

Using Design-Based Adaptive Sampling Procedures in Site Decontamination

Myron J. Katzoff, Abera Wouhib, Joe Fred Gonzalez, Jr.
National Center for Health Statistics
3311 Toledo Road
Hyattsville, MD 20782

June 25, 2003

Abstract

We consider the application of finite-population design-based sampling procedures in a spatial context to decontamination of a site where there is a significant public health risk of anthrax exposure. Through computer simulation, we study the properties of adaptive sampling procedures employed in the search of a bounded three-dimensional space that serves as a model of the site. For a finite set of designs, we compare the operational efficiency of procedures, as measured by percent of contamination eliminated, and examine the variation in coverage proportions with choices of initial sample selection parameters, cloud-density and design complexity.

1 Introduction

In this paper, we report some initial results from computer simulation studies of sampling procedures that could be useful in the removal of anthrax from buildings. The anthrax pathogen is transported by spores which produce active bacteria when they come into contact with animals or humans. In our simulations, we have conceived the intentional release of spores as occurring at “point sources” from which they spread into a closed environment. The location of the point-sources could very well depend upon the use of the structure or uses of portions of the structure (*e.g.*, mail sorting, business-type offices and so forth). For this reason, the mechanisms by which spores are dispersed could be the air-flow in a ventilation system (or other air movement attributable to air-pressure differences between building units), inter-office mail distribution or, even, ordinary foot traffic.

The general goals of our statistical systems and methods development are the synthesis of sampling procedures or plans which:

- (1) provide for a mathematically correct estimation of risk;

- (2) result in significant improvements in the numbers of sample units contained in the final sample that are important because they have lethal concentrations of a contaminant; and
- (3) take advantage of statistical information about particle distributions that is available before or becomes available after selection of an initial sample (in which case, the sampling procedure is called adaptive).

The following collection of modules for computer simulations evolved from our early study of a few of these plans:

- (1) define and label sampling units;
- (2) create spore-clouds or clusters;
- (3) select an initial sample;
- (4) define and apply criteria for linking and adding sampling units to the initial sample; and
- (5) formulate and quantify estimators

This modular structure has already proven serviceable in studying several plans and we anticipate that this will continue to be the case for our still growing collection of plans. In particular, we expect that it will also be helpful in the exploration of adaptive procedures in which more refined use is made of the information in the initial sample through the application of advanced statistical methods to that data. One motivation for such refinements is, of course, increasing the odds of capturing the “important” units in the sample. The monetary and psychological costs associated with the time delays in completing decontamination and the need of the skills of highly trained personnel are additional motivation for concentrating effort on extracting as much useful information from initial sample data as possible.

2 Detailed Descriptions of the Computer Simulation Modules

In the first two subsections of this section, we describe modules that we have repeatedly used for (1) defining and labeling sampling units and (2) creating spore clusters. These two modules have been used without change in all of the sampling plans that we have studied to date. We indicate by example what is needed to specify the selection of an initial sample and to formulate and quantify estimators. The modules for carrying out these tasks are the ones that we have changed from one simulation to the next. We want to call attention to the fact that, up to this point in our exploratory work, we have defined and applied only one criterion for linking and adding sampling units to the initial sample. We began our work with this narrower approach hoping to get a better understanding of the statistical properties of several of the sampling procedures

that one might consider for anthrax remediation than might otherwise be possible. In future work, we will want to consider criteria for adding sampling units that embody a broader conceptualization of the term “adaptive”.

2.1 Mathematical Representation of the Building and Its Subunits

We assume that our “building” or structure is a three-dimensional rectangular parallelepiped throughout which we impose a grid of smaller rectangular parallelepipeds to represent its subunits. For the purpose of identifying and labeling the subunits, we refer to a rectangular coordinate system with its origin positioned at one corner of the lower plane of the building. Let L_x denote the total width of the building; L_y , its length; and L_z , its height. In what is intended to be an obvious notation, let N_x , N_y and N_z be the numbers of increments of uniform length in each dimension of the grid with widths δx , δy and δz respectively. The following relationships must hold

$$\begin{aligned}\delta x &= L_x/N_x \\ \delta y &= L_y/N_y\end{aligned}$$

and

$$\delta z = L_z/N_z.$$

We have thus described a system of $N_x \cdot N_y \cdot N_z$ volumetric subunits such that each has volume $\delta x \cdot \delta y \cdot \delta z$. Each such unit can be uniquely associated with a point (k_x, k_y, k_z) where k_x , k_y and k_z are nonnegative integers such that $0 \leq k_x \leq N_x - 1$, $0 \leq k_y \leq N_y - 1$ and $0 \leq k_z \leq N_z - 1$ and the corresponding grid-unit has these vertices:

$$\begin{array}{cc} (k_x, k_y, k_z) & (k_x, k_y, k_z + 1) \\ (k_x + 1, k_y, k_z) & (k_x + 1, k_y, k_z + 1) \\ (k_x + 1, k_y + 1, k_z) & (k_x + 1, k_y + 1, k_z + 1) \\ (k_x, k_y + 1, k_z) & (k_x, k_y + 1, k_z + 1). \end{array}$$

2.2 Creation of Spore Distributions

This subsection describes a procedure for creating “clouds” or clusters of spores in the building. For this purpose, we draw random samples from trivariate normal distributions constructed in the manner described in the paragraphs that follow.

The location parameter for a cloud is determined by this simple approach:

Let r_x , r_y and r_z be independent random selections from a $U(0, 1)$ distribution. Let

$$\mu_x = r_x \cdot L_x$$

$$\begin{aligned}\mu_y &= r_y \cdot L_y \\ \mu_z &= r_z \cdot L_z\end{aligned}$$

The vector $\boldsymbol{\mu} = (\mu_x, \mu_y, \mu_z)'$ is the mean vector or location parameter for the trivariate normal distribution.

This procedure is repeated for as many clouds of spores as we choose to generate.

For the next step of the process, let \mathbf{A} be a predefined lower-triangular matrix. That is, let

$$\mathbf{A} = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

If v_1 , v_2 and v_3 are independent normal random variables with zero means and unit variances, we generate a point $(x, y, z)'$ from a trivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}\mathbf{A}'$ when the coordinates of the point conform with

$$\begin{aligned}x &= a_{11} \cdot v_1 + \mu_x \\ y &= a_{21} \cdot v_1 + a_{22} \cdot v_2 + \mu_y \\ z &= a_{31} \cdot v_1 + a_{32} \cdot v_2 + a_{33} \cdot v_3 + \mu_z\end{aligned}$$

If N_s is the size of (*i.e.*, number of points in) a spore cloud, we repeat this process N_s times. This method of generating trivariate normal random variables provides a very convenient way to control variation in the creation of the variables since the determinant of the covariance matrix is the generalized variance and is just the square of the product of the diagonal elements of \mathbf{A} .

One slight refinement concerning the position of $(x, y, z)'$ is necessary: we must have

$$\begin{aligned}0 &< x < L_x \\ 0 &< y < L_y \\ 0 &< z < L_z\end{aligned}$$

Therefore, whenever any one of these conditions is violated, we generate additional points in the above manner until we have N_s points each of which satisfies these conditions.

As we generate each point at the second step of the process, we keep a running total of the numbers of points contained in each parallelepiped subunit (or cell) of the building. The cell in which we should find the point $(x, y, z)'$ is (k_x, k_y, k_z) where

$$\begin{aligned}k_x &= [x/\delta x] \\ k_y &= [y/\delta y] \\ k_z &= [z/\delta z]\end{aligned}$$

where $[\cdot]$ denotes the function which extracts the integer part of the number enclosed in square brackets.

2.3 Initial Sample Design and Selection: An Example

Suppose that the initial sample is to be drawn in accordance with a three-stage cluster design with simple random sampling at each stage of selection so that there are n_x first stage units, n_y second stage units and n_z third stage units. To select a three-stage sample of size $n_x \cdot n_y \cdot n_z$ employing simple random sampling at each stage, we proceed the same way at each stage of selection. Let $n = n_x, n_y$ or n_z and $N = N_x, N_y$ or N_z as appropriate for the dimension in which sample selection is to be considered. Then, in each dimension:

- (a)select the first increment with probability $\frac{n}{N}$;
- (b)if that increment is chosen, select the second increment with probability $\frac{n-1}{N-1}$; otherwise, select that increment with probability $\frac{n}{N-1}$;
- (c)at the j^{th} increment, if k_j increments have already been selected, choose the j^{th} increment with probability $\frac{n-k_j}{N-j+1}$.

The result of applying this procedure will be to produce a random sample of subunits with inclusion probabilities $\frac{n_x \cdot n_y \cdot n_z}{N_x \cdot N_y \cdot N_z}$.

2.4 Adaptive Addition of Subunits to the Initial Sample

The literature on anthrax indicates that an exposure of 2,000 - 2,500 spores may be enough to cause anthrax in humans. Our simulation studies make use of the notion of a critical minimum number (CMN) of spores but, solely for the purpose of keeping the computational burden manageable, we have used a smaller value. In any case, subunits that share a common boundary with an initial sample unit are added to the sample when the initial sample unit contains at least the CMN spore count. The newly added subunits are then examined to determine if they have a spore count that exceeds or equals the CMN and for each subunit that does, the subunits sharing a common boundary with them are added to the sample. This process continues until no further subunits are added.

The geometry chosen for the building and subunits within the building was especially fortunate because it facilitates describing the relationship between subunits of sharing a common boundary. Let the point which corresponds to the unit of interest be (k_x, k_y, k_z) . (Subsequently, we will refer to the unit in an obvious short-hand as the (k_x, k_y, k_z) -unit.) Assume that $0 < k_x < N_x - 1$, $0 < k_y < N_y - 1$ and $0 < k_z < N_z - 1$. In this case, no side of the (k_x, k_y, k_z) -unit has points in common with a boundary surface of the building. The six units which share a common boundary with the (k_x, k_y, k_z) -unit are the ones corresponding to these points:

$$(k_x \pm 1, k_y, k_z)$$

$$(k_x, k_y \pm 1, k_z)$$

$$(k_x, k_y, k_z \pm 1).$$

If the (k_x, k_y, k_z) -unit has points in common with a boundary surface of the building, then at least one of the following conditions must hold:

- (1) either $k_x - 1 = -1$ or $k_x + 1 = N_x$
- (2) either $k_y - 1 = -1$ or $k_y + 1 = N_y$
- (3) either $k_z - 1 = -1$ or $k_z + 1 = N_z$

Whenever one of these conditions holds, there is no corresponding grid-unit that would share a common boundary with the (k_x, k_y, k_z) -unit. For example, if $k_x - 1 = -1$, the $(k_x - 1, k_y, k_z)$ -unit does not exist. Likewise, if $k_x + 1 = N_x$, the $(k_x + 1, k_y, k_z)$ -unit does not exist. (Note that in each of these cases, if the rectangular parallelepiped subunits did exist they would have to lie outside the boundary surfaces of the building.) To state this more concisely, consider the point (k'_x, k'_y, k'_z) . If any of the following three conditions holds, then there is no subunit of the grid corresponding to that point:

- (1) $k'_x = -1$ or N_x
- (2) $k'_y = -1$ or N_y
- (3) $k'_z = -1$ or N_z .

2.5 Estimators: Continuation of The Example

To avoid unnecessary complexity in notation, in this section, we use (i, j, k) and (i', j', k') in places where we might have used (k_x, k_y, k_z) and (k'_x, k'_y, k'_z) , respectively.

Let the three-tuple (i, j, k) denote a subunit, box or cell of the initial sample of selection units. By ψ_{ijk} we denote the collection of subunits that include (i, j, k) and all of the subunits added to the sample which contain at least the CMN spore count. We refer to ψ_{ijk} as the network that includes unit (i, j, k) of the initial sample and let m_{ijk} denote the number of units in that network. By convention, $m_{ijk} = 1$ when, and only when, the (i, j, k) -unit contains no spores, in which case we have no further interest in it. Therefore, let us consider what is to be done when m_{ijk} is a positive integer greater than one. In that case, for each characteristic that is of interest, form

$$w_{ijk} = \frac{1}{m_{ijk}} \sum_{(i', j', k') \in \psi_{ijk}} y_{i' j' k'} \quad ,$$

where $y_{i' j' k'}$ is the sampling unit value corresponding to the population characteristic of interest. An estimator of the population mean is then

$$\hat{\mu} = \frac{1}{n_x} \sum_{i=1}^{n_x} \frac{1}{n_y} \sum_{j=1}^{n_y} \frac{1}{n_z} \sum_{k=1}^{n_z} w_{ijk} \quad (1)$$

and an unbiased estimator of its variance is

$$\widehat{Var}(\hat{\mu}) = \frac{1-f_x}{n_x} s_1^2 + \frac{f_x(1-f_y)}{n_x n_y} s_2^2 + \frac{f_x f_y (1-f_z)}{n_x n_y n_z} s_3^2 \quad (2)$$

where

$$f_x = \frac{n_x}{N_x}, \quad f_y = \frac{n_y}{N_y} \quad \text{and} \quad f_z = \frac{n_z}{N_z}$$

and if

$$\begin{aligned} \bar{w}_{ij} &= \frac{1}{n_z} \sum_{k=1}^{n_z} w_{ijk} \\ \bar{w}_i &= \frac{1}{n_y n_z} \sum_{j=1}^{n_y} \sum_{k=1}^{n_z} w_{ijk} \end{aligned}$$

then

$$\begin{aligned} s_1^2 &= \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (\bar{w}_i - \hat{\mu})^2 \\ s_2^2 &= \frac{1}{n_x(n_y - 1)} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} (\bar{w}_{ij} - \bar{w}_i)^2 \\ s_3^2 &= \frac{1}{n_x n_y (n_z - 1)} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{k=1}^{n_z} (w_{ijk} - \bar{w}_{ij})^2 \end{aligned}$$

The reader should note that the estimators in equations (1) and (2) have the form of the classical design-unbiased estimators for a three-stage design with simple random sampling. The simple intermediate step of creating the w_{ijk} enables the development of mean-per-unit and variance estimators for an adaptive procedure of the type considered here by recasting the new estimation problem in the form of one that has been dealt with previously. The estimation of other population characteristics may not always be handled by this device.

3 Some Simulation Outputs

For the simulation results discussed in this section, we consider the situation of a hypothetical building that has ten floors on which there are, for simplicity, ten rooms of identical dimensions along each side of five corridors. By ignoring the separations of the rows of rooms by the corridors, and conveyances and

passageways for getting from one floor to another, we intend that the building conform with the rectangular parallelepiped conceptualization given in earlier sections, which enables the use of the coordinate system for locating rooms also described earlier. (Note that this structural choice facilitates a base 10 numbering of the building subunits.) [To avoid giving a “hot spot” in a corridor no chance of being tested, one might adopt the rule that the corridor area faced by the entrance of a room that is in the sample is always to be tested. Likewise, one might also require that all stairways and elevators be tested.]

The spore distribution for the simulation outputs we shall examine consists of two clouds each of which is generated according to the procedure outlined in section 2.2 with a minor deviation. We chose very large samples from each trivariate normal and simply discarded the points representing spores that fell outside the boundaries of the building. We made 200,000 random draws from a trivariate normal with one location parameter and 300,000 draws from a second trivariate normal with a different location parameter but the same covariance matrix. After eliminating points that fell outside the building, there were 479,000 points distributed over the building interior.

Before reporting our observations, we list the remaining details on the simulation runs we made:

- (1) The CMN spore count was set at 500.
- (2) For some comparisons, we considered specifications of the covariance matrix such that the determinant was either 0.49 or 7.29.
- (3) To simulate prior knowledge of the sources of contamination, we examined some consequences of choosing with certainty the subunits that contained the mean vectors.

For initial sample sizes of 36, 80, 150 and 294, we calculated estimates of mean spore-count per sample unit excluding those units for which the spore count was less than the CMN for both traditional (*i.e.*, without adaptively adding sample units) and adaptive sampling procedures. Table 1 shows results for 30 iterations of the model for an initial sample size of 150. As one might expect from other studies (as reported, for example, in [4]), the ranges of the estimates of mean values and standard deviations indicate that, under adaptive sampling, estimates of means are less spread out than under traditional sampling.

Table 1: Mean Spore-Count Estimates and Standard Deviations for Initial Samples of 150

	Without Adaptive Sampling	With Adaptive Sampling
Means	113.37 - 893.07	141.40 - 775.41
Std. Dev.	30.787 - 61.2512	21.8340 - 54.6769

For a problem like anthrax remediation, the capability of an adaptive sampling procedure to capture a large portion of the subunits containing lethal

concentrations of the pathogen is of the greatest interest. In Table 2, we show results for 30 iterations of the model for the situation where the determinant of the covariance matrix is 0.49. The “Initial” column indicates the number of subunits of the initial sample that contained more than the CMN spore-concentration; the “Adaptive” column gives the final result of adaptively adding subunits to the ones in the “Initial” column; and the “Traditional” column gives the sampling theory expansion estimate for the total number of subunits in the building that contain more than the CMN spore-concentration. The estimated means and variances at the bottom of the table are based on the 30 sample iteration values. With the same definitions of the column and row labels, the situation is much the same in Table 3 except that the subunits containing the mean-vectors for the two clouds were included with certainty in each sample. (It is intriguing to conjecture, at this point, that inclusion of the subunits that contain the point sources will ensure that the greatest portion of the “important” subunits are included in the sample.) For the cases examined in Tables 2 and 3, the true number of subunits with spore-counts exceeding the CMN was 145.

In Table 4, it is enough to show results for an initial sample of size 80 to indicate that not much changes when we allow for an increase in cloud-dispersion of several orders of magnitude. Here, the covariance matrix was constructed so that its determinant equaled 7.29.

4 Summary and Possible Future Directions

The discussion of simulation results suggests that adaptive sampling procedures might be preferred to traditional ones for the anthrax remediation problem because the final sample yield of subunits with lethal spore counts was much greater for the adaptive procedures than for the corresponding nonadaptive classical ones. The simulations also indicated that significant improvements in such coverage would accrue from knowledge of the sources of contamination. We conjecture that similar improvements would be realized from careful analyses of initial sample results even when knowledge of contamination sources is not available. It is interesting that increasing the amount of dispersion did not alter these conclusions.

Although somewhat subtle, the three-stage procedure outcomes suggest that more consistent and stable results might be achieved by employing systematic sampling and stratification. Accordingly, in subsequent research we expect to study the consequences of stratification, systematic sample selection and the use of initial sample analyses.

References

- [1] Katzoff, Myron J.; Sirken, Monroe G. and Thompson, Steven K.(2002). *Proceedings of the Section on Survey Methods Research, American Statistical Association*, pp.
- [2] Sirken, Monroe G.(1997). Network Sampling. *Encyclopedia of Statistics*. Wiley & Sons, **4**, 2977-2986.
- [3] Sirken, Monroe G. and Shimizu, Iris(1999). Population-based establishment sample surveys: the Horvitz-Thompson estimator. *Survey Methodology*, **25**, 187-191.
- [4] Thompson, S.K.(1992). *Sampling*. New York: John Wiley & Sons.
- [5] Thompson, Steven K.(1998). Adaptive sampling in graphs. *Proceedings of the Section on Survey Methods Research, American Statistical Association*, pp.13-22.
- [6] Thompson, S. and Frank, O.(2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, **26**, 87-98.
- [7] Thompson, S.K. and Seber, G.A.F.(1996). *Adaptive Sampling*. New York: Wiley