Statistical Methods for Decontamination Sampling

Myron J. Katzoff, Abera Wouhib and Joe Fred Gonzalez, *Jr.*
Office of Research and Methodology
National Center for Health Statistics (NCHS)

James Bennett, Stanley Shulman and William K. Sieber
National Institute for Occupational Safety and Health (NIOSH)

December, 2005

Problem Statement

Suppose that a biological agent (for example, anthrax) is released in a closed environment such as an office building. What types of (statistical) spatial sampling procedures might we use in decontaminating the building?

- This problem is similar to that of choosing a spatial sampling design in the removal of hazardous materials in environmental statistics

- As in environmental statistics, we might consider adaptive and multiphase sampling

- For now, we have decided to concentrate on the application of adaptive sampling procedures

Statistical Approaches to Adaptive Spatial Inference

• Design based - the building is "subdivided" in some natural or geometrically convenient way (*e.g.*, rooms, offices, work-areas, a three-dimensional grid, *etc.*) into subunits to create a finite population from which samples are drawn in accordance with a probability structure chosen by the sampler. (Inference based on structure chosen.)

• Model based - utilizes parametric stochastic models for which the parameters are to be estimated. Then, given a location vector **s**, the value of the spatial variable at **s** is predicted; or a functional of the spatial process is predicted

• a mixture of the design and model based approaches

Design Based Sampling Procedures for Handling Microparticles - 1

• Provide for mathematically correct statistical inferences; estimators with smaller variances than would be obtained with traditional sampling procedures

• Are likely to result in significant improvements in the proportion of sample units captured that contain lethal concentrations of a pathogen or contaminant

• Take advantage of statistical information about particle (spore) distribution that becomes available either before or after the selection of an initial sample (makes it adaptive)

• Are compatible with well-established field practices in current use for contaminant removal

Design Based Sampling Procedures for Handling Microparticles - 2

Lessons learned from computer simulation studies and scientific/engineering literature reviews:

- Allow for direct use of the location of point sources when these are known

- Control the spread of the initial sample through extensive stratification (or, in the language of environmental sampling, the creation of large "remediation" units)

- Provide for the direct use of the physics of interzone airflow and contaminant transport in the linking of sampling units (likely to be very important in large building problems)

Characteristics of *Bacillus Anthracis* Contamination Relevant in Adaptive Sampling

• Anthrax, caused by this bacterium, has the ability to produce spores which behave like microparticles

• Spores can produce active bacteria when they come into contact with animals or humans

• In a closed (indoor) environment (like a building), a ventilation system, ordinary foot traffic or inter-office mail distribution can disperse spores into areas adjacent to a point source where an initial release occurred

• Spores/contaminants can collect on surfaces (desk tops, tables, file cabinets) and can re-aerosolize to further penetrate a work environment

Example: An Adaptive Sampling Procedure

• Assume we are going to draw a sample of $2k$ units, where k is a positive integer

• Form k strata and select two units with arbitrary probabilities without replacement in each stratum

• If a sample unit is found to contain a lethal concentration of spores, add units that share a common boundary with it to the sample; for each of those units, add units adjacent to them when they too contain a lethal concentration of spores, and so forth, until no further units can be added to the sample (linking mechanism). [Note that there are two ways in which a sample unit can enter the final sample: (1) it can be drawn in the initial sample; (2) it is adjacent to a unit that contains a lethal spore count.]

Example *continued*

In the final sample, retain only those units that were in the initial sample or that were added and contained a lethal concentration of spores. If unit j is a member of the final sample, let m_j denote the number of retained units to which it is linked in the final sample and let A_j denote the set of those units. (Given distinct units i and j in the final sample, note that if unit i is in A_j, then j belongs to A_i also. Further, we can have j in A_j but $A_j \setminus \{j\}$ can be empty, [that is, the set A_j can contain only unit j] in which case we adopt the convention that m_j is one even if that unit did not contain a lethal spore count.)

Example *continued*

- Define the variable w_j for unit j by

$$w_j = \sum_{i \text{ in } A_j} \frac{y_i}{m_i}$$

- If units i and j are members of the initial sample drawn from the same stratum with inclusion probabilities π_i and π_j, then when drawing two units per stratum,

$$\frac{w_i}{\pi_i} + \frac{w_j}{\pi_j}$$

is the contribution from the stratum to the population total under the adaptive scheme and the estimator for the total is the sum of terms of this form, one for each stratum.

The variance of this estimator is estimated by the sum of the terms of the following form for each stratum:

$$\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{w_i}{\pi_i} - \frac{w_j}{\pi_j} \right)^2$$

where

$$\pi_{ij} = \frac{\pi_i\nu_j + \pi_j\nu_i}{2 + \sum_{i=1}^{N} \nu_i} \ ,$$

N is the number of units in the stratum and $\nu_i \stackrel{def}{=} \frac{\pi_i}{1-\pi_i}$.

Example: Some Quantities That Might Be of Interest

Let

$$y_i = \begin{cases} 1, & \text{if unit } i \text{ contains a} \\ & \text{lethal spore concentration} \\ 0, & \text{otherwise} \end{cases}$$

and define variables

$$\begin{aligned} w_{1j} &= \sum_{i \text{ in } A_j} y_i \\ w_{2j} &= \sum_{i \text{ in } A_j} \frac{y_i}{m_i} \\ w_{3j} &= \sum_{i \text{ in } A_j} \frac{y_i}{{m_i}^2} \end{aligned}$$

Some Quantities That Might Be of Interest *continued*

If, for $k = 1, 2, 3$

$$\widehat{\tau}_k = \sum_{j=1}^{n} \frac{w_{kj}}{\pi_j} \ ,$$

where the summation is over j in the original sample, then $\widehat{\tau}_1$ is an estimator of the total of the sizes of the networks linked to units containing lethal spore concentrations, $\widehat{\tau}_2$ is an estimator of the number of units containing lethal counts and $\widehat{\tau}_3$ is an estimator of the number of networks composed of units containing lethal spore concentrations. Also, $\widehat{\tau}_1/\widehat{\tau}_2$ is an estimator of the average network size for units containing lethal spore counts.

Some Quantities That Might Be of Interest *continued*

Finally, if we redefine y_i to be

$$y_i = \begin{cases} \text{spore count,} & \text{when the unit contains} \\ & \text{a lethal count of spores} \\ 0, & \text{otherwise} \end{cases}$$

and put

$$w_{4j} = \sum_{i \text{ in } A_j} \frac{y_i}{m_i} \ ,$$

then $\widehat{\tau}_4 = \sum_{j=1}^{n} \frac{w_{4j}}{\pi_j}$ is an estimator of the total spore count in units containing lethal concentrations of spores and $\widehat{\tau}_4/\widehat{\tau}_2$ is an estimator of the average number of spores in units containing lethal counts. Also, $\widehat{\tau}_4/\widehat{\tau}_3$ is a similar average over networks.

A Few (Transitional) Comments

- Use of the "proximity" of units to link them may be very effective in situations where there is little or no information concerning where to look next

- Better choices of "linked" units may be made when there is explicit knowledge of the dynamics of airflows and particle transport mechanisms of the environment under study as could be determined through CFD studies

- Some studies have posited approximately 70% of particle transport through airflow and 30% as a result of tracking, resuspension and redeposition

Computational Fluid Dynamics (CFD)

FOCUS: mathematical (deterministic) models of airflow and particle transport

• CFD is quantitatively predicting what will happen when fluids (such as air, water, etc.) flow, with such complications as mass transfer and mechanical movement. (Provided by Concentration Heat and Momentum, Ltd. – a British software development firm)

• CFD is the process of modeling fluid flows by numerically solving the governing partial differential equations or other mathematical equations of motion. (CSIRO)

Both of these definitions capture relevant aspects of our application of CFD

A Few Points to Note on Building Interior Flow and Transport Models

• CFD models for buildings have been a scientific research topic since at least 1960

• Up to 1996, one emphasis for buildings appeared to be on developing accurate information on airflow patterns and rates for providing good indoor air quality and calculating space-conditioning loads

• Since 1996, DoE has supported a multi-laboratory effort to improve response to terrorist attack through its Chemical and Biological Nonproliferation Program. We are interested in the mathematical models that have been developed for predicting what happens to gases and particles as they travel through buildings, subways and urban areas

A Few Points to Note on Building Interior Flow and Transport Models *continued*

• Up to now, much analysis for buildings has relied upon lumped-parameter models instead of models based on "first principles" approaches to fluid dynamics because of the extremely high computational demands of fluid dynamics models

• Many of the lumped-parameter models in current use are of the multizone (MZ) type. These models describe a building as a collection of well-mixed spaces (or zones) connected by discrete flow paths. A zone may, for example, correspond to a single room, a portion of a room or several well-coupled rooms. Using zones, most buildings have been characterized as MZ structures

even when no internal partitions have been present (*e.g.*, postal facility processing centers, airplane hangers)

• Different from the MZ idea is the control-volume method of CFD which is under detailed study at NIOSH. This method involves division of the physical space into a large number of discrete control volumes called cells. The Navier-Stokes equations that govern fluid motion are then integrated over each control volume to form simplified algebraic equations

General Form of Governing Equations

Following S.V. Patankar, *Numerical Heat Transfer and Fluid Flow* (p.15), the momentum, energy and mass conservation equations all obey a differential equation of the form

$$\frac{\partial}{\partial t}(\rho\phi) + div(\rho\vec{v}\,\phi) = div(\Gamma_\phi \nabla\phi) + S_\phi \ ,$$

where ϕ is a general scalar function, ρ denotes fluid density, $\vec{v}$ is the velocity vector, S_ϕ denotes the source of ϕ per unit volume and Γ_ϕ is the diffusion coefficient for ϕ.

If dV denotes the infinitesimal for volume and $d\vec{A}$ is the surface infinitesimal multiplied by the unit vector normal to the surface, then, integrating this differential equation on both sides over the

cell volume and applying the divergence theorem of Gauss, yields the steady-state conservation equation

$$\int_A \rho \overrightarrow{v} \phi \cdot d\overrightarrow{A} = \int_A \Gamma_\phi \nabla \phi \cdot d\overrightarrow{A} + \int_V S_\phi dV \ .$$

The discrete form of the last equation is then just

$$\sum_f^{N_{faces}} v_f \phi_f A_f = \sum_f^{N_{faces}} \Gamma_\phi (\nabla \phi)_n A_f + S_\phi V \ ,$$

where f is the cell face index, N_{faces} is the number of faces enclosing cell, ϕ_f is the value of ϕ convected through face f, v_f is the mass flux through f, A_f is area of face f, $(\nabla \phi)_n$ is the magnitude of $\nabla \phi$ projected on the normal to f and V is cell volume.

Remarks on Model Equations

• Solving these equations using numerical methods will give us, for example, tracer gas concentration values at each point (centroid) of a three-dimensional grid for a ventilation chamber.

• Hence, given the opportunity to conduct physical experiments with a ventilation chamber, we can compare the experimentally measured concentration values with those obtained by solving the applicable system of equations and, thereby, validate a CFD model.

Verification and Validation

• Verification: the process of determining that a model implementation accurately represents the developer's conceptual description of the model

• Validation: the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model

(Definitions supplied by the American Institute of Aeronautics and Astronautics)

Next, we consider statistical methods for application in model validation.

Strategies and Spatio-Temporal Analysis Procedures for CFD Model Validation

- Needed: A parametric description (model) of the universe of ventilation chamber scenarios that we expect to encounter

  One possible objective: estimate the parameters of that model from a representative sample since it is impossible to examine the entire universe

- A general statistical model for CFD model validation:

  Let $\mathbf{t}$ denote a member of the space-time index set for an observable variable $Y(\mathbf{t})$ so that $\mathbf{t}$ denotes three coordinates of location and time. Assume that the coordinates

of $\mathbf{t}$ are integer-valued. We might consider the general model to be

$$Y(\mathbf{t}) = S(\mathbf{t}) + \varepsilon(\mathbf{t}),$$

where $S(\mathbf{t})$ is a "signal" process representing tracer-gas concentration values associated with index set coordinates $\mathbf{t}$ and $\varepsilon(\mathbf{t})$ denotes an error process which, under a null hypothesis, we might assume to be wide-sense stationary and have a zero mean vector. We might also assume that $\varepsilon(\mathbf{t})$ is a Gaussian process with some spectral density matrix.

A First Model

• Suppose that we have chosen a finite number of locations, N, that we are going to keep fixed throughout the analysis. In this case, we assume that our observations can be described by the basic model

$$Y(t) = S(t) + \varepsilon(t) \qquad t = 0, \pm 1, \pm 2, \ldots$$

where $Y(t)$ is an N-vector of observed values, $S(t)$ now represents an N-vector of tracer-gas concentration field values and $\varepsilon(t)$ is an N-dimensional wide-sense zero mean stationary Gaussian process with $N \times N$ spectral density matrix $f(\lambda)$.

• From the CFD model, $S(t) = S_0(t)$, the numerical solution of the system of equations describing airflow in the chamber.

Thus, one statistical formulation of the validation problem entails testing the hypothesis H: $S(t) = S_0(t)$.

• The N locations have been carefully chosen to challenge the validity of H on the basis of engineering experience with the phenomenon under study and extensive experimentation in the ventilation chamber

• To test H, one might employ discrete Fourier transforms (DFTs) and consider the dual testing problem in the frequency domain: utilize the asymptotic independence of the transformed differences $Y(t) - S_0(t)$ under H to form a test statistic which, under independence, has a known multivariate distribution

Some Points to Note

• The parts of the experimental design that are related to numbers of locations and time points for observation could benefit from a judicious application of appropriate sampling theorems

• One should be prepared to apply statistical procedures to further analyze the experimental data if H is rejected. For this purpose, provision should be made to produce enough data to estimate bispectra and trispectra (i.e., higher-order spectra) for the investigation of nonGaussianity, nonlinearity or nonstationarity

• Since ventilation chambers have finite extent, there may be edge effects at boundaries/walls which present problems that must be taken into account

Some Points to Note *continued*

• Statistical tests can be designed which exploit the properties of HOS and the properties of the equations governing fluid motion in various situations

• Note that the formulation of the PDEs of CFD used here have the property that the highest order derivatives occur linearly; hence, they constitute a quasi-linear system, which suggests that we might want to test for linearity in the error process at some point

• We have assumed Gaussianity of the error process so we might also want to formulate this as an hypothesis to be tested

Thoughts on A Test for Gaussianity:
Observations, Definitions and Items to Note

Let $\overrightarrow{X} = (X_1, X_2, X_3, X_4)'$ be a zero-mean Gaussian r.v. with covariance matrix $\Sigma = \{\sigma_{jk} : j, k = 1, 2, 3, 4\}$. Since the c.f. for $\overrightarrow{X}$ is $\Phi_X(\overrightarrow{t}) = \exp\{-\frac{1}{2}\overrightarrow{t}'\Sigma\overrightarrow{t}\}$ where $\overrightarrow{t} = (t_1, t_2, t_3, t_4)'$, if $\imath = \sqrt{-1}$, it may be seen that

$$E(X_1X_2X_3) = (-\imath)^3 \left.\frac{\partial^3\Phi_X(\overrightarrow{t})}{\partial t_1\partial t_2\partial t_3}\right|_{\overrightarrow{t}=\mathbf{0}} = 0$$

and

$$\begin{aligned} E(X_1X_2X_3X_4) &= (-\imath)^4 \left.\frac{\partial^4\Phi_X(\overrightarrow{t})}{\partial t_1\partial t_2\partial t_3\partial t_4}\right|_{\overrightarrow{t}=\mathbf{0}} \\ &= \sigma_{23}\sigma_{14} + \sigma_{13}\sigma_{24} + \sigma_{12}\sigma_{34} \end{aligned}$$

Thoughts on A Test for Gaussianity: Observations, Definitions and Items to Note *continued*

Next, let $S = \{1, 2, \ldots, k\}$ be a set of positive integers and let $\{I_p\}$ be a collection of nonempty disjoint subsets of S whose union is S. (The collection $\{I_p\}$ is called a partition of S.) If q is the number of elements in the collection $\{I_p\}$, the k-th order cumulant for $\underset{\rightharpoonup}{X}^{k\times 1}$ is given by the expression

$$\text{cum}(X_1, \ldots, X_k) = \sum (-1)^{q-1} (q-1)! \prod_{p=1}^{q} m_{\underset{\rightharpoonup}{X}}(I_p) \ ,$$

where the summation extends over all the partitions of S and $m_{\underset{\rightharpoonup}{X}}(I_p) = E(\prod_{j \epsilon I_p} X_j)$. Using what we already know when $k = 4$, it may now readily be seen that the third and fourth order cumulants of a zero-mean Gaussian r.v. are <u>both</u> zero.

Thoughts on A Test for Gaussianity: Observations, Definitions and Items to Note *continued*

Now suppose that $X(t)$ is a one-dimensional discrete-time stationary time series and let $\{\tau_j\}_{j=1}^{k-1}$ be a set of displacements from t. Since the time series is stationary, the cumulants are only functions of the τ_j and <u>not</u> t; that is, fixing the τ_j, the k-th order cumulant

$$\text{cum}(X(t), X(t+\tau_1), \ldots, X(t+\tau_{k-1}) = C_{k,X}(\tau_1, \tau_2, \ldots, \tau_{k-1})$$

for all t. If, further, $\{C_{k,X}(\tau_1, \tau_2, \ldots, \tau_{k-1})|\text{for } j = 1, 2, \ldots, k-1 \text{ and } \tau_j = 0, \pm 1, \pm 2, \ldots\}$ are absolutely summable, we can form

Thoughts on A Test for Gaussianity: Observations, Definitions and Items to Note *continued*

the discrete-time Fourier transform of the k-th order cumulants

$$S_{k,X}(\lambda_1, \lambda_2, \ldots, \lambda_{k-1}) = \sum_{|\tau_1|<\infty,\ldots,|\tau_{k-1}|<\infty} C_{k,X}(\tau_1, \tau_2, \ldots, \tau_{k-1}) \times \exp[-\imath \sum_{j=1}^{k-1} \lambda_j \tau_j]$$

called the k-th order polyspectra for $X(t)$. When $k = 3$, $S_{k,X}(\cdot)$ is known as the bispectrum and when $k = 4$, $S_{k,X}(\cdot)$ is called the trispectrum.

Since the bispectrum is zero for a zero-mean stationary Gaussian process, it might seem reasonable to anticipate that a test for

Thoughts on A Test for Gaussianity: Observations, Definitions and Items to Note *continued*

Gaussianity can be based solely on how "close" the bispectrum is to zero. However, something more may be needed. From what we have already observed, the $C_{X,3}(\tau_1, \tau_2)$ and $C_{X,4}(\tau_1, \tau_2, \tau_4)$ are all equal to zero when $X(t)$ is zero-mean Gaussian but for a zero-mean double-exponential, logistic or uniform distribution, $C_{X,3}(0,0) = 0$ while $C_{X,4}(0,0,0) \neq 0$. This might be viewed as an indication that a test for Gaussianity must be based on both the bispectrum and the trispectrum.