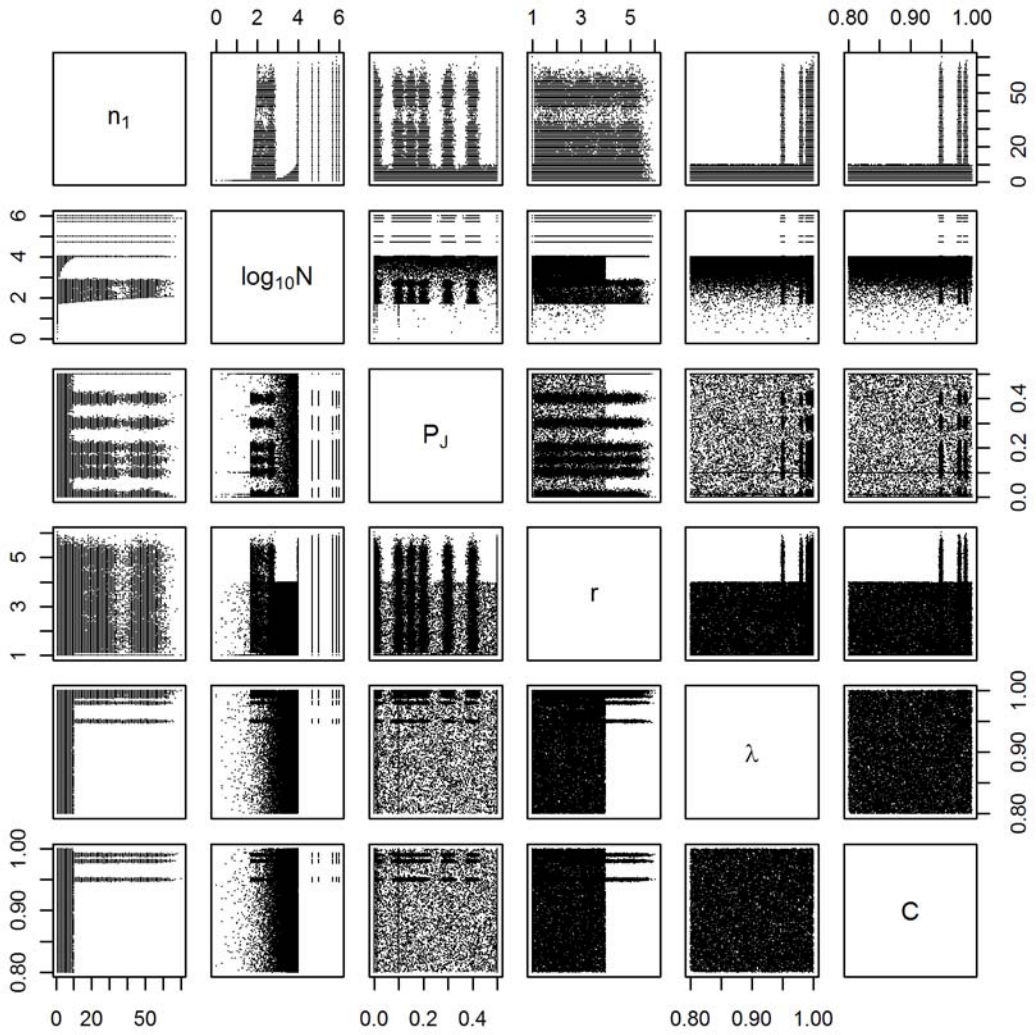


Figure D.2. Scatter Plots which Illustrate the Pair-Wise Relationships of the Six Input Parameters for the 64,561 Cases

D.2.1 Approach 1: Performance when Assumptions are Met

Two approaches were used to validate that the CJR module in VSP provides sample sizes which do, in fact, result in the stated level of confidence provided that 1) neither the judgmental nor any of the random samples detect contamination, and 2) all of the assumptions of the CJR model are met. The first approach was to perform a Monte Carlo simulation of the confidence for each of the cases by taking draws from the posterior predictive distribution. This approach is described in the following paragraphs. The second approach was to simply calculate the confidence numerically using Equations (4) and (5) in Section D.1.3. To simulate the confidence, for each of the 64,561 cases, we took 50,000 draws from the posterior distribution which arises from applying Bayes' theorem to the prior distribution and the likelihood that neither the judgmental nor any of the random samples detect contamination. Each draw from the posterior is a number specifying the probability that a high-risk judgmental sample contains detectable contamination which is then divided by *r* to produce the posterior probability that a low-risk cell contains detectable contamination. This probability is then used to take a single draw from a

continuous binomial[2] distribution, which represents the number of unsampled cells that contain detectable contamination. The simulated confidence is calculated as the percentage of the 50,000 draws which result in the number of contaminated cells being less than or equal to $(1 - \lambda)N$, the number of cells which we will tolerate being contaminated (in order to have a specified confidence that λ x 100% of the decision area is clean). The standard error for each simulated confidence was also computed, making it possible to statistically compare the simulated quantities to the exact confidence which can be calculated directly using numerical integration, as shown in Equations (4) and (5) of in Section D.1.3.

In what follows, we demonstrate that the simulated confidence agrees with the calculated confidence. This is important because it provides justification for using the more precise calculated confidence to validate the sample size calculations performed by VSP.

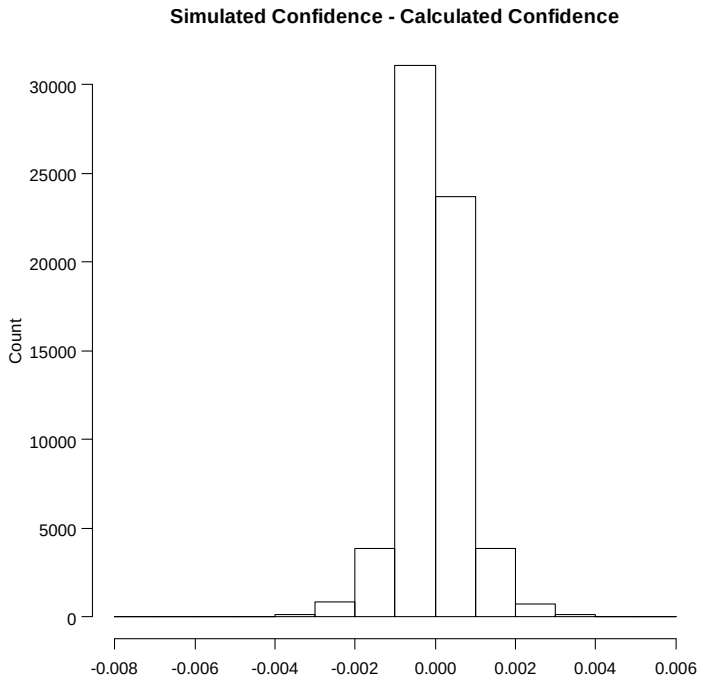


Figure D.3. Histogram of the Difference Between Simulated Confidence and Calculated Confidence for the 64,561 Cases

The difference between the simulated confidence in the calculated confidence is shown in Figure D.3. The standard numerical summary of the same data is given below:

```
     Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-7.603e-03 -2.711e-04  0.000e+00 -1.140e-06  2.762e-04  5.115e-03
```

Note that the very small minimum and maximum values suggest that the simulated confidence agrees very well with the calculated confidence. Simply stated, the simulated confidence is always within ±1% of the calculated confidence. Furthermore, we expect the simulated values of the confidence should be normally distributed, per the central limit theorem. This is clearly illustrated by Figure D.4.

[2] More precisely, a continuous approximation of the Binomial cumulative distribution function (cdf) is used instead of the discrete Binomial cdf.

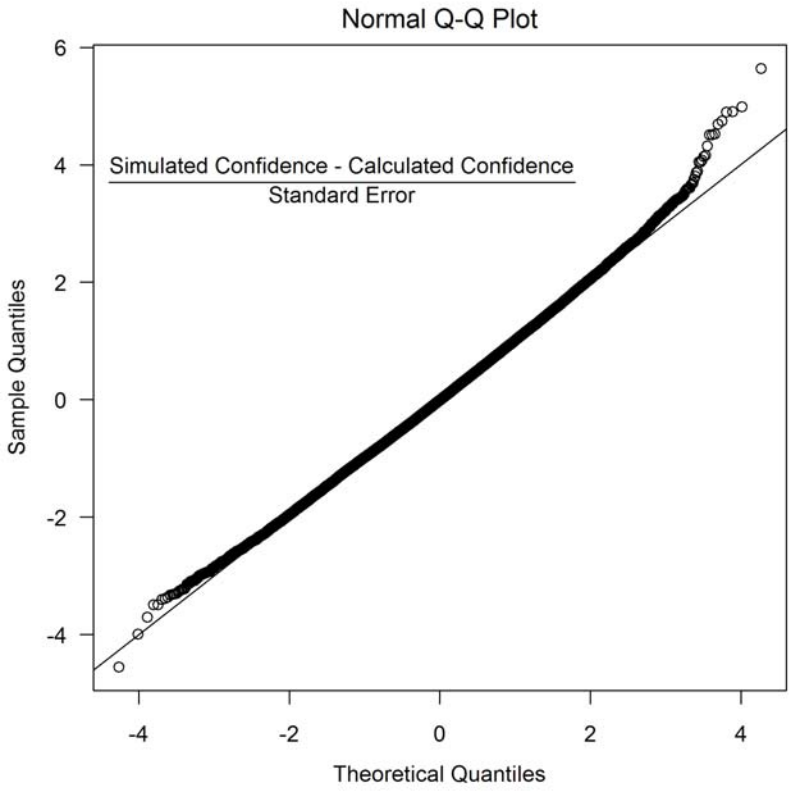


Figure D.4. Normal Q-Q Plot which Demonstrates that the Standardized Simulated Confidence Values of the 64,561 Cases Follow the Standard Normal Distribution

Consequently, after accounting for simulation error, the simulated confidence values agree as expected with the calculated confidence values. As a result, in subsequent investigations to validate the CJR module when the assumptions are met, we can use the more precise numerical calculations of the confidence, rather than the less precise simulated confidence, since both approaches have been shown to be statistically equivalent.

At this point, we distinguish between the calculated (observed) confidence and the nominal, or desired, confidence. When investigators implement a sampling design, they must specify the nominal level of confidence that they wish the sampling design to achieve. This nominal confidence is used (along with other parameters) to determine the required number of samples. Ideally, the number of required samples is the smallest sample size that satisfies the nominal confidence level. Consequently, the observed confidence achieved by the specified sample size should always be at least as large as the nominal confidence. In most cases, the observed confidence is slightly larger than the nominal confidence, because sample sizes are rounded up to the nearest whole integer.

A histogram of the difference between the calculated and the nominal confidence for the cases that required[3] random samples in order to achieve the nominal confidence level is shown in Figure D.5. Figure D.5 shows that the difference between the calculated and nominal confidence is always positive, which conclusively demonstrates that the random sample size required in the CJR module of VSP does result in achieving the nominal, desired confidence level. In approximately 1.4% of the cases that did not require random sampling, the observed confidence was substantially larger than the nominal confidence because a larger than usual random sample size was required by VSP to ensure that the number of random samples required, n_2^*, be a non-decreasing function of *N*, the total number of grid cells in the decision area[4]. For this reason, the histogram in Figure D.5 is skewed to the right.

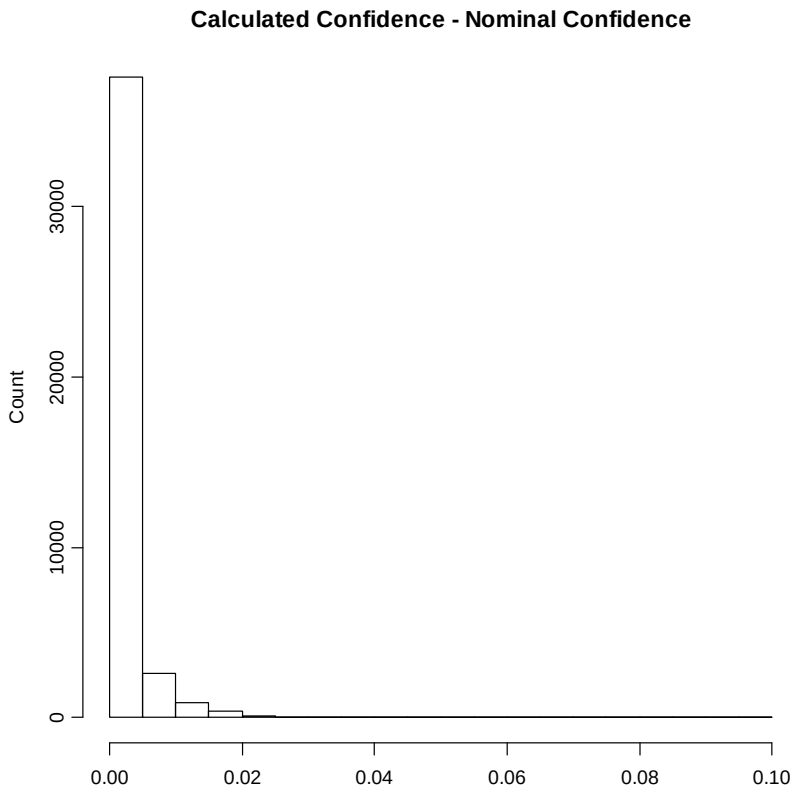


Figure D.5. Histogram of the Difference in Calculated Confidence and Nominal Confidence for Cases where Random Samples were Required to Achieve the Confidence Level

A statistical summary for the difference between the calculated and nominal confidence is shown below:

```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
3.018e-10 3.391e-05 2.289e-04 1.584e-03 1.377e-03 9.529e-02
```

[3] In approximately 35% of the cases, the parameter inputs for the CJR model resulted in no random samples being required because the goal confidence was already achieved by only taking judgmental samples. This can occur when the number of judgmental samples is large, the percentage of the decision area that needs to be clean is relatively low, the goal confidence is low, and/or the prior likelihood of contamination is low. When this occurs, VSP warns the user that the chosen CJR design may rely too heavily on the model assumptions.

[4] This issue is discussed in Section 1.1.1.1.1D.1.3 and is represented mathematically by Equations (6) and (7).

Clearly, in almost all the cases, the calculated confidence is no more than 2% higher than the nominal confidence—and the calculated confidence was at least as large as the nominal confidence for all cases.

D.2.2 Approach 2: Performance when Assumptions are Violated

The validation for this approach was conducted under the premise that some of the assumptions for the CJR method were invalid. It should be noted that all mathematical models require assumptions—and the performance of those models deteriorates when those assumptions are violated. The same is true for the CJR model. However, evaluating the CJR method when the assumptions are violated is a useful exercise in understanding the influence of the design parameters and the consequences of the worst case scenario where those input parameters are *severely* misspecified.

We calculated the confidence for each of the 64,561 cases under the hypothetical scenario that an investigator collected the number of samples specified by the CJR method and that none of the samples indicated the presence of contamination—however, unbeknownst to the investigator, 1) there really was no credible prior knowledge regarding the likelihood of a judgment sample location being contaminated, and 2) judgmental and randomly located sample locations were equally likely to contain detectable contamination. Thus, for each case, we used the hypergeometric model to calculate the confidence that would result from observing n_1 negative judgmental samples and n_2 negative random samples. (This hypergeometric model is the same one used in compliance sampling). We then compared the hypergeometric confidence to the nominal (or desired) confidence that the investigator would have chosen when originally implementing the CJR design.

Because the required sample size for the CJR method is almost always less than the sample size required by compliance sampling, we expect that the hypergeometric confidence will almost always be less than the calculated confidence. Consequently, we are most interested in understanding how the input parameters and their values influence the difference between the hypergeometric confidence and nominal confidence. In Figures D.7 through D.12, we only show the results for roughly half of the 64,561 cases because these cases had ideal, uniform coverage over the values of P_J, r, λ, and C. This was done to improve the readability of the plots. One distinguishing feature of this subset of 30,001 cases is that the maximum value for N was 10,000. However, the conclusions that we draw regarding the relationships of P_J, r, λ, C, and the difference between the hypergeometric confidence and the nominal confidence by using the smaller subset are applicable to the complete set of cases. Figure D.6 illustrates the pair-wise relationship among the six input parameters, as well as how those parameters influence the quantity

$$Diff = hypergeometric\ confidence - nominal\ confidence.$$

As mentioned previously, the value of *Diff* is almost always negative because the CJR assumptions regarding P_J and r result in smaller sample sizes than those required by the hypergeometric model for the same values of λ and C. Figure D.6 demonstrates that P_J, λ, and C do have a relationship with *Diff*, but the influence of r is not readily apparent from Figure D.6. The complex relationships of P_J, r, λ, C, and *Diff* are best understood by examining their simultaneous influence in a three-dimensional plot. Various perspectives of such 3D plots are given in Figure D.7 through D.12. A color scheme was used to represent the difference between the hypergeometric confidence and the nominal confidence. Yellow dots represent cases where the hypergeometric confidence was closest to the nominal confidence. Green, blue, and red dots represent cases where the hypergeometric confidence is progressively smaller than the

model confidence. The four color categories in these figures were chosen so that ¼ of all of the dots plotted were yellow, ¼ were green, ¼ were blue, and the remaining ¼ were red.

A number of interesting patterns are visible in the plots. However, the most pronounced, easily interpretable, patterns occur in the areas of high concentrations of yellow and red dots. In Figures D.7 through D.10, the yellow dots are most concentrated when P_J is high (i.e., near 0.50), C is high (i.e., >0.95), and λ is high (i.e., >0.95). This result is intuitive, because the high concentration of yellow dots corresponds to parameter values which give rise to larger samples sizes—sample sizes that would be very similar to those required by hypergeometric compliance sampling. The yellow dots also extend along the P_J axis—but as P_J decreases, larger and larger values of C and λ are required in order to achieve a hypergeometric confidence that differs from the nominal confidence by no more than 5%. In Figures D.7 through D.10, the red dots are most concentrated when P_J is low (i.e., near 0) and λ is high (i.e., >0.95). The high concentration of red dots correspond to parameter values which give rise to much smaller sample sizes than would be required by the hypergeometric compliance sampling, thus resulting in a much greater disparity between the hypergeometric and the nominal confidence.

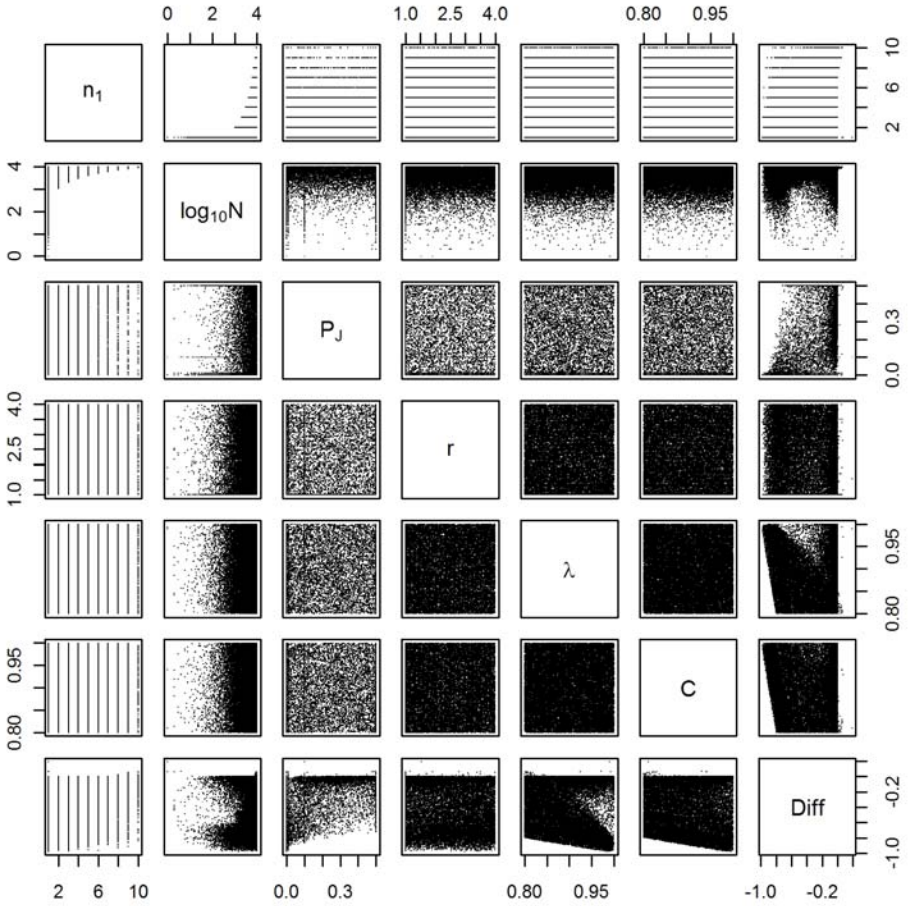


Figure D.6. Scatter Plots which Illustrate the Pair-Wise Relationships of the Six Input Parameters and the Difference ("Diff") Between the Hypergeometric and the Nominal Confidence, C, for the 30,001 Cases which had Uniform Coverage over the Values of PJ, r, λ, and C

Figures D.11 and D.12 illustrate the influence of P_J, r, and λ on *Diff*. Yellow dots are most concentrated in the region where r is relatively small (< 2.5), P_J is fairly large (> 0.25) and λ is fairly high (> 0.9). Red dots are most concentrated in the region where P_J is small (near 0) and λ is high (> 0.95).

While there is a substantial amount of overlap in the various color groups in Figures D.7 through D.12, some general trends are identifiable. Even though small values of P_J and large values of r drive down the required sample sizes for CJR designs, if the assumptions regarding P_J or r are wrong, too few samples may be taken which may result (unbeknownst to the investigator) in a confidence level much lower than the nominal level. Consequently, if the values of P_J or r are only guesses or very rough estimates, it would be prudent to err on the side of overestimating the value of P_J and underestimating the value of r. The three-dimensional plots demonstrate that some protection against misspecification of P_J and r is provided simply by requiring high levels of confidence, C, and high levels of cleanliness, λ. However, for very small values of P_J (< 0.01), requiring high levels of confidence still does not result sample sizes large enough to compensate for severely underestimating the value of P_J.

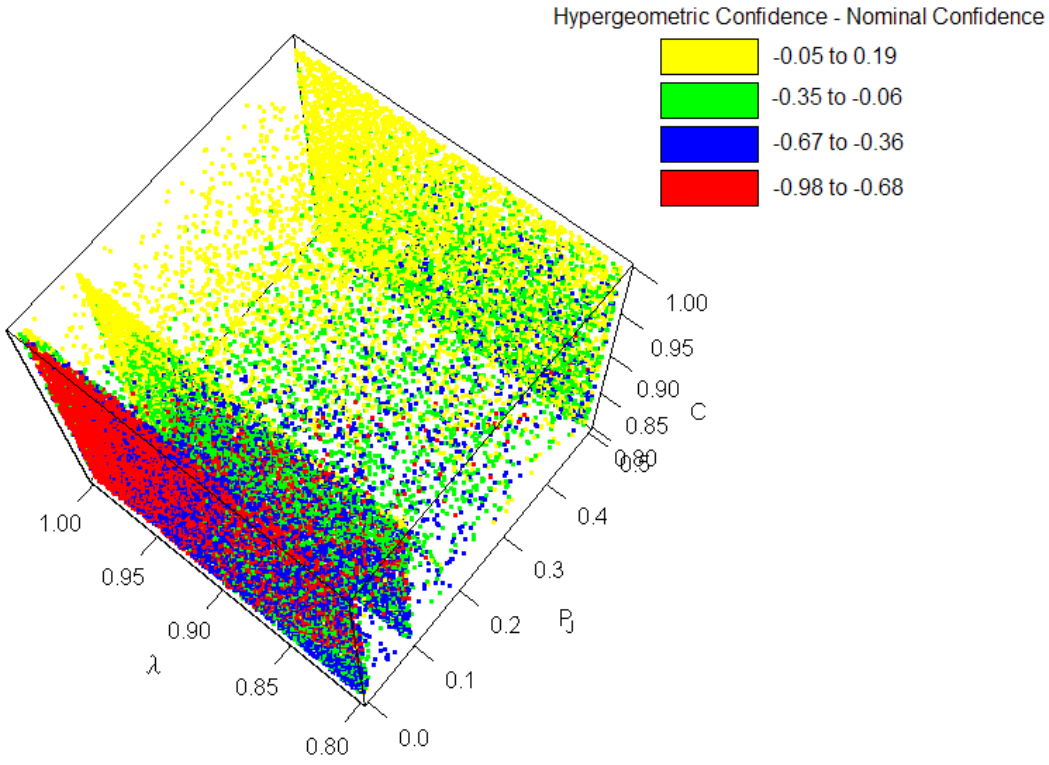


Figure D.7. Influence of λ, C, and P_J on the Difference in Hypergeometric and Nominal Confidence, Perspective 1. Color categorizations were chosen so that ¼ of all of the dots plotted were yellow, ¼ were green, ¼ were blue, and the remaining ¼ were red. The same holds for Figures D.7 through D.9.

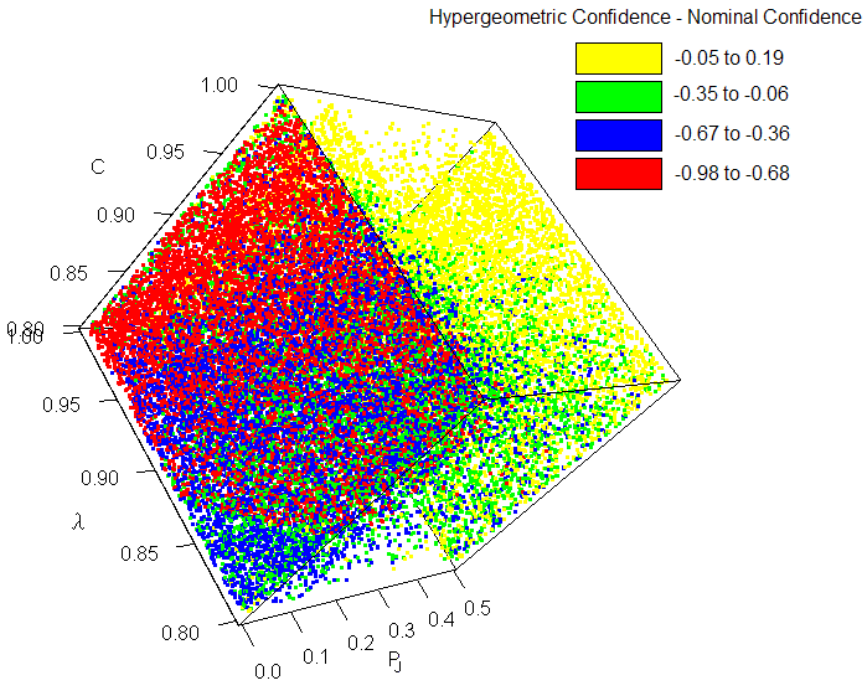


Figure D.8. Influence of λ, C, and P_J on the Difference in Hypergeometric and Nominal Confidence, Perspective 2.

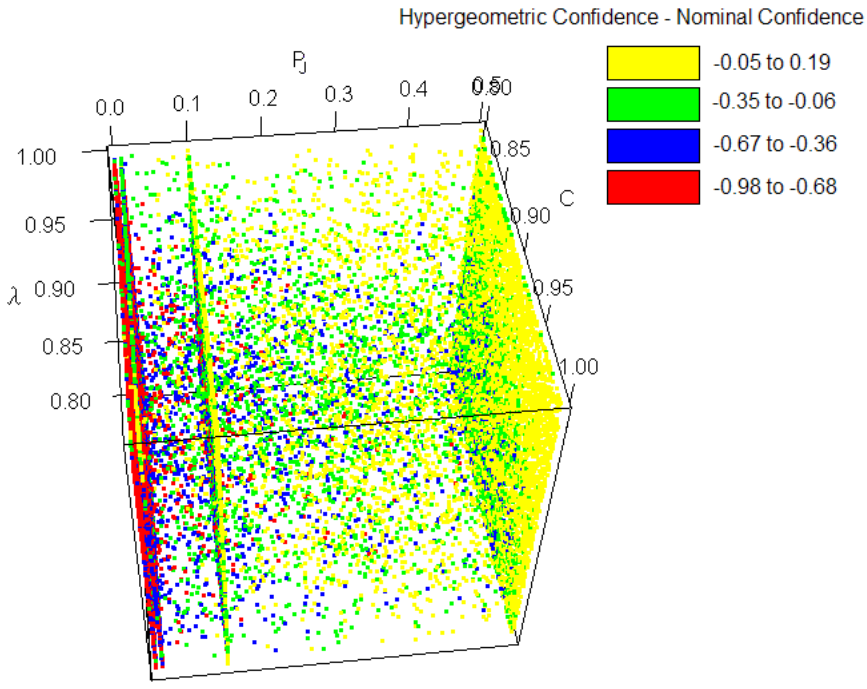


Figure D.9. Influence of λ, C, and P_J on the Difference in Hypergeometric and Nominal Confidence, Perspective 3

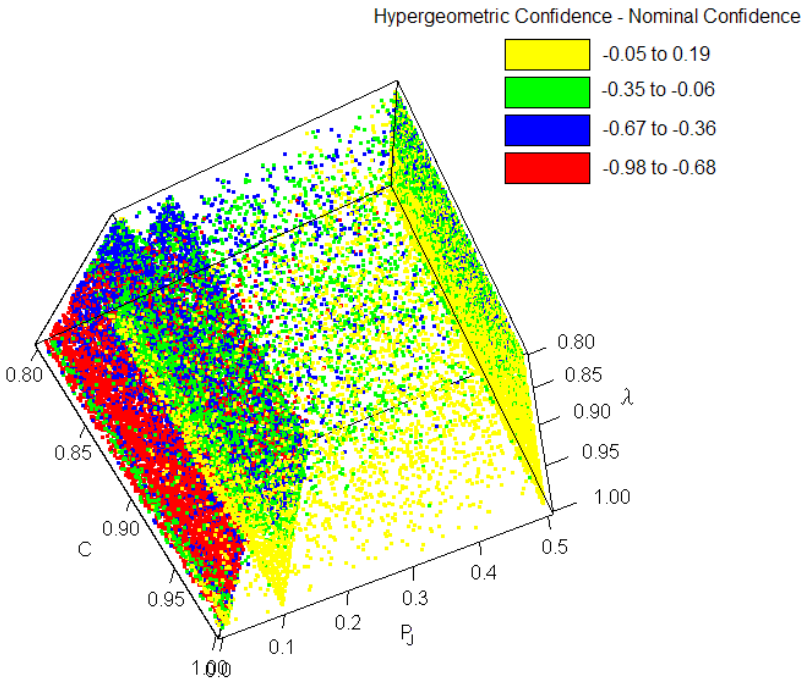


Figure D.10. Influence of λ, C, and P_J on the Difference in Hypergeometric and Nominal Confidence, Perspective 4

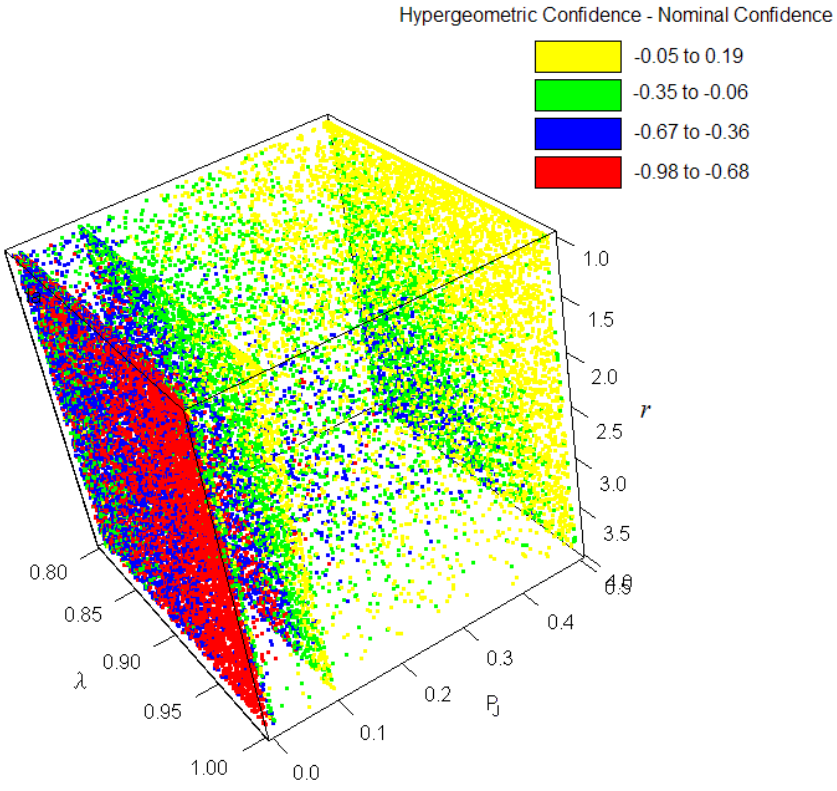


Figure D.11. Influence of λ, r, and P_J on the Difference in Hypergeometric and Nominal Confidence, Perspective 1

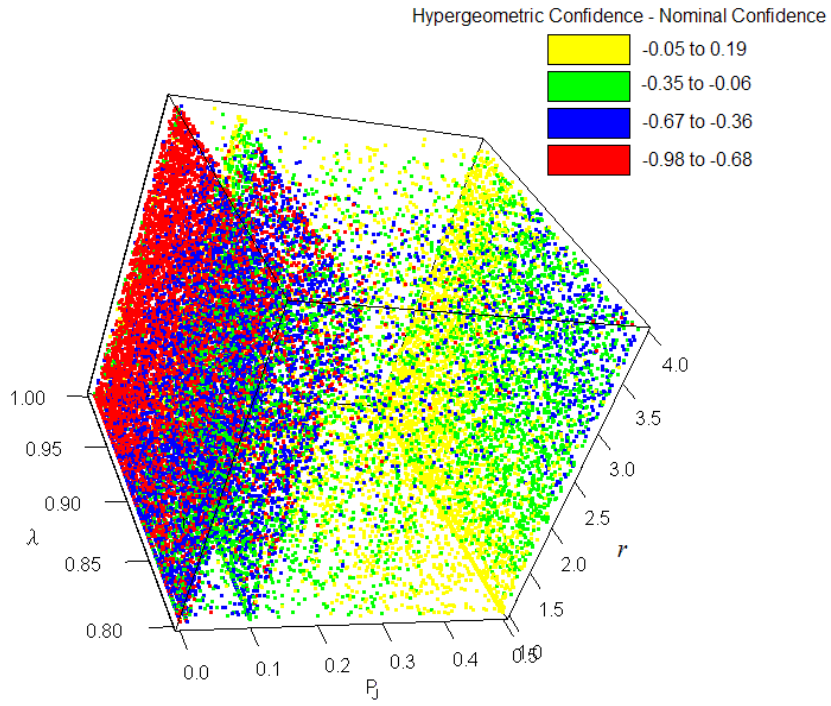


Figure D.12. Influence of λ, r, and P_J on the Difference in Hypergeometric and Nominal Confidence, Perspective 2

In addition to understanding the influence of P_J, r, λ, C, on *Diff*, it is also illustrative to consider the cumulative distribution of the *Diff* values for nearly all the cases[5]. The series of red and blue vertical lines in Figure D.13 show that:

1. Approximately 46% of the cases resulted in hypergeometric confidence that differed by no more than 5% from the nominal confidence.
2. Approximately 54% of the cases resulted in a geometric confidence that differed by no more than 10% from the nominal confidence.
3. Approximately 60% of the cases resulted in a geometric confidence that differed by no more than 20% from the nominal confidence.

Consequently, in a majority of the cases considered, the loss in confidence due to completely misspecifying the values of P_J and r was no more than 10%.

[5] In the analysis shown in Figure D.13, 2,169 cases were excluded (~3.4%) because they had values of r =1 and P_J = 0.5, which result in a model that is very closely related to the hypergeometric model. Consequently, these cases would be expected to have virtually the same hypergeometric and CJR confidence.

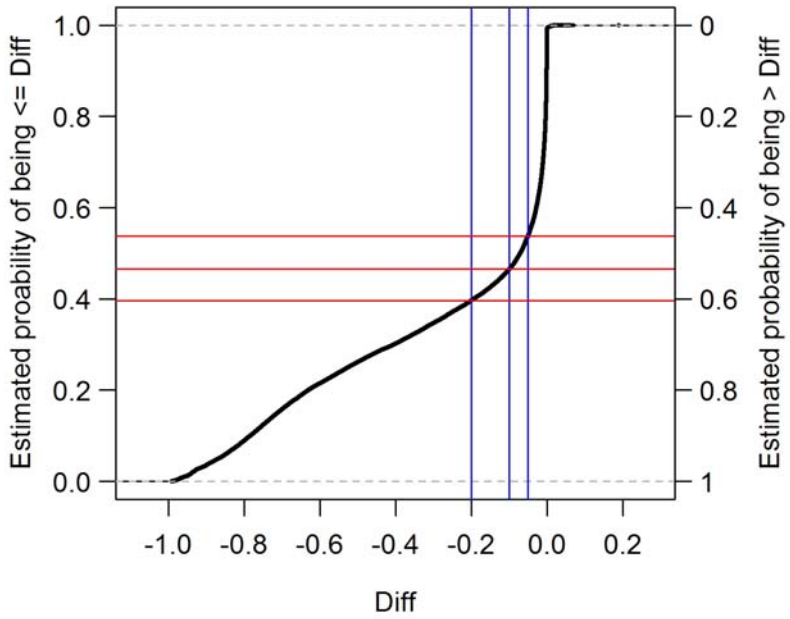


Figure D.13. Estimated cumulative distribution function of the difference in hypergeometric confidence and nominal confidence for almost all of the 64,561 cases.

D.2.3 Validating the Correct Calculation of Sample Size

Calculating the sample size for the CJR design is computationally intensive. It involves the use of non-trivial numerical integration and optimization routines that are can be difficult to execute correctly and are prone to producing incorrect results without special modifications. Specifically, these calculations are performed as realizations of Equations (4), (5), (6), and (7) in Section D.1.3. To verify that these calculations are preformed correctly by VSP, the algorithms used to calculate these quantities were developed and implemented independently by Landon Sego and John Wilson. Thus, for each of the 64,561 cases, the sample size was calculated using algorithms programmed and developed by Sego using R, a statistical programming language, and also by using algorithms and programs developed by Wilson using C++ in VSP.

Calculating the integrals in Equations (4) and (5) proved particularly challenging, since efficient numerical integration algorithms based on quadrature tended to fail in a number of cases because the integrands are practically constant over a large portion of the interval of integration from 0 to 1. This required parsing the interval of integration into subsets that would be less prone to failure. Calculating the maximization in Equation (7) also proved to be challenging, as simple optimization routines sometimes failed to find the global maximum over the interval $\left(n_1/\lambda, N\right]$, especially for large N.

Sego and Wilson independently developed and programmed the following approaches to calculate sample sizes for the CJR model:

Approach 1: Developed by Wilson and programmed in VSP, with integrations accomplished using adaptations of quadrature routines.

Approach 2: Developed by Sego and programmed in R, with integrations accomplished using Simpson's method, which is much less efficient than quadrature but relatively simple to program and highly accurate and dependable.

Approach 3: Developed by Sego and programmed in R, with integrations accomplished using adaptations of quadrature routines.

The number of required random samples was calculated using all three approaches for each of the 64,561 cases. The differences observed in the required random sample amongst the three approaches are summarized in Table D.2. For over 99.9% of the cases, all three methods agree exactly. When the methods do differ, they differ by no more than 1 random sample. The minor differences between the three methods occur under extreme conditions. Specifically, they occur when N approaches sizes of 50,000 or more and when λ is greater than 0.998, resulting in large values of n_2^* (in excess of 3,500)—and for such large sample sizes, the difference of one sample is virtually meaningless.

Table D.2. Summary of the Agreement in Sample Size Determination for the Three Approaches.

	Value of Difference in n_2^* Between the Two Approaches			Total Number of
Compared Approaches	-1	0	1	Cases
Approach 2 – Approach 3	9	64,521	31	64,561
Approach 2 – Approach 1	36	64,524	1	64,561
Approach 1 – Approach 3	1	64,502	58	64,561

The lightly shaded cells represent the number of cases corresponding to the various observed differences between the required sample sizes of the three approaches.

In summary, we conclude that the three approaches agree almost perfectly, and when they do not, the discrepancy between them can be attributed to machine error. Consequently, we have strong reason to believe that the sample size algorithms in the CJR module of VSP are coded properly and functioning as they should.

D.3 Summary of CJR Validation Results

The following points summarize results of the CJR validation:

- To explore a broad range of possible values for the CJR input parameters, the confidence was calculated for an extensive number of cases (64,561).
- Simulated values of the confidence were demonstrated to be statistically equivalent to calculated values of the confidence, thus corroborating the algorithms used to calculate the confidence exactly.
- The calculated confidence (based on the sample size indicated by the CJR module in VSP) always meets or exceeds the nominal confidence level.

- Even when the assumptions regarding the a priori probability of contamination and/or the risk ratio between judgmental and randomly placed samples is/are completely wrong, the loss in confidence was no more than 10% in more than half the cases we considered.

- The complex algorithms used to calculate the confidence and determine the required number of random samples have been correctly implemented in VSP and they agree with extensive tests against independently written code.

Distribution

No. of Copies

5 **Department of Homeland Security**
Science & Technology Directorate
245 Murray Lane SW Bldg 410
Washington, DC 20528
Elizabeth George
Don Bansleben
Erik Lucas
Lance Brooks
Randy Long

3 **United States Environmental Protection Agency**
USEPA Facilities
26 West Martin Luther King Drive
Cincinnati, OH 45268
Dino Mattorano
Oba Vincent
Larry Kaelin

5 **Center for Disease Control/National Institute for Occupational Safety and Health**
4676 Columbia Parkway
Cincinnati, OH 45226
Ken Martinez
Matt Gillen
Karl Sieber
James Bennett
Stanley Shulman

2 **Lawrence Livermore National Laboratory**
P.O. Box 808
Livermore, CA 95330
Ellen Raber
Don MacQueen

2 **Sandia National Laboratory**
1515 Eubank Se
Albuquerque, NM 87185
Mark Tucker
Bob Knowlton

No. of Copies

Oak Ridge National Laboratory
116 Virginia Road
Oak Ridge, TN 37830
Cyril Thompson

U.S. Department of Energy
1000 Independence Ave, SW
Washington, DC 20585
George Detsis

Local Distribution
Pacific Northwest National Laboratory

Deborah Gracio	(PDF)
Steve Martin	(PDF)
Cindy Bruckner-Lea	(PDF)
Larry Chilton	(PDF)
Name	(PDF)

Pacific Northwest
NATIONAL LABORATORY

U.S. DEPARTMENT OF
ENERGY

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99352
1-888-375-PNNL (7665)
www.pnl.gov