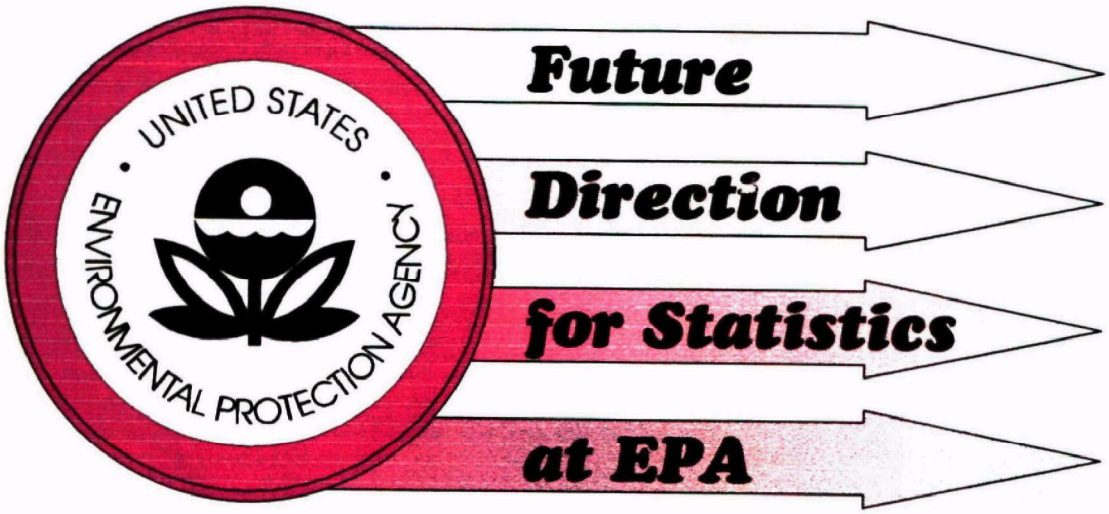


Twelfth Annual EPA Conference on Environmental Statistics

Richmond, Virginia
April 1 - 3, 1997

Chapter 40

Geostatistical Sampling Designs for Hazardous Waste Sites

George T. Flatman and Angelo A. Yfantis

This chapter discusses field sampling design for environmental sites and hazardous waste sites with respect to random variable sampling theory, Gy's sampling theory, and geostatistical (kriging) sampling theory. The literature often presents these sampling methods as an adversarial "either/or" philosophy; this chapter emphasizes when each should be used with a cooperative "both/and" philosophy. The intrasample variances, biases, or correlations must be taken care of by the use of Gy's sampling theory for both independent random variable sampling and analysis and correlated random variable sampling and analysis. The deciding factors in the choice of sampling design and analysis are not just intersample variances, biases, or correlations but also the discreteness of the waste under investigation, remediation as a unit, and the relative cost of samples versus the cost of remediation.

ENVIRONMENTAL SAMPLING is a multidisciplinary science. It requires chemists, media experts, risk assessors, and even statisticians. The sampling design is an integral part of the experimental design and data analysis, and most importantly, the data analysis cannot recover more information than the samples contain. Thus the statistician needs to be on the project from its inception. Optimal environmental sampling requires consideration of at least three branches of statistics. Classical random variable statistics (*1*) are needed in quality assurance (QA) and in the analysis of data that are reasonably independent (little or no process, spatial, or chronological correlation). Gy's theory of sampling (*2*) is needed for the definition of correctness for the "field sample" [determination of amount (mass or volume) sampled] and any samples taken in heterogeneous media (almost all environmental samples). Geostatistics, and its most used form, kriging (*3*), is needed for field sites with a spatial structure. The choice of sampling designs—

3152–4/96/0779$15.75/0

when to use classical random design or kriging's regular grid design—is a difficult decision. Even statisticians differ on such a question. This chapter discusses the statistical rules that enter into the decision. The decision depends on specifics of the site and remediation plan as well as statistical aspects. For example, Gy's theory must be used to take a correct sample for either random variable statistics (sampling or analysis) or geostatistics (sampling or analysis).

When I discussed the role of statistics in sampling design with a manager of a chemical laboratory, the manager confided in me that his statistician's recommendations were always illogical and irrational and contradicted common sense. We did not have time to discuss specifics, but I suspect the advice he received was also poor statistically because it confused the use of random variable statistics with the use of spatial statistics. If the correct branch of statistics has been chosen, statistical requirements can be explained from statistical theory in a logical and reasonable manner that does not defy common sense. It is important in a multidisciplinary project for all to be comfortable with the soundness of the decisions. Statisticians should be asked to explain the statistical requirements they recommend until all feel comfortable with the design.

Random Variable Statistics

A random variable has both magnitude and probability. It may come from a symmetric distribution such as normal or uniform, or from a skewed distribution such as lognormal or Poisson. Chemical environmental data sets are often assumed lognormal, and radioactive data sets are often assumed Poisson. Because both distributions are positively skewed, the estimate of the mean based on few samples has a higher probability of being underestimated than the mean of a normal distribution or any symmetric distribution with a strong central tendency. Random errors as monitored by QA are often assumed normal. The branch of statistics that deals with random variables gives us the statistical inferences that have tools for QA. Random variables provide measures of central tendency (such as mean, median, and mode), dispersion (such as range and standard deviation), and statistical inference (such as confidence intervals, prediction intervals, and tolerance intervals).

The mean and standard deviation are the statistics usually sought by a sampling campaign; they are sufficient statistics (i.e., completely define the distribution) for the normal distribution. The mean of any distribution becomes normal as the number of samples, n, becomes large. This property justifies the use of confidence intervals for the mean if, and only if, n is large enough ($n > 16$ for a symmetric distribution, and $n > 50$ for a skewed distribution). However, if the number of samples is much fewer than 50 samples from the typical environmental distribution in a confidence interval, then these limits are not to be trusted. Either knowledge of the distribution or transformation to normality is required for statistical inference about the variable, its distribution, or future samples. A listing of means and standard deviations or intervals, without investigating the distribution, is misleading and has the

potential of inviting wrong decisions because the readers will assume normality. Nonparametric intervals and tests are available, but they lack power. For example, the critical values for one-sided intervals for probabilities $(1 - \alpha)$ of 0.95 and 0.99 using the Tchebycheff inequality are 4.472 (square root of 20) and 10.000 instead of the standard normal distribution values of 1.64 and 2.33. Most regulators will cringe at 4 or 10 in a compliance hypothesis test. Another consideration is that random variable sampling design requires rigorous definitions of the *population* and *sampling unit,* so that the design can give each sampling unit an equal probability of being chosen. This requirement will be discussed further.

Population Defined

In environmental samples, population is not as obvious or as well-defined a term as it is in statistical textbooks (e.g., all the cards in a deck, or the two sides of a coin). In site evaluation, the most obvious population is the waste site as a whole, but the usual site has more than one population of interest. It may have population(s) of plume(s) and background population(s). The population of interest is the population(s) of the plume(s). Waste plumes seldom honor property boundaries or travel in politically defined shapes such as city blocks. Thus the populations of interest are the plume(s) and the background, not a mixture of these. To average all the samples from the site would give an estimate of a mean from a mixture of populations, a "fruit salad" of plume(s) and background(s). If the location and extent of the plume or background are not known, but a map of mean contours (isopleths or isarithmic lines) is wanted for multiple remediations, then this situation would require geostatistical sampling and analysis. If the waste to be evaluated is well-defined and confined, such as liquid waste stored in 55-gallon drums or a waste pile on a tarp that will be disposed of as a unit, then the population of interest is the drum or pile and therefore classical statistics (a mean value) will be adequate for the decision.

Sampling Unit Defined

For textbook statistics, a sampling unit is a draw of a card or a flip of a coin, but for an environmental sampling the unit is complicated by natural variation (e.g., media heterogeneity or pollutant characteristics) and sampling tool variation and biases (*4*). In laboratory QA the unit may be the contents of every ith vial in the queue of the analyzing instrument. At an environmental site, the "sampling unit" is ambiguously used to refer to both the sample and the sample support. The sample is much smaller in volume or mass than the sample support, but if it is representative, it has approximately the same concentrations of the pollutant or the same values of some measured characteristic. The sample, simple or composite, is a small critical mass that is taken from the sample support for measurement. The sample support is the larger volume or mass of in situ media that is to be represented by the measure of the sample. The sampling support is often the same volume as the remediation unit. These two units are determined by the goal of the sampling campaign or the reme-

diation option(s), but they must meet the requirements of Gy's theory of sampling and geostatistics, which are each discussed in subsequent sections of this chapter. The extractable mass or volume (field sample) cannot be dictated by the size of the sampling tool or the size of the official container. It should be determined by the heterogeneity of the media in accord with Gy's theory. Differing amounts of media of interest, because they are ambiguously called "sample", should be identified by size and use. The analysis sample (i.e., aliquot or split), used in its entirety by the chemist for analysis, has a mass less than that of the preparation sample, which has a mass less than or equal to that of the field sample. Each change of scale or reduction of sample mass must pass Gy's requirements (*see* the subsection ***Analytical Error***). The name of the sample is unimportant, but the change of mass is important. Any change in volume (mass) must be checked using a monogram made up for the current site. Extraction(s) for the field sample from the in situ sample support (i.e., sampling unit) must satisfy both Gy's theory requirements and geostatistical requirements.

Dealing with Correlation in Practice

In theory, the difference between an independent random variable and a random variable correlated in time or space is clear, but this difference is not so clear in practice. In practice, most environmental samples are correlated in either time or space, and possibly in both time and space, yet a random sampling or analysis is done. Even the analyses of the samples in the queue of a mass spectrometer (MS) are correlated somewhat in time, but this correlation is weak enough and the QA samples are spaced far enough apart that the correlation can be ignored. Correlation in space or time can be taken into account by slightly more complicated formulas in random variable statistics; Gilbert (5) gives relevant sediment and groundwater examples of how correlated sample units require more samples to be taken (larger n) than if the observations were independent. The critical criterion for using a spatial sampling and data analysis is the management decision or need to see a contour (isopleth) map of the pollutant location as well as concentration (these are kriging results) in place of a list or histogram of chemical analyses with a confidence interval about an estimate of some mean (random variable output).

Pierre Gy's Sampling Theory

Pierre Gy is a mining engineer and Francis Pitard is a chemist. Both men have had brilliant careers in process and mining quality control. Pitard has written a two-volume work (2) that captures and communicates their experiences in the sampling of heterogeneous media. These volumes are valuable for environmental sampling of soils or sediments. Pitard organizes the taking of "correct" samples with correct sampling tools, according to seven "errors". The emphasis on correct samples and tools is analogous to the emphasis from the U.S. Environmental Protection Agency (EPA)

on representative samples. Because of the potential of one, some, or all seven of these errors to erode the correctness or the representativeness of an environmental sample, this chapter will refer to them as "variances" to stress their additivity for a component-of-variance model. "Variance" emphasizes the intrinsic nature of these errors or biases in heterogeneous material sampling, in contrast to the negative connotations of these terms in the vernacular ("error" as a careless mistake; "bias" as an intentional dishonesty). *Variance, error,* and *bias* are technical terms that describe differing problems with different solutions. An "error variance" is often thought of as symmetric with a mean (expectation) of zero and as reducible by taking more samples; a "bias variance" is one-sided (e.g., always too high or too low) and is reducible not by taking more samples but only through a correct sampling design. The symmetry or one-sidedness must be carefully thought out and often field-tested for all potential variance in any sampling design and QA plan.

This theory sounds like any QA plan talking about errors, but it refers to a different type of error and needs to be discussed in its own part of the QA plan. Specifically, it deals with intrasample error (errors within the sample) rather than intersample error (errors between samples). The various components of variance of this sampling theory sound trivially obvious when pointed out, but they are easily overlooked in the stress of formulating a QA or sampling plan. Leaving them out can be disastrous for QA and data quality objectives. Even though these sources of variation sometimes are obvious and trivial, they must be taken into account in every environmental sampling plan.

The Fundamental Error

This component of variance is a natural property of heterogeneous material. It is not an error in the sense of an avoidable mistake; however, if the sample planner does not take it into consideration it will generate unnecessary (avoidable) variance in the laboratory analyses. The variance is caused by the range of particle sizes in the medium and the fact that often only certain sized particles contain the pollutant of interest. This situation is illustrated in Figure 1; the shaded or lined particles are assumed to contain or carry the pollutant, and the other particles are the heterogeneous medium. Thus the chemical analysis depends on two values: the number of solid particles (percentage composition), and their concentration. This dependence adds another variance term or component of variance (percentage composition) to the analytic variance. The magnitude of this error is small in a fine or homogeneous soil or sediment but becomes larger as the medium becomes more heterogeneous in particle size and particle affinity for the pollutant of interest. This fundamental component of variance can be reduced by increasing the mass of the sample or by reducing the particle size of the sampling material by appropriate digestion.

To maintain the original level of accuracy, the sample material must always be reduced in maximum particle size before being reduced in mass or volume (split or aliquot). The mass of a sample required for a given relative variance [relative standard deviation (RSD) squared] can be read from Pitard's nomograms as a function of

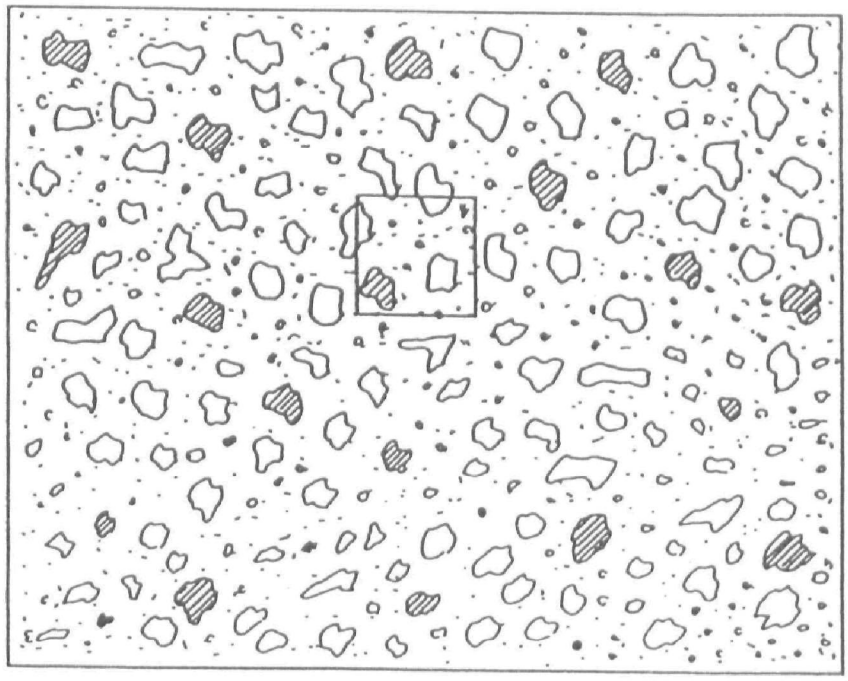


Figure 1. Heterogeneous material: fundamental error. (Reproduced with permission from reference 2, Vol 1. Copyright 1989 CRC Press.)

various physical properties, the most important one being maximum particle size of the medium (6). This relationship will be directly applicable to waste monitoring if the pollutants of interest are heavy metals, but the application to volatile chemicals or semivolatile chemicals remains to be developed. The EPA has a very readable document on this subject that presents an example nomogram for soil properties (7). The extension of Gy's theory to volatile chemicals and semivolatile chemicals is a very important but as yet undeveloped part of environmental sampling.

Grouping and Segregation Error

There is potential for this variance in any heterogeneous media. The grouping and segregation error develops through movement of samples through processing, handling, shipping, or mixing. The heterogeneity may be in density or size (also adhesion, cohesion, magnetism, affinity for moisture, and angle of repose of crystalline structure) so that the particles come together by groups during any movement or vibration. Figure 2 illustrates this type of error for the pile at the end of a conveyer belt. If the black particles contain the pollutant of interest, then a sample from the right side of the pile will be biased high and a sample from the left side will be biased low. In taking a sample of a waste stream or pile, the potential variance can be minimized by sampling along the gradient of grouping and segregation. For soil, gravel, or sediment being carried on a conveyor belt, the gradient of grouping and segregation would be across the belt orthogonal to the direction of motion, and thus a correct sample would be a rectangular (not a trapezoidal) section oriented across the belt. Sampling a pile, a truck, or a railroad car of waste in a correct manner is

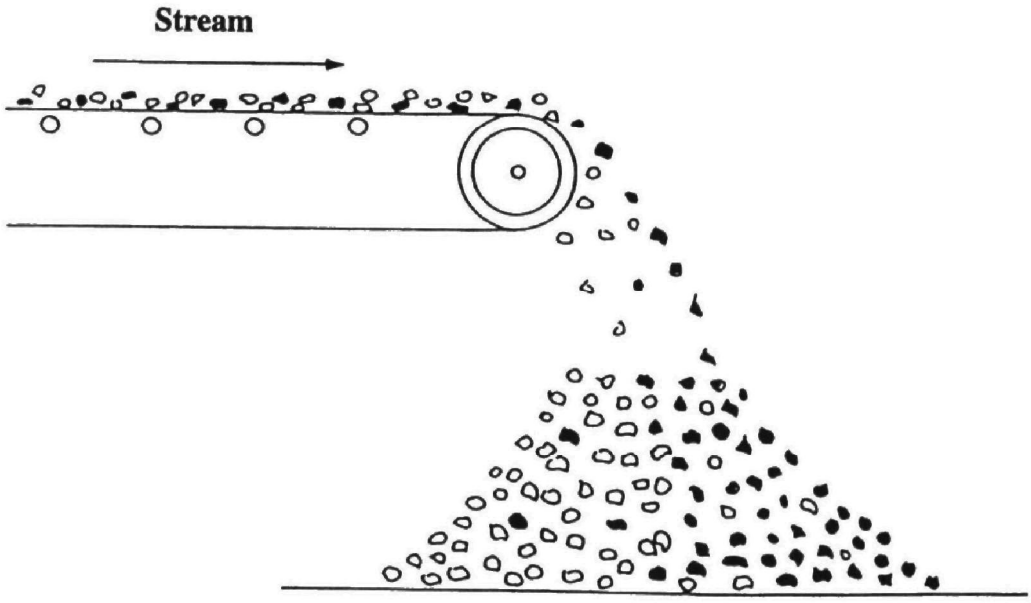


Figure 2. Grouping and segregation error. (Reproduced with permission from reference 2, Vol. 1. Copyright 1989 CRC Press.)

very difficult because of this component of variance. The correct time to sample is before the pile is built or the truck or railroad car is loaded. In sample preparation, Pitard suggests that the pouring of the well-mixed material from the V-blender, especially if the particulate material is allowed free fall of any distance, can undo (defeat) the blending (8). Aliquoting increases this error. The general rule is that as aliquot size decreases, the variance increases. Theoretically, as the size of the aliquot approaches the size of the grains of the sample, this error grows larger without bounds. The corollary to this theorem is the fact that the chemist, aliquoting to get the relatively small amount of material (analytical sample) actually required for the analysis, can turn the analytic equipment into a random number generator if the sample material has not been ground to the required fineness and aliquoted correctly.

Spatial and Periodic Errors

These error sources could be periodic and/or spatial structures on the scale of the extracted sample or the sample support (the in situ area or volume represented by the sample). If they were of a larger scale they would be studied by a time series analysis or a geostatistical analysis, but they are not of interest, and the decision statistic is the mean of the unit and not the means of the subunits. In the preceding discussion of classical statistics, the 55-gallon drum was assigned to a classical statistical analysis instead of a geostatistical analysis, even though there may have been a structure in concentration in the vertical dimension of the drum. No one wants a contour map of the concentration of pollutant inside of a drum because the drum will be remediated (disposed of) as a unit. However, this gradient cannot be ignored; instead it must be representatively sampled by sampling each layer proportional to

its volume. This sampling is accomplished by the choice of sampling tools. To minimize the microspatial variance, a "composite liquid waste sampler" (COLIWASA) must be used. The name of the sampling tool tells an important principle. Compositing is an important tool in random variable statistics to save chemical analysis costs, but in spatial statistics it is used to ensure that the sample is representative of the in situ sample support. Subsample compositing is physically doing the same thing that statistical averaging does to the numerical values of replicate samples, except compositing loses the information about the variance or standard deviation, with the benefit of saving the cost of $(n - 1)$ chemical analyses. These are two quite different and important uses of compositing.

Increment Delimitation and Extraction Errors

These two variances arise from the interaction of a sampling tool with the heterogeneity of the media sampled. The circles in Figures 3 and 4 can represent the cutting edge of a plugging or coring device descending on the media to take a soil or sediment plug or core. In Figure 3, taking the shaded area of the larger particles would be the correct sample, but if the larger particles are hard compared to the softer interstitial material, the tool will not cut through the harder particles to give the desired correct sample. Rather, the large hard particles will be pushed out of the sample if their centers of gravity lie outside the corer, as illustrated by the white par-

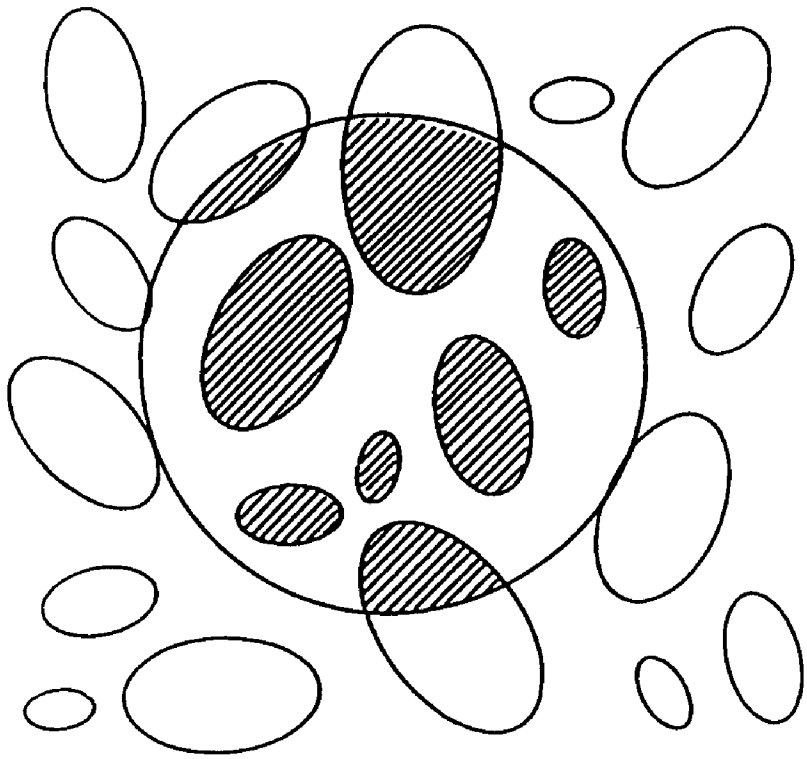


Figure 3. Increment delimitation error. (Reproduced with permission from reference 2, Vol. 2. Copyright 1989 CRC Press.)

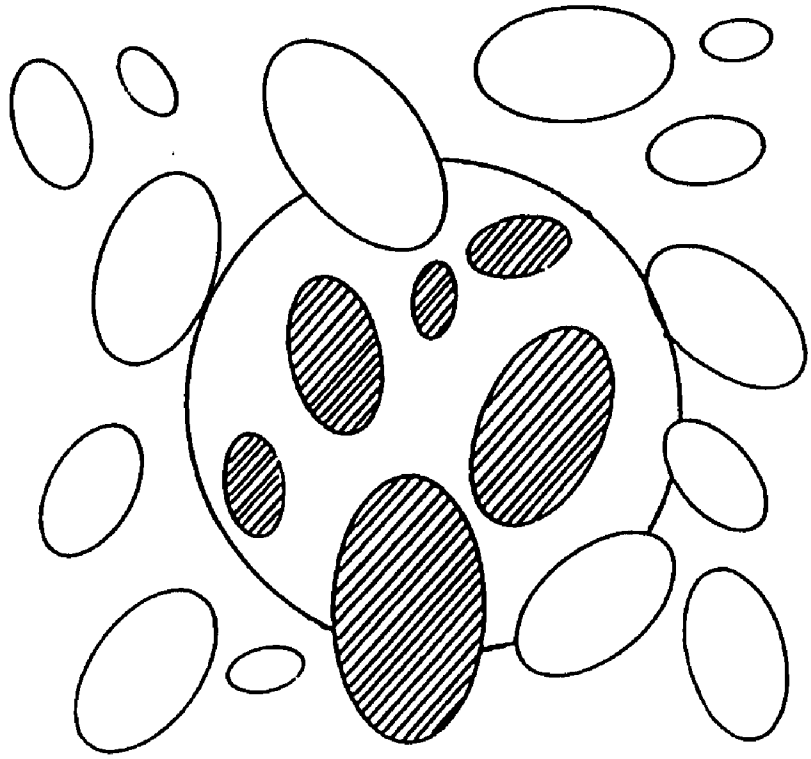


Figure 4. Increment extraction error. (Reproduced with permission from reference 2, Vol. 2. Copyright 1989 CRC Press.)

ticles in Figure 4. If their centers of gravity fall within the corer as illustrated by the shaded particles in Figure 4, then the particles in their entirety will be included in the sample. Either case is incorrect, but the two cases tend to average out. It is important to distinguish these two concepts: (1) the *delimitation error* is the variation caused by the inability to cut through all the heterogeneous media and take the part included in the circle of the coring or plugging device, and (2) the *extraction error* is the variation caused by taking or pushing out of the way the whole hard particle as a function of whether its center of gravity falls in or out of the circle of the corer or plugger. If the cylinder could be cut out exactly by a laser and then taken out intact by levitation as in science fiction, these two errors could be avoided. Today's solution to these problems is to have a corer or plugger that is at least two or three times the diameter of the largest particle size.

Analytical Error

The EPA and the American Chemical Society have published many excellent papers, proceedings, and books on this interdisciplinary subject. Therefore, to avoid duplication, we wish to speak only to the chemist's method of abstracting a much smaller sample (analytical sample) from the prepared sample. This step, because of the smallness of the mass of the analytical sample compared to the mass of the sample from which it comes, is the sample most apt to incur an unacceptable magnitude of Gy's fundamental error. If the analytical sample is taken by sticking a spatula ran-

domly into the top of the material in the bottle and taking out the desired amount, such a sample is a grab sample and not an aliquot or split; the chemical analysis is very apt to give a value that is incorrect for Gy's theory and unrepresentative for regulatory use.

For an example of the grinding and splitting or aliquoting needed to acquire a correct and representative analytical sample, the critical path (A→B→C→D→E) should be traced through Figure 5, a nomogram adapted from references 2 (Vol. 1) and 7. (Grinding cannot be done, however, for volatile and semivolatile pollutants or to the media for a leach test.) First the nomogram must be made for the specific site (e.g., particle sizes and particle characteristics). The horizontal or *x*-axis is the sample weight in grams, and the vertical or *y*-axis is the RSD of the fundamental error; both axes are in log scale. In the center of the nomogram is a family of linear graphs that introduces the third variable, maximum particle size. Each particle size has its own line, and each line represents one and only one particle size.

The two ways to reduce the *y*-axis intercept, the RSD of the fundamental error, are: (1) to take a line with smaller particle size from the family of graphs, or (2) to take a larger weight of sample on the *x*-axis. First, in the family of linear graphs, the top line of the family represents the largest particle size, namely 75 mm, and intercepts the largest RSD on the *y*-axis. The next lower line is 25.4 mm, and so on down to the line with the lowest RSD, which is for a particle size of 0.2 mm. The 0.2-mm line is probably representative of QA internal standards, in contrast with Superfund's definition of soil as <2 mm and the definition from the Resource Conservation and Recovery Act (RCRA) of soil as <9 mm. These disparities in sizes might explain some of the bench chemist's problems with increasing variance or RSD (e.g.,

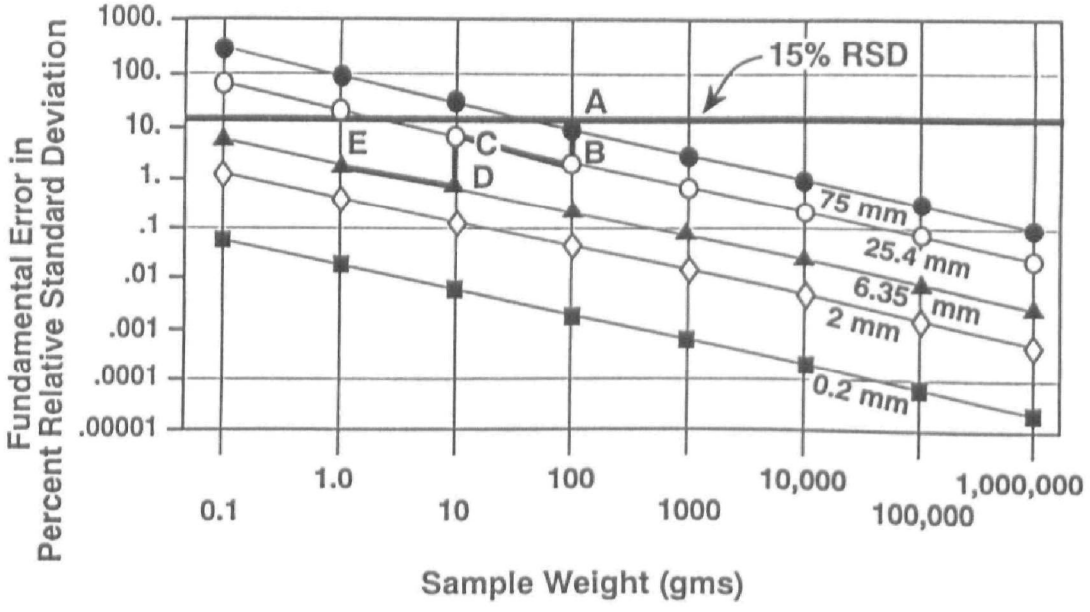


Figure 5. Maximum particle size: preparation error. (Adapted from reference 2, Vol. 1, and reference 7.)

square root of relative variance) as samples come from QA internal standards, Superfund samples, and RCRA samples. Second, each graph has a negative slope, which shows that as the mass of sample on the horizontal axis for a given particle size increases, the relative variance intercepted on the *y*-axis decreases. The horizontal line labeled 15% RSD represents the target accuracy or maximum acceptable RSD. If the maximum particle size of the material of interest is measured empirically to be 75 mm, and the pollutant of interest is one that can be pulverized or grown without loss, such as a heavy metal (Pb), then from the intersection of the horizontal line of maximum RSD = 15% and the downward sloping line of particle size 75 mm, the necessary minimum sample weight can be read on the horizontal or *x*-axis as 100 g. Thus the field technician or scientist must take a sample or composite subsamples so that a field sample of 100 g or more is obtained.

If the chemist is going to take an aliquot of 1 g for the analysis (analytical sample), then the preparation procedure must follow a path such as A→B→C→D→E in Figure 5. To maintain the accuracy of the 100 g of field sample whose maximum particle size is 75 mm, the digestion process must first grind and then split. Grinding reduces particle size and splitting reduces the mass of the sample. Grinding is going down on the nomogram from A to B representing pulverizing from a maximum particle size of 75 mm to a particle size of 25.4 mm, and aliquoting or splitting is moving to the left along the 25.4-mm line on the nomogram from B to C, representing aliquoting or splitting the sample of 100 g to a sample of 10 g. The new critical mass due to the particle reduction or the location of C on the new smaller particle line is the last integer weight tick line that intersects the new particle line just below the 15% relative error line. The amount of information in the 100 g of material of maximum particle size 25.4 mm at B appears to have an order of magnitude (axes in log scale) decrease in RSD. This apparent decrease is not true, because variance of an extracted sample is not reduced by grinding, or information is not created by digestion, but it does mean that now we can split the sample mass down to the new critical mass (10 g) on the current line (25.4-mm line) and still have the original RSD of the 100-g sample, namely 15%. A nomogram path has no lower RSD than its highest point (in this example, point A). Again, more digestion moves the sample from C on the 25.4-mm line to D on the 6.35-mm line. No information is created by grinding, but now the information in 100 g of 75-mm particles, namely 15% RSD, can be carried by a new critical mass of only 1 g as splitting or aliquoting moves us along the 6.35-mm line to E.

The process makes sense if the would-be user remembers that grinding reduces the critical mass needed to carry the same RSD and that the aliquoting or splitting removes only the unneeded mass. One might well ask, "Why the broken path? Wouldn't it be easier to grind all the way in one step and then split?" Yes, it would be simpler, but it would require the grinding of a larger mass of sample; the stepwise path minimizes the mass of material digested. In the interest of minimizing grinding or preparation, Pitard suggests sieving the material so the part less than the new maximum particle size falls through, and then grinding only the part that did not fall through, remembering to recombine the two.

This process sounds a little complicated because it is complicated, but with particle size analyses of the media of interest and with statistically guided preparation (pulverizing and splitting or aliquoting), a correct and representative analytical sample can be prepared for the chemical analysis.

Spatial Variable Statistics

The old adage that a chain is as strong as its weakest link implies that the prudent blacksmith will strengthen the weakest link and try to make all links equally strong. The application to environmental sampling is that error variances are a chain: the analytical variance, the sampling and handling variance, and the field variance are links. The goal of quality improvement is to make the sum of the variances as small as possible, and the cost-effective way to minimize this sum is to spend more resources on the variance link that is improved most cheaply. Because of diminishing returns in variance reduction, the optimal variance to reduce is often the biggest one. The field sampling variance is often the appropriate link or variance to reduce. Variance reduction is most obviously accomplished by taking more samples, but if sampling or analytic costs are high, increasing samples may be too expensive. In many cases, the field sampling variance is economically reduced by going from a random to a spatial variable sampling design.

The term *geostatistics* was coined by Matheron (*9*) to describe the study of regionalized or spatially correlated variables. In the past 20 years, the geostatistical literature has grown enormously, and many significant developments in theory and methodology have been presented. The practice of geostatistics has also spread from its original applications in the mining industry to such fields as soil science, forestry, meteorology, and environmental science.

The geostatistical methods described in this chapter, namely semivariograms and ordinary kriging, represent two of the approaches available to us, and we selected them primarily to illustrate geostatistical concepts and their implications for sampling programs. A discussion of the pros and cons of alternate approaches, such as generalized covariance and universal kriging, is beyond the scope of this chapter. More extensive treatments of the subject can be found in references 3 and 10.

Random or Spatial Variables

Most field sampling plans are based on random variable statistics and assume that the sample observations are independent and identically distributed (IID). However, field samples are usually spatially correlated. *Correlation* is a statistical measurement of the intuitive physical fact that samples taken close together are more similar in value than samples taken farther apart. Neglecting this correlation can make the statistics, tests, and sampling procedures that assume independence (IID) inappropriate (*11, 12*); using this correlation makes the statistics, tests, and sampling procedures of spatial statistics more appropriate and powerful. A truly random variable is

completely described by its probability distribution. Samples are used to estimate this distribution and to estimate statistical descriptors such as mean, median, and standard deviation. In addition, spatial variables must be described by a measure of the correlation between each value and the values at nearby locations. Samples can be used to estimate the spatial correlation function and are frequently used to estimate localized mean values for remediation units or exposure units.

Localized mean estimates are often displayed in the form of isopleths or contour maps. A practical rule for the investigator is that if a contour map is a desired or even a plausible end product of a proposed study, geostatistical methods should be considered.

The implications for the design of a sampling program can be significant. Although random sampling is appropriate for random variables, Olea (*13*) demonstrated that the most effective sampling pattern for local estimation of spatial variables is the regular grid. Yfantis (*14*) evaluated triangular, square, and hexagonal grids. Also, geostatistical studies commonly use a multiphase approach, and the first sampling phase is oriented primarily toward estimating the spatial correlation (*15*).

Semivariograms for Quantifying Spatial Correlation

One way in which spatial correlation can be measured and displayed is by a *semivariogram*, or graph of the type shown in Figure 6. The dots are the empirical semivariogram representing experimental values computed from sample data; the fitted curve is a theoretical semivariogram or an estimation of a spatial correlation function

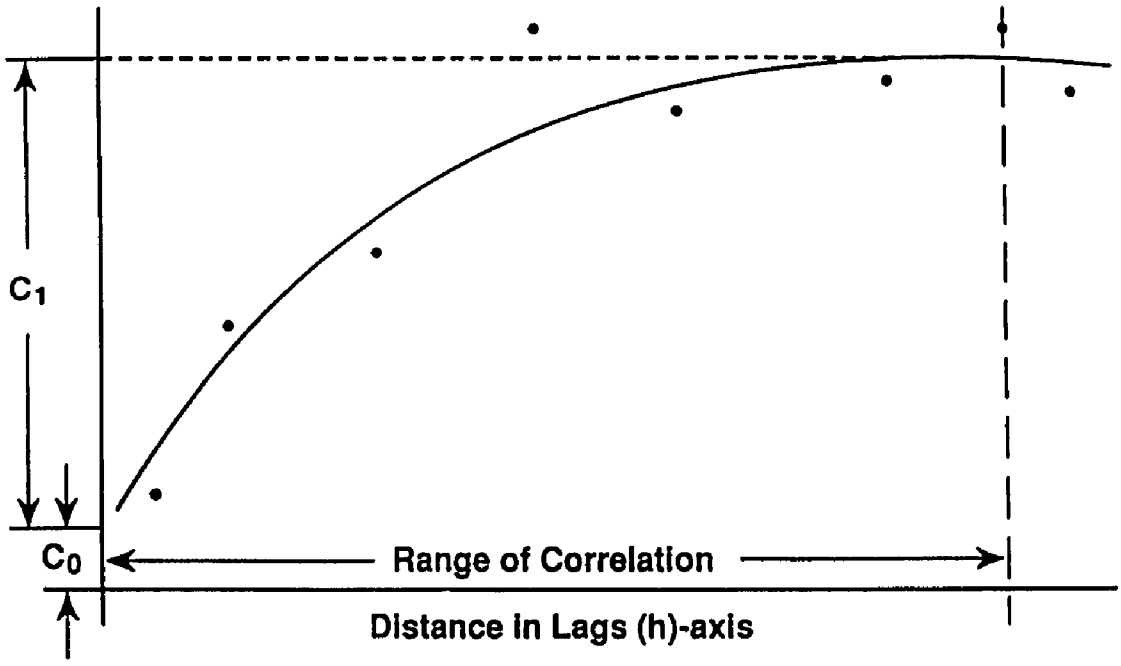


Figure 6. A typical semivariogram. (Reproduced from reference 16. Copyright 1988 American Chemical Society.)

assumed to be characteristic of the sampled area. The horizontal axis, called the *lag axis*, is the distance between points in linear units such as meters or kilometers; the vertical axis, called the *gamma axis*, is the variance of differences in pollution units squared, such as parts per million squared. The experimental points are computed by averaging data grouped into distance class intervals. Variance is a function of lag. The rising nature of the points and curve follows the principle of sampling that states the variance or difference between observations increases as the distance between their locations increases.

Sill and Range of Correlation. Figure 6 is typical of many semivariograms of chemical concentrations in the environment; the rise in variance has an upper bound known as the *sill*. When the variance reaches the sill, sample locations are far enough apart to make the samples independent. The distance on the lag axis at which the semivariogram's curve reaches the sill is the *range of correlation*. This distance is important to the sampling plan, the estimation of pollution over the area under investigation, and the interpolation error. The range of correlation explains a practical relationship between spatial variables and random variables; *random variables* are field samples that are farther apart than the range of correlation, and *spatial variables* are field samples closer together than the range of correlation. This range of correlation is important for choosing the correct analysis; if a classical random variable statistic is wanted, such as the mean or variance, then one type of sampling design that would ensure spatial independence of the samples would be any systematic random design requiring that all samples are at least the range of correlation apart (*17*). If a contour map of pollution isopleths or interpolation variance is wanted, then as the sampling locations get closer together, the local interpolation error decreases. Depending on the information wanted and the spacing of the sample locations, either random or spatial variance statistical analysis can be used on field samples.

Variance Model. In Figure 6, on the vertical axis of the fitted model the variance has two components, C_0 and C_1. The C_1 component of the variance is the measure of structural variation and has the characteristic of increasing variance between sample observations as the distance between sample locations increases. The C_0 component of the variance combines random variance factors, such as sampling and analytical error, along with any unmeasured spatial variance that may exist at distances smaller than the sampling interval; C_0 is constant for all lags. The relationship of C_0 to the need for compositing samples and the relationship of C_1 to the distance between sample locations will be discussed in a later section.

Anisotropy and Directional Semivariograms. The variance structure, as measured by the semivariogram, is often different in the range of correlation in different directions. This condition is called *anisotropy* and must be measured by directional semivariograms. Directional semivariograms are computed experimentally by grouping sample pairs into directional classes, or windows, as well as into distance

classes. The directional ranges of correlation can change the geometry of the sampling grid and the orientation of the grid. Often, not enough preliminary data are available to compute directional semivariograms, and thus the sampling design must work with only an omnidirectional range of correlation. However, an omnidirectional range of correlation and a sampling design from it honor the variance–covariance structure more than conventional random variable methods that consider only a scalar variance.

Kriging for Surface Estimation

Kriging is a linear-weighted average interpolation technique used in geostatistics to estimate unknown points or blocks from surrounding sample data. By assuming that the spatial correlation function inferred from the experimental semivariogram is representative of the points to be estimated as well as those sampled, the interpolation error (kriging error or kriging standard deviation) associated with any estimate that is a linear-weighted average of sample values can be computed. The kriging algorithm computes the set of sample weights that minimize the interpolation error.

Kriging software usually offers both punctual and block output options. *Punctual kriging* treats the input values as located at points and output estimates as values at points. *Block kriging* estimates the output for an area or volume (called block) by averaging multiple points estimated over that area or volume. This difference is determined in the sampling and becomes important in the data analysis (*see* the subsection ***Sample Support and Estimation Blocks***).

Kriging has a number of characteristics of a desirable estimation method: sample weights can be adjusted for anisotropy; samples in correlated clusters can be down-weighted; the degree of smoothing increases as the random component (C_0) of the semivariogram model increases; and, when the semivariogram model is completely random ($C_1 = 0$), the kriging estimator becomes the sample mean, as in independent random sample statistics.

Spatial Outliers

Spatial outliers can be found by examining a geographical plot of the data; they may fit into a random variable histogram of all the data very well. In other words, a *spatial outlier* is a sample value that does not agree in magnitude with the values of its neighboring samples, especially the samples within a range of correlation. For example, a high (polluted) value in a low (background) neighborhood might be a spatial outlier but not a random variable outlier because the high value agrees with other polluted values. Once these outliers are identified, their location descriptions should be looked up in the sampling diary. If they are obviously from different sources that do not have the same correlation structure, they should be excluded from the semivariogram evaluation. The question of whether to include a spatial outlier in the final local estimate of concentration must be answered on a case-by-

case basis. This matter involves the investigator's judgment, just as in the case of random variables.

The following discussion exemplifies an analysis of spatial outliers. Investigating the data from a city-wide sampling campaign for Pb, exploratory data analysts showed an empirical semivariogram with a range of correlation of at least 6 miles and two hot spots that were one order of magnitude higher in concentration than the rest of the data. The data set was printed out on a geographical plot that showed the two hot spots to be in sharp contradiction to their individual local neighborhoods, that is, every neighboring point and every point one neighbor out was at least one order of magnitude lower in concentration. The geographical map that identified the freeway system and the data showed that both points seemed very close to the freeways. In checking the sample log book this conclusion was confirmed; one of the aberrant samples was taken under a freeway overpass and the other at a freeway on-ramp. Freeway Pb is said to have a range of about 500 feet. Thus, because the two points represented a different source of Pb and had a much shorter range, they were excluded from the semivariogram computations. However, what was to be done with them in the kriging and mapping? If they were included in the kriging, they would spread their high values over circular areas of 6 miles in radius. This representation would be grossly untrue because the outliers are known to have a different source and a shorter range of correlation. The mapping would show a large area needing remediation that, in fact, did not need remediation. Nevertheless, the values had been found, and users of the map (risk assessors) needed to know of the hot spots. The compromise was to krige and contour the Pb concentrations of the other samples onto a kriging map and then just print the magnitude of such outliers at their respective locations on the map.

Spatial Soil Sampling

The growing number and complexity of toxic chemicals and hazardous waste sites call for a new statistical technique for monitoring with more efficient sampling designs and more precise data analysis. Geostatistics is a promising tool for these needs. This section traces the logic sequence of geostatistical analysis and then draws together the implications of geostatistical sampling design for soil pollution monitoring. Geostatistical sampling design has at least two phases: (1) the survey or the preliminary sampling to find the extent of the plume and to estimate a semivariogram, and (2) the census to take as many samples as needed to estimate the surface within the desired accuracy as calculated from the semivariogram model.

Sample Support and Estimation Blocks

The basic assumption of geostatistical sampling is to define and assign area or volume to all inputs and outputs. In monitoring for environmental protection, the spatial quantities to be defined and assigned are the *sampling unit* (area or volume), the

remediation unit, and the *exposure unit*. Geostatisticians call the sampling unit the sample support. The sampling unit or support is ambiguous: it is used to refer to both the amount of medium extracted for the sample and the in situ area or volume represented by the sample. The context usually identifies whether the extracted support or the in situ support is meant. The remediation unit is determined by the method of remediation, and the exposure unit is determined by the risk assessor. For example, an appropriate remediation block might be a volume 250 ft long, 16 ft wide, and 0.5 ft thick, because this amount was the minimum volume to move economically. The shape is dictated by the up and back pass of a bulldozer with an 8-ft blade that scrapes up one truckload of contaminated soil. Sample unit, remediation unit, and exposure unit need to be defined (*18*) and then incorporated by a geostatistician into the sampling plan.

The critical mass of a correct sample should be calculated as previously explained (*see* the subsection ***Analytical Error***). The spatial variance of the sampling unit should be measured by taking "too many" equally spaced samples in several units in an exploratory sampling trip to the site. If the sampling unit is larger in spatial variance (a large spatial variance can be encountered in a small area), then the field samples design will have to use composite samples. In spatial compositing the geometry as well as the mass of the subsamples (samples to be composited) is important. The general rule is that subsamples should be equally spaced on the sampling unit. For example, if four subsamples can be afforded, then one should be taken from each quarter of the in situ support. Each subsample for the compositing should be a correct sample (*see* the subsection ***Analytical Error***). All samples for all analyses, even eye-balling, should have the same representativeness, which for composite samples means the same number of subsamples. The composite field sample, just like any other average, has a variance divided by the number of subsamples. *Homoscedasticity* (equality of variances) is a requirement for every data analysis even when eye-balling the data. If the quantity to be estimated (e.g., remediation or exposure unit) equals the sampling unit, punctual kriging analysis may be used because there is no change of scale or support. If the desired area or volume of estimation is larger than the sampling unit, block kriging will have to be used.

Survey or Semivariogram Sampling

In a multiphase sampling program using spatial statistics, the primary goal in the initial exploratory sampling is the collection of enough data to compute an empirical semivariogram and to determine the extent of the plume. These goals may conflict if limited resources are available. Widely spaced samples are needed to define extent, and closely spaced samples are often needed for semivariogram analysis. Approaches to this problem include regular grids (i.e., radial, square, or rectangular), transects, and combinations.

Burgess et al. (*19*) suggested transect sampling for variogram input, and this idea led to very good variograms in agricultural applications. However, in pollution monitoring, transects alone have given very noisy variograms. This result is probably

due to intrinsic noise in pollution data, which is often highly skewed and contains high coefficients of variation. A combination exploratory grid, consisting of a grid of square sampling units having extended transects in the directions of the major axis and minor axis of the estimated plume (*20*), is illustrated in Figure 7. Prior information may be used to select the best grid orientation. For example, if the plume to be investigated was made by aerial deposition from an identifiable source, then wind roses can be examined for wind direction and magnitude, and topographic maps can be examined for natural barriers. Only the relatively regular grid concept is important in Figure 7; the orientation is site-specific.

If the extent of the plume must be found, and funds are limited, then the transect samples should be variably spaced closer together at the grid center and farther apart at the grid extremes. The purpose of this sampling is to capture the correlation structure of the plume. Inhabited areas have a high occurrence of disturbed sampling sites and local pollution from secondary sources, which are only stochastic noise to the semivariogram's calculation. Therefore, this noise should be avoided by this sampling. For example, aerially deposited smelter Pb should not be mixed with auto Pb by taking samples along the freeways. The samples from the semivariogram sampling can be pooled with the secondary mapping samples if they have the same support.

However, the semivariogram sampling often is the sampling that tests for the need for more compositing. If the support is changed between the samplings and we wish to pool the samples for analysis, then the change in support must be corrected before pooling. The sampling team must be aware of the need to keep all samples on the same support. When compositing, the same number and mass of

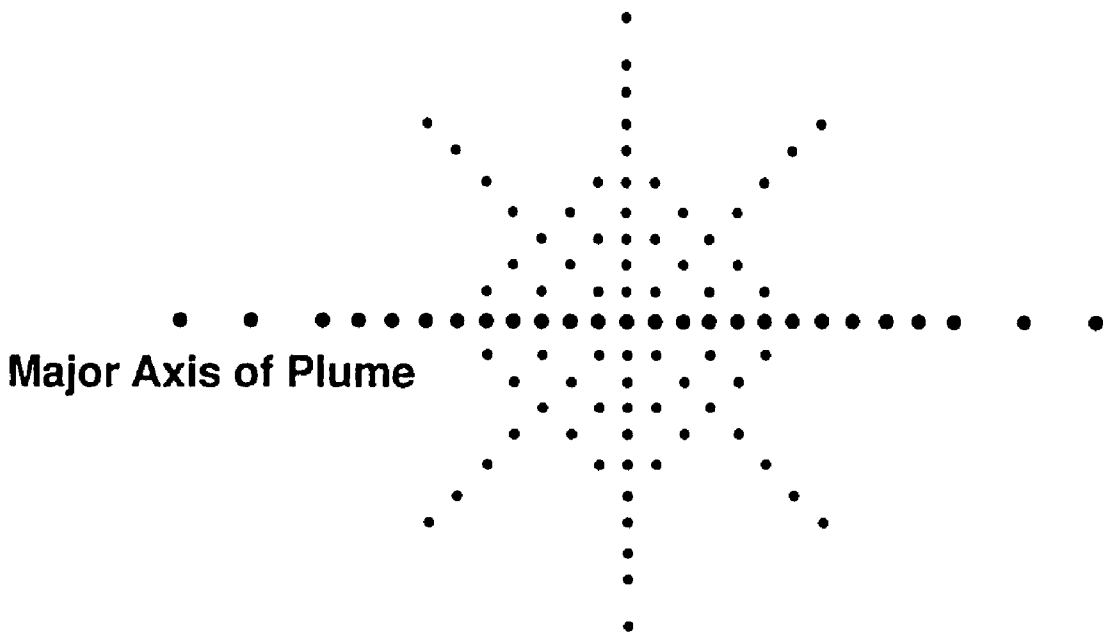


Figure 7. An exploratory grid design. (Reproduced from reference 16. Copyright 1988 American Chemical Society.)

subsamples and the same spacing or geometry must be maintained. When the sampled locations must move from the regular grid to avoid cultural improvements or natural barriers, then the spatial analysis program is corrected for this movement by the true coordinates of the new sample locations; however, no easy method is available for the program to correct for change of support. If the microvariation could be sampled and the support established before the semivariogram sampling, then a complex statistical problem could be avoided in the pooling of samples for the spatial analysis.

Some samples should be taken close together (in the scale of the sampling unit) to determine the need for composite samples. This sampling can be combined with field duplicates for quality analysis and control. Gy's fundamental error and compositing become more important as coring volumes decrease. These microvariation samples should be taken at a distance of a few multiples of the core's diameter apart. The distance between sample locations or grid unit's length needs to be estimated from the sample unit of interest (e.g., residential yard, city block, or square mile section) and the desired output unit (e.g., remediation unit, that is, the minimum volume of surface soil to be removed). The optimum exploratory sampling distance is a proper fraction of these measurements, but it is often determined by money available for sampling.

Census or Sampling for Map Making

In spatial statistics, the goal of secondary sampling is to uniformly cover the area in question with a density of samples sufficient to contour the plume with an acceptable error of interpolation. This sample coverage is accomplished by using the directional semivariograms to determine the orientation, shape, and size of the grid cell. Independent random variable statistics, in which the number of samples is computed, differs from spatial statistics, in which orientation, shape, and size of the grid are calculated and the number of samples is determined from the number of grid cells needed to cover the area.

If the directional semivariograms have a marked difference in their respective ranges of correlation, then the optimum cell geometry is not a square but a rectangle with the longer side in the direction of the longer range of correlation, and the ratio of the sides should be the ratio of the ranges of correlation. Thus the grid cell sides are of equal correlation or kriging (interpolation) variances rather than equal distance. This characteristic will save a lot of samples while retaining the same accuracy in both directions.

Boundary. For secondary sampling, the extent of the sampling grid must first be chosen. The sampling grid must extend beyond the suspected plume or area in question. The area in question must be bounded by sampling locations to avoid extrapolation in the kriging estimation algorithm for contouring. Extrapolation, which is estimating a value from data on only one side of the location of the point to be estimated, is likely to lead to unrealistically high or low values. If an action level

has been set and a part of the plume has been adequately proven to be above or below the acting level, then that part of the plume need not be resampled. The sampling may be guided more by population areas or critical receptors than by the actual plume. The goals of the sampling must be written, and the areas of interest, action levels, and action areas (sampling unit, remediation unit, and exposure unit) must be defined before the optimum grid design can be made.

Compositing Samples Reduces Nugget. The next step in secondary sampling is choosing the sample support (*21*). If a residential yard is the sampling unit, then the ideal sampling process would be to take the entire yard, blend it to homogeneity, and remove the appropriate number of aliquots or splits to meet the volume needed by the laboratory for analysis. However, because few residents would donate their whole yard to science, and laboratory mixing equipment such as V-blenders or ball mills cannot homogenize so large a volume, this sampling unit must be represented by a few symmetrically laid out subsamples composited together. The number of subsamples is a compromise between the size of the microvariance and the amount of time and money allowed for the digestion of the subsamples. The subsamples are laid out symmetrically because a structural or spatial correlation may exist.

The mixing of the subsamples to achieve homogeneity is essential for compositing. If the medium is water, then the task is relatively easy; for soils or sediments, the task is difficult. Aliquots or splits should be taken after the mixing to make the final sample more representative. If a large nugget (e.g., $C_0 > 0.3$ relative variance) persists after Gy's critical mass calculations and compositing within the support, then the relative sizes of the field sampling and the laboratory analysis errors must be identified. The analysis of some pollutants has an analytical error that overwhelms the field sampling error and accounts for approximately all the semivariogram nugget.

The minimum volume at each step and especially the aliquot used by the chemical analyst in the lab must exceed the critical mass referred to in Gy's theory (*see* the section **Pierre Gy's Sampling Theory**).

Grid Unit Length or Distance Between Sample Locations. The range of correlations, the nugget (C_0), and the sampling budget determine the *grid unit length*, or the distance between sample locations. This length determination was discussed in mathematical detail by Yfantis et al. (*14*). Figure 8 shows the graphs of interpolation variance as a function of the ratio of grid spacing to range of correlation for a family of semivariograms. The model variograms each have relative C_1 and C_0 so that their sum equals 100%. The variograms differ only in the fraction of the sill ($C_0 + C_1$) represented by the nugget component (C_0) and the structure (C_1). If the semivariogram has a big nugget like the top graph of $C_1 = 10\%$ and $C_0 = 90\%$, then diminishing returns (the curve has less rapid vertical drop and becomes more horizontal) start and increase if the sample distance is less than two-thirds of the range of correlation. For a very low nugget, such as the lowest graph ($C_1 = 100\%$ and C_0

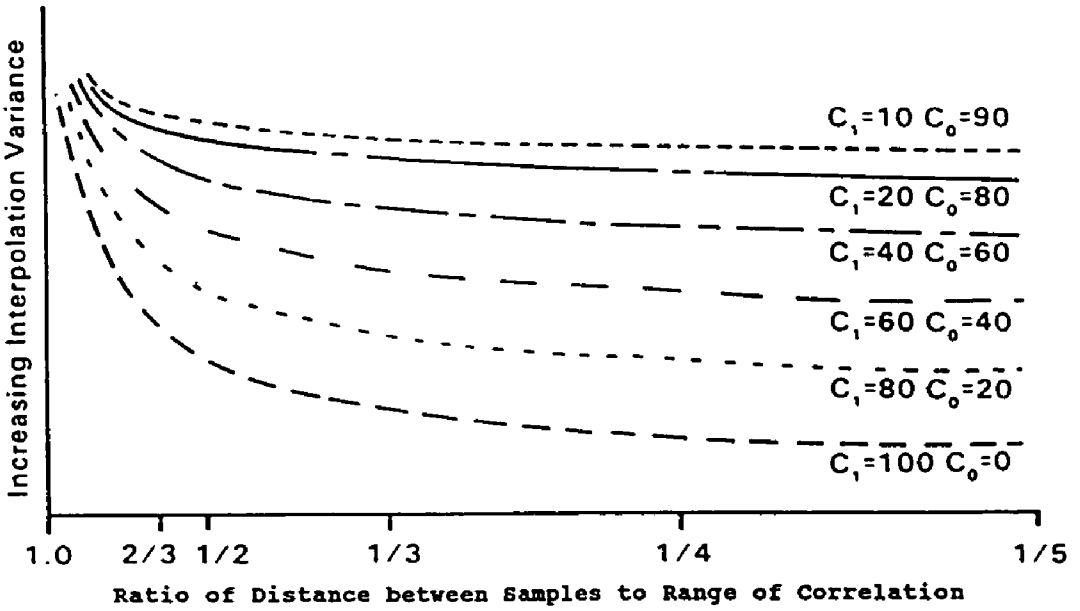


Figure 8. Diminishing information for additional samples. (Reproduced from reference 16. Copyright 1988 American Chemical Society.)

= 0%), diminishing returns do not start and increase until the sampling distance is less than one-half of the range of correlation. The general rule is that for smaller nuggets (C_0), the distance between sampling points on the sampling grid gets smaller. The grid should be laid out with no vertices unsampled. If this design exceeds budget, then the whole grid size should be adjusted, not just certain vertices left unsampled as in systematic random sampling.

Some real-world examples can clarify how the magnitude of the nugget (C_0) and the range of correlation determine the optimum cell size or distance between samples. One Pb smelter had a nugget of about 40% and a range of correlation of 3200 ft. In Figure 8, the family of diminishing return curves and the graph (for C_0 = 40%) indicates by observation and judgment that the point of diminishing returns is between one-third (33.3%) and one-fourth (25%) of the range of correlation, or 29% for the sake of argument. The sampling distance should not be less than 29% × 3200 ft, or 928 ft. Expressed as a function of money, the sampling distance should be the shortest affordable distance in keeping with the toxicity of the pollutant, but not less than 928 ft between samples. In contrast, a second Pb smelter had a semivariogram with a nugget of zero (0) and a range of correlation of 2400 ft. In Figure 8, the curve of diminishing returns for $C_0 = 0$ indicates by observation and judgment that the point of diminishing returns is between one-fourth (25%) and one-fifth (20%) of the range of correlation, or 22.5% for the sake of argument. In this case, the sampling distance should not be less than 22.5% × 2400 ft, or 540 ft. Expressed as a function of money, the sampling distance should be the shortest affordable distance in keeping with the toxicity of the pollutant, but not less than 540 ft. If the funding is adequate and the pollutants are of extreme toxicity,

then the distance indicated by the point of diminishing returns should be used in minimum interpolation variance. If there is less money and the pollutant is less toxic, then a longer distance should be used for the grid cell's side. The directional semivariograms should orient the sampling grid.

The Pb smelters mentioned previously worked well with an east–west and north–south grid because the plume was formed by 80 years of aerial deposition. A third set of data, dioxin along a highway, gave a readable semivariogram in a direction of 13 degrees from east to west. This discovery took much searching because we started with the default directions (0, 45, 90, 135 degrees) of the semivariogram software; these default semivariograms showed no structure [pure nugget semivariograms (C_0 = 100%)]. After we discovered the semivariogram at 13 degrees the reason became obvious, because the road that was the transport of the pollutant ran at that angle and so should any sampling grid.

In the field, some vertices cannot be sampled because of man-made improvements or natural barriers, but these vertices must be sampled as closely as possible, and the actual coordinates should be used in the spatial analysis program.

Grid Orientation and Shape Versus Anisotropy. If the ranges of correlation are extremely different on the directional semivariograms, then the correlation structure is anisotropic. Optimum sampling patterns reflect this anisotropy. For example, the sides of a rectangular grid would be in the same ratio as the ranges of correlation for the corresponding directional semivariograms. This ratio was explained in detail by David (22), and a sampling design for logarithmic anisotropy was derived by Barnes (23). Anisotropy is a frequent occurrence, but often the semivariogram sampling gathers too few samples to measure it. Thus, more samples may be used cost-effectively in the semivariogram sampling in order to save samples in the larger census (or mapping) sampling by identifying and taking advantage of anisotropy.

Use of the triangular grid as opposed to the rectangular grid has been discussed (*13*, *14*). If the nugget is large ($C_0 >> C_1$), little is gained by the triangular grid. Also, the triangular grid makes taking advantage of anisotropy more difficult. If a triangular grid is chosen, a theodolite, which is a surveying instrument, is not needed in the field; instead every other row of samples must be offset by one-half of a grid length. In practice, this action is easier than it sounds and almost as easy as the traditional square grid.

Beyond Anisotropy

Numerous additional geostatistical considerations affect environmental sampling. These considerations include spatial drift or trend, multivariate analysis, mixed or overlapping populations, concentration-dependent variances, and specification of confidence limits. Geostatistical techniques have been developed over the years to deal with these various problems, but an adequate discussion is beyond the scope of this chapter.

Acknowledgments

The EPA, through its Office of Research and Development, funded and performed the research described here.

References

1. Gilbert, R. O. *Statistical Methods for Environmental Pollution Monitoring;* Van Nostrand Reinhold: New York, 1987.
2. Pitard, F. F. *Pierre Gy's Sampling Theory and Sampling Practice;* CRC Press: Boca Raton, FL, 1989; Vol. 1 & 2.
3. Isaaks, E. H.; Srivastava, R. M. *An Introduction to Applied Geostatistics;* Oxford University: New York, 1990; pp 1–592.
4. Pitard, F. F. In *Pierre Gy's Sampling Theory and Sampling Practice;* CRC Press: Boca Raton, FL, 1989; Vol. 2, p 36.
5. Gilbert, R. O. In *Statistical Methods for Environmental Pollution Monitoring;* Van Nostrand Reinhold: New York, 1987; pp 35–42.
6. Pitard, F. F. In *Pierre Gy's Sampling Theory and Sampling Practice;* CRC Press: Boca Raton, FL, 1989; Vol. 1, pp 169–183.
7. *Preparation of Soil Sampling Protocols: Sampling Techniques and Strategies;* Center for Environmental Research Information: Cincinnati, OH, 1992; pp A1–A16; EPA/600/R–92/128.
8. Pitard, F. F. In *Pierre Gy's Sampling Theory and Sampling Practice;* CRC Press; Boca Raton, FL, 1989; Vol. 1, p 190.
9. Matheron, G. *Econ. Geol.* **1963,** *58,* 1246–1266.
10. Journel, A. G. *Geostatistics for the Environmental Sciences;* Stanford University: Stanford, CA, 1986.
11. Palmer, M. W. *Vegetation (Dordrecht, Netherlands)* **1988,** *75,* 91–102.
12. Cliff, A. D.; Ord, J. K. *Spatial Processes: Models and Applications;* Pion: London, 1981; pp 1–266.
13. Olea, R. A. *Math. Geol.* **1984,** *16*(4), 369–392.
14. Yfantis, E. A.; Flatman, G. T.; Behar, J. V. *Math. Geol.* **1987,** *19*(3), 183–205.
15. Flatman, G. T.; Yfantis, E. A. *Environ. Monit. Assess.* **1984,** *4,* 335–349.
16. Flatman, G. T.; Englund, E. J.; Yfantis, A. A. In *Principles of Environmental Sampling;* Keith, L. H., Ed.; ACS Professional Reference Book; American Chemical Society: Washington, DC, 1988; pp 73–84.
17. Borgman, L. E.; Quimby, W. F. In *Principles of Environmental Sampling;* Keith, L. H., Ed.; ACS Professional Reference Book; American Chemical Society: Washington, DC, 1988; pp 25–43.
18. Neptune, D.; Brantly, E. P.; Messner, M. J.; Michael, D. I. *Hazard. Mater. Control* **1990,** *May/June,* 19–25.
19. Burgess, T. M.; Webster, R.; McBratney, A. B. *J. Soil Sci.* **1981,** *32,* 643–659.
20. Starks, T. H.; Brown, K. W.; Fisher, N. J. In *Quality Control in Remedial Site Investigation;* American Society for Testing and Materials: Philadelphia, PA, 1986; Vol. 5, pp 57–66; ASTM STP925.
21. Starks, T. H. *Math. Geol.* **1986,** *18*(6), 529–537.
22. David, M. *Geostatistical Ore Reserve Estimation;* Elsevier Scientific: Amsterdam, 1977.
23. Barnes, M. G. *Statistical Design and Analysis in the Cleanup of Environmental Radionuclide and Other Spatial Phenomena;* TRAN STAT (Statistics for Environmental Studies) No. 13, Battelle Memorial Institute: Richland, WA, 1980; pp 1–21.

Reprinted from ACS Professional Book
Principles of Environmental Sampling
Lawrence H. Keith, Editor
Published 1996 by the American Chemical Society

Welcome to the 12th Annual EPA Conference on Statistics

After a year's hiatus, it is a pleasure to welcome you to the 1997 EPA Conference on Statistics. What a difference a year can make! Last year, we postponed the conference and held a series of "local" training sessions. While these were a great success, they were local and accessible only to those folks in the Washington or Raleigh/Durham areas. I heard from many people that while we may have successfully circumvented the travel issue, the opportunity was not available to everyone.

So here we are all together again in Richmond for a conference that will include all the elements that we have heard you want. There will be formal sessions, plenary sessions, workshops, training, round tables, rectangular tables, panel discussions, poster sessions, and outstanding speakers. The theme this year is "Future Directions for Statistics at EPA". And what could be a better time to discuss future directions than right after the Administrator's announcement of the creation of the new EPA Center for Environmental Information and Statistics. In fact, we have changed the schedule around in a last-minute upheaval to arrange for Agency officials responsible for the CEIS to be here to present their vision and respond to your comments and suggestions.

I also want to encourage you to avail yourself of the informal opportunities here to discuss common questions and concerns with fellow statisticians. It's no secret that some of the best information is garnered in the hallways, over dinner, or while waiting for the elevator. We are always anxious to make it even better. I want to thank the planning and arrangements committees for their efforts to organize this conference and Margaret Conomos for her assistance in coordinating transportation. It always looks deceptively easy, and we owe it to the hard work of these people that it is easy for the rest of us. Special thanks are in order for Marcia Gardner of SRA Technologies, Inc., who handled all the details so well.

Barry D. Nussbaum
Conference Chairman

Conference Planning Committee

John Fox
Henry Kahn
Elizabeth Margosches
John Warren

EJBD
ARCHIVE
EPA
230-
R-
97-
003

Arrangements Committee

Joan Bundy
Pat Wilkinson

AGENDA

TWELFTH ANNUAL EPA CONFERENCE ON ENVIRONMENTAL STATISTICS

Richmond, VA - April 1-3, 1997

AGENDA

Tuesday, April 1, 1997

REGISTRATION - Conference Area Foyer — *9 30 - 10 30 am*

OPENING SESSION - Grand Ballroom Section B — *10 30 am - 12 00 pm*

Welcome & Introductions - Barry Nussbaum, Chair, Conference Planning Committee

Keynote Speaker: N. Phillip Ross, Director, Center for Environmental Statistics

"Center for Environmental Information and Statistics (CEIS)"

Followed by an Interactive Discussion with the CEIS Development Staff

Tom Curran / Bill Hunt

Lunch Break — *12 00 - 1 15 pm*

TRAINING SESSION 1-A - Georgian/Elizabethan Rooms — *1 15 - 4 45 pm*

EnvironmentalStats for S-PLUS: Software for Environmental Statistics

Presenters: Steven Millard, PSI, and Nagaraj Neerchal, University of Maryland Baltimore Campus

SESSION I - Raleigh/Drake Rooms — *1 15 - 2 45 pm*

Cancer Statistics, Epidemiology and Genetics — *Chair Ruth Allen, NCI*

"Atlas of Cancer Mortality in the United States, 1970-92" — *Presenter Susan Devesa, NCI*

"Evaluating Disease Cluster Alarms" — *Presenter Martin Kulldorff, NCI*

Break *- Conference Area Foyer* — *2 45 - 3 00 pm*

TRAINING SESSION 1-A - (continued) — *3 00 - 4 45 pm*

SESSION II - Raleigh/Drake Rooms — *3 00 - 4 45 pm*

Representativeness in Statistics and Quality Assurance — *Chair John Warren, ORD*

Presenters John Warren, ORD, and Malcolm J Bertoni, RTI

ROUNDTABLE DISCUSSIONS — *4 45 - 6 00 pm*

GROUP A - Georgian/Elizabethan Rooms

Statistics & Health — *Facilitators Ruth Allen, NCI, and Elizabeth Margosches, OPPTS*

GROUP B - Raliegh/Drake Rooms

Quality Assurance — *Facilitator John Warren, ORD*

GROUP C - Grand Ballroom, Section B

Statistical Research — *Facilitators Barry Nussbaum, OPPE, and Larry Cox, ORD*

GROUP D - Hilliard Room

Risk & Uncertainty — *Facilitator: Barnes Johnson, OSWER*

Wednesday, April 2, 1997

SESSION III - Georgian/Elizabethan Rooms *8 45 - 10·15 am*

How Severe Is It? *Chair ElizabethMargosches, OPPTS*

"Toxic Severity for a Useful and Understandable Benchmark Dose" Presenter Linda Teuschler, ORD

"Severity Analysis Using Ridits" Presenter Mary Marion, OPPTS

SESSION IV - Raleigh/Drake Rooms *8 45 - 10 15 am*

Exposure Assessment *Chair John Fox, OW*

"Interpreting Data from a National Survey of Protozoan in Drinking-Water Sources"

Presenter. John Fox, OW

"Relationships Between Dioxins in Soil, Air, Ash, and Emissions from a Municipal Solid Waste Incinerator Emitting Large Amounts of Dioxins"

Presenter Matthew Lorber, ORD

"Statistical Modeling of Dioxin Concentration Data from Sediment Cores"

Presenter Paul Pinsky, ORD

Break *- Conference Area Foyer* *10 15 - 10 30 am*

PANEL DISCUSSION - Grand Ballroom, Section B *10 30 am - 12 00 pm*

EPA Cooperative Agreements *Chair Barry Nussbaum, OPPE*

Participants Larry Cox, ORD, Peter Guttorp, University of Washington, G P Patil, Pennsylvania State University

www.stat.washington.edu/~~MRCES~~ MRCSE National Research Center for Statistics & Environment

envstat list

Lunch Break *12 00 - 1 15 pm*

TRAINING SESSION 1-B - Georgian/Elizabethan Rooms *1 15 - 4 45 pm*

EnvrionmentalStats for S-PLUS: Software for Environmental Statistics

Presenters· Steven Millard, PSI, and Nagaraj Neerchal, University of Maryland Baltimore Campus

TRAINING SESSION 2 - Raleigh/Drake Rooms *1 15 - 3 30 pm*

Spatial Statistics Sampling *Chair George Flatman, EPA, Las Vegas*

"Spatial Sample Design" Presenter Evan Englund, EPA, Las Vegas

"Skewed Frequency Distributions" Presenter. George Flatman, EPA, Las Vegas

Break *- Conference Area Foyer* *2.15 - 2 30 pm*

TRAINING SESSION 1-B (continued) *2 30 - 4 45 pm*

TRAINING SESSION 2 - (continued) *2 30 - 3 30 pm*

SESSION V - Raleigh/Drake Rooms *3 30 - 4 45 pm*

Applications of Statistical Calibration Techniques in Analyzing Environmental Data

Chair Bimal Sinha, University of Maryland Baltimore Campus (UMBC)

"Confidence Regions and Tests in a Calibration Problem"

Presenter Thomas Mathew, UMBC

MINI SESSION A - Hilliard Room *4 45 - 5 30 pm*

Water Quality & Fishy Statistics *Chair/Presenter Henry Kahn, OW*

"Recent Developments in the Estimation of U S Fish Consumption"

MINI SESSION B - Georgian/Elizabethan Rooms *4 45 - 5 45 pm*

Statistics and the Internet *Chair/Presenter Chapman Gleason, OPPE*

"Using the Web and other Networking Technologies in Support SAS for the Enterprise"

Wednesday, April 2, 1997 (continued)

RECEPTION & POSTER PRESENTATIONS - Capitol Room . 5 30 - 6 45 pm

Pesticide Residue Monitoring Data
Presenter Edward Brandt, EAB, OPP, OPPTS

A Master Sampling Frame for the Collection of Non-Agricultural Pesticide Usage Data
Presenter· Alan R. Goozner, EAB, OPP, OPPTS

The National Air Quality and Emissions Trends Report, 1995
Presenter· David Mintz, OAR

Thursday, April 3, 1997

SESSION VI - Grand Ballroom, Section B . . . . . 8 45 - 10 15 am

Statistics of Measurement in Analytical Chemistry
Chair Henry Kahn, OW

"A Two Component Model for Error in Analytical Chemistry and Issues of Detection and Quantification"
Presenter. David M Rocke, Director, Center for Statistics in Science and Technology, University of California, Davis

"Estimation of Precision of Low Concentration Chemical Analytical Measurements and Establishment of Detection and Quantification Limits"
Presenters· Henry Kahn, OW, Kathleen Stralka and Raphael Kuznetsovski, SAIC

Break -Conference Area Foyer . . 10 15 - 10 45 am

CLOSING SESSION - Grand Ballroom, Section B . . . . . 10 45 - 12 30 pm

Featured Speaker. Daniel B. Carr, School of Information Technology and Engineering, George Mason University

"Statistical Graphics for Environmental Applications Developments and Challenges"

Bus to EPA Headquarters leaves at 1:30 pm

ATTENDEE LIST

Twelth Annual EPA Conference
on Environmental Statistics
List of Attendees

Ruth Allen
National Cancer Institute
Division of Cancer Epidemiology
and Genetics
6130 Executive Boulevard, MSC 7395
EPN Room 535
Bethesda, MD 20852-7395
(301) 496-1609
Fax: (301) 402-4279
Allenr@epndcc.nci.nih.gov

Robin Anderson
OAR/ORIA
U.S. EPA (6603J)
401 M Street, SW
Washington, DC 20460
(202) 233-9385
Fax: (202) 233-9650
Anderson.Robin@epamail.epa.gov

David Annett
Support Contrator for NCI (SEER Program)
IMS, Inc.
12501 Prosperity Drive, Suite 200
Silver Spring, MD 20904
(301) 680-9770
Fax: (301) 680-8304
David_Annett@nih.gov

Lara Autry
OAR/OAQPS/EMAD
U.S. EPA (MD-19)
Research Triangle Park, NC 27711
(919) 541-5544
Fax: (919) 541-1039
Autry.Lara@epamail.epa.gov

Jeff Beaubur, Ph.D.
OPPTS/HERD
U.S. EPA (7403)
401 M Street, SW
Washington, DC 20460
(202) 260-2263
Fax: (202) 260-1279

Malcolm Bertoni
Research Triangle Institute
401 M Street, NW, Suite 740
Washington, DC 20460
(202) 728-2067
Fax: (202) 728-2095
mjb@rti.org

Ed Brandt
OPPTS/OPP
U.S. EPA (7503W)
Office of Pesticides
401 M Street, SW
Washington, DC 20460
(703) 308-8050
Fax: (703) 308-8151
Brandt.Edward@epamail.epa.gov

Lori Brunsman
OPPTS/OPP/HED
U.S. EPA (7509C)
401 M Street, SW
Washington, DC 20460
(703) 308-2902
Fax: (703) 305-5147
Brunsman.Lori@epamail.epa.gov

Judy Calem
OW/OGWDW
U.S. EPA (4607)
401 M Street, SW
Washington, DC 20460
(202) 260-8638
Fax: (202) 260-3762
Calem.Judy@epamail.epa.gov

Daniel Carr
George Mason University
School of Information Technology
and Engineering
Fairfax, VA 22030-4444
(703) 993-1671
Fax: (703) 993-1521

Steven Chang
OSWER/OERR
U.S. EPA (5204G)
401 M Street, SW
Washington, DC 20460
(703) 603-9017
Fax: (703) 603-9104
Chang.Steven@epamail.epa.gov

Darlene Cockfield
OPPE/OSPED/EID
U.S. EPA (2163)
401 M Street, SW
Washington, DC 20460
Fax: (202) 260-4903

Margaret Conomos
OPPE
U.S. EPA (2164)
401 M Street, SW
Washington, DC 20460
(202) 260-3958
Fax: (202) 260-4968
Conomos.Margaret@epamail.epa.gov

Lawrence Cox
ORD/NERL
U.S. EPA (MD-75)
Research Triangle Park, NC 27711
(919) 541-2648
Fax: (919) 541-7588
Cox.Larry@epamail.epa.gov

John Creason
ORD/NHEERL
U.S. EPA Room 215 ERC (MD-55)
Research Triangle Park, NC 27711
(919) 541-2598
Fax: (919) 541-5394
Creason.John@epamail.epa.gov

David Crosby
American University
Department of Mathematics
and Statistics
4400 Massachusetts Avenue, NW
Washington, DC 20016
(202) 885-3135
Fax: (202) 885-3155
Crosby@nzms.wwb.noaa.gov

Thomas Curran
OAR/OAQPS
U.S. EPA (MD-12)
Research Triangle Park, NC 27711
(919) 541-5694
Fax: (919) 541-4028
Curran.Thomas@epamail.epa.gov

Susan Devesa
National Cancer Institute
EPN, Room 415
Bethesda, MD 20892
(301) 496-8104
Fax: (301) 402-0081
Devesas@epndce.nci.nih.gov

Donald Doerfler
ORD/ERC
U.S. EPA (MD-55)
Research Triangle Park, NC 27711
(919) 541-7741
Doerfler.Donald@epamail.epa.gov

Evan Englund
ORD/NERL-CRD (CAP)
U.S. EPA
P.O. Box 93478
Las Vegas, NV 89193-3478
(702) 798-2248
Fax: (702) 798-2107
Englund.Evan@epamail.epa.gov

George Flatman
ORD/NERL-CRD
U.S. EPA
P.O. Box 93478
Las Vegas, NV 89193-3478
(702) 798-2528
Fax: (702) 798-2208
Flatman.George@epamail.epa.gov

John Fox
OW
U.S. EPA (MC-4303)
401 M Street, SW
Washington, DC 20460
(202) 260-9889
Fax: (202) 260-7185
Fox.John@epamail.epa.gov

Mary Frankenberry
OPPTS/OPP/EFED
U.S. EPA (7507C)
401 M Street, SW
Washington, DC 20460
(703) 305-5694
Fax: (703) 305-6309
Frankenberry.Mary@epamail.epa.gov

Chapman Gleason
OPPE
U.S. EPA (2163)
401 M Street, SW
Washington, DC 20460
Gleason.Chapman@epamail.epa.gov

Alan Goozner
OPPTS/OPP/BEAD
U.S. EPA (7503W)
401 M Street, SW
Washington, DC 20460
(703) 308-8147
Fax: (703) 308-8151
Goozner.Alan@epamail.epa.gov

Peter Guttorp
University of Washington
National Research Center for Statistics
and the Environment
Box 351720
Seattle, WA 98195-1720
(206) 616-9262
Fax: (206) 616-9443
Peter@stat.washington.edu

Karen Hogan
OPPTS/OPPT
U.S. EPA (7403)
401 M Street, SW
Washington, DC 20460
(202) 260-3895
Fax: (202) 260-1279
Hogan.Karen@epamail.epa.gov

David Holland
ORD/NHEERL
U.S. EPA (MD-56)
ERC Annex
Research Triangle Park, NC 27711
(919) 541-3126
Fax: (919) 541-1486
Holland.David@epamail.epa.gov

William F. Hunt, Jr.
OAR/OAQPS/EMAD
U.S. EPA (MD-14)
Research Triangle Park, NC 27709
(919) 541-5536
Fax: (919) 541-2357
Hunt.Bill@epamail.epa.gov

Helen Jacobs
OW
U.S. EPA (4303)
401 M Street, SW
Washington, DC 20460
(202) 260-5412
Fax: (202) 260-7185
Jacobs.Helen@epamail.epa.gov

Barnes Johnson
OSWER/OSW
U.S. EPA (5307W)
401 M Street, SW
Washington, DC 20460
(703) 308-8855
Fax: (703) 308-0511
Johnson.Barnes@epamail.epa.gov

Henry Kahn
OW/EAD
U.S. EPA (MC-4303)
401 M Street, SW
Washington, DC 20460
(202) 260-5408
Fax: (202) 260-7185
Kahn.Henry@epamail.epa.gov

Douglas Kendall
U.S. EPA Region VIII
NEIC/OECA, Building 53, Box 25227
Denver Federal Center
Denver, CO 80225
(303) 236-5132 x281
Fax: (303) 236-5116
Kendall.Douglas@epamail.epa.gov

Mel Kollander
Temple University
Institute for Survey Research
2300 M Street, NW, Suite 800
Washington, DC 20037
(202) 973-2820
Fax: (202) 973-2821
Melk@gwis2.circ.gwu.edu

Martin Kulldorff
National Cancer Institute
Biometry Branch, DCPC
EPN 344, 6130 Executive Boulevard
Bethesda, MD 20892
(301) 496-7519
Fax: (301) 402-0816
MartinK@helix.nih.gov

Raphael Kuznetsovski
SAIC/Reston Facility Directory
11251 Roger Bacon Drive
Reston, VA 20190
(703) 318-4553
Fax: (703) 709-1040
Rkuznetsovski@lan813.ehsg.saic.com

James R. Lee
American University
School of International Service
Washington, DC 20016
(202) 885-1691
Fax: (202) 885-2494
Jlee@American.edu

Matthew Lorber
ORD/NCEA
U.S. EPA (8623)
401 M Street, SW
Washington, DC 20460
(202) 260-3924
Fax: (202) 260-6370
Lorber.Matthew@epamail.epa.gov

Arthur Lubin
U.S. EPA Region V
Office of Strategic Environmental Analysis
77 West Jackson Boulevard
Chicago, IL 60604-3507
(312) 886-6226
Fax: (312) 353-0374

Elizabeth Margosches
OPPTS/OPPT
U.S. EPA (7403)
401 M Street, SW
Washington, DC 20460
(202) 260-1511
Fax: (202) 260-1279
Margosches.Elizabeth@epamail.epa.gov

Mary Marion
OPPTS/OPP/HED
U.S. EPA
401 M Street, SW
Washington, DC 20460
(703) 308-2854
Marion.Mary@epamail.epa.gov

Thomas Mathew
University of Maryland
Department of Mathematics and Statistics
1000 Hilltop Circle
Baltimore, MD 21250
(410) 455-2418
Fax: (410) 455-1066
Mathew@umbc2.umbc.edu

Steven P. Millard
Probability, Statistics and Information (PSI)
7723 44th Avenue, NE
Seattle, WA 98115-5117
(206) 528-4877
Fax: (206) 528-4802
Smillard@nwlink.com

David Mintz
OAR/OAQPS
U.S. EPA (MD-14)
Research Triangle Park, NC 27711
(919) 541-5224
Fax: (919) 541-1903
Mintz.David@epamail.epa.gov

Nagaraj Neerchal
University of Maryland Baltimore Campus
Department of Mathematics and Statistics
1000 Hilltop Circle
Baltimore, MD 21250
(410) 455-2637
Fax: (410) 455-1066
Nagaraj@math.umbc.edu

Barry Nussbaum
OPPE/CES
U.S. EPA (2163)
401 M Street, SW
Washington, DC 20460
(202) 260-1493
Fax: (202) 460-4968
Nussbaum.Barry@epamail.epa.gov

Brenda Odom
ORD/QAD
U.S. EPA (8724)
401 M Street, SW
Washington, DC 20460
(202) 260-8194
Fax: (202) 401-7002
Odom.Brenda@epamail.epa.gov

G. Patil
The Pennsylvania State University
Center for Statistical Ecology and
Environmental Statistics
421 Thomas Building
University Park, PA 16802
(814) 865-9442
Fax: (814) 863-7114
Gpp@stat.psu.edu

Hugh Pettigrew
OPPTS/OPP
U.S. EPA (MC-7509C)
401 M Street, SW
Washington, DC 20460
(703) 305-5699
Fax: (703) 305-5147
Pettigrew.Hugh@epamail.epa.gov

Andrea Pfahles-Hutchens
OPPTS/OPPT
U.S. EPA (7403)
401 M Street, SW
Washington, DC 20460
(202) 260-0288
Fax: (202) 260-1279

Paul Pinsky
ORD/NCEA
U.S. EPA (8623)
401 M Street, SW
Washington, DC 20460
(202) 260-1079
Fax: (202) 260-3803
Pinsky.Paul@epamail.epa.gov

Esperanza Renard
ORD/NCERQA/QAD
U.S. EPA (MS-104)
2890 Woodbridge Avenue
Edison, NJ 08837
(908) 321-4355
Fax: (908) 321-6640
Renard.Esperanza@epamail.epa.gov

David Rocke
University of California, Davis
Graduate School of Management
Davis, CA 95616
(916) 752-7368
Fax: (916) 752-2924
dmrocke@ucdavis.edu

Randall Romig, Ph.D.
U.S. EPA Region VI
(6MD-HX)
1445 Ross Avenue
Dallas, TX 75202-2733
(214) 665-8346
Fax: (214) 665-8072
Romig.Randall@epamail.epa.gov

N. Phillip Ross
OPPE/OSPED/CES
U.S. EPA Room 3101 M (2163)
401 M Street, SW
Washington, DC 20460
(202) 260-5244
Fax: (202) 260-8550
Ross.Nphillip@epamail.epa.gov

Robert Runyon
U.S. EPA, Region II
2890 Woodbridge Avenue
Edison, NJ 08837
(908) 321-6645
Fax: (908) 906-6824
Runyon.Robert@epamail.epa.gov

Judy Schmid
ORD/NHEERL
U.S. EPA (MD-55)
ERC
Research Triangle Park, NC 27711
(919) 541-0486
Fax: (919) 541-5394
Schmid.Judy@epamail.epa.gov

Mark Schmidt
OAR/OAQPS/EMAD/AQTAG
U.S. EPA (MD-14)
AQTAG, EMAD
Research Triangle Park, NC 27711
(919) 541-2416
Schmidt.Mark@epamail.epa.gov

R. Woodrow Setzer
ORD/NHEERL
U.S. EPA (MD-55)
Research Triangle Park, NC 27711
(919) 541-0128
Fax: (919) 541-5394
Setzer.Woodrow@epamail.epa.gov

Ronald Shafer
OPPE
U.S. EPA (2163)
401 M Street, SW
Washington, DC 20460
(202) 260-6766
Fax: (202) 260-4968
Shafer.Ronald@epamail.epa.gov

Bimal Sinha
OPPE/CES
U.S. EPA (2163)
401 M Street, SW
Washington, DC 20460
(202) 260-2680

Marla Smith
OW
U.S. EPA (4303)
401 M Street, SW
Washington, DC 20460
(202) 260-8639
Fax: (202) 260-7185
Smith.Marla@epamail.epa.gov

William P. Smith
OPPE/CES
U.S. EPA Room 3201 (2163)
401 M Street, SW
Washington, DC 20460
(202) 260-2697
Fax: (202) 260-4968
Smith.Will@epamail.epa.gov

Kathleen Stralka
SAIC
11251 Roger Bacon Drive
Reston, VA 20190
(703) 318-4583
Kathleen.A.Stralka@cpmx.saic.com

Linda Teuschler
ORD/NCEA-CIN
U.S. EPA (MS-190)
26 W. Martin Luther King Drive
Cincinnati, OH 45268
(513) 569-7573
Fax: (513) 569-7916
Teuschler.Linda@epamail.epa.gov

John Warren
ORD/NCERQA/QAD (8724)
U.S. EPA (8724)
401 M Street, SW
Washington, DC 20460
(202) 260-9464
Fax: (202) 401-7992
Warren.John@epamail.epa.gov

Charles White
OW/OST/EAD
U.S. EPA (MC-4303)
401 M Street, SW
Washington, DC 20460
(202) 260-5411
Fax: (202) 260-7185
White.Chuck@epamail.epa.gov

Conference Support Staff

Patricia Crocker
SRA Technologies, Inc.
8110 Gatehouse Road, Suite 600W
Falls Church, VA 22042
(703) 205-8500
Fax: (703) 205-6260

Marcia Gardner
SRA Technologies, Inc.
8110 Gatehouse Road, Suite 600W
Falls Church, VA 22042
(703) 205-8500
Fax: (703) 205-6260
Marcia.Gardner@sratech.com

Maryce Jacobs
SRA Technologies, Inc.
8110 Gatehouse Road, Suite 600W
Falls Church, VA 22042
(703) 205-8500
Fax: (703) 205-6260

Hale Vandemer
SRA Technologies, Inc.
8110 Gatehouse Road, Suite 600W
Falls Church, VA 22042
(703) 205-8500
Fax: (703) 205-6260

ABSTRACTS

Index of Presentations Listed Alphabetically by Presenter(s)

Note: Complete abstracts for each conference presentation appear on the pages that follow. These include the name and type of session, and the date and time of presentation (in the upper right hand corner of the page), as well as the title of the presentation, the name and affiliation of each author, and the name and affiliation of each presenter.

Title: EnvironmentalStats for S-PLUS: Software for Environmental Statistics

Author: Steven P. Millard, Ph.D., Probability Statistics & Information (PSI)

Presenters: Steven Millard, PSI, and Nagaraj Neerchal, Department of Mathematics and Statistics, University of Maryland Baltimore Campus

Abstract

S-PLUS is a premier statistics and graphics software package that is rapidly being adopted by practitioners in fields ranging from pharmaceuticals to finance. ENVIRONMENTAL STATS for S-PLUS is a new S-PLUS module designed specifically for environmental statistics. Developed over the past three years, it covers all the major statistical methods found in the environmental monitoring literature and includes an extensively detailed hypertext help system to guide you through the background and application of each method. This training course will cover basic ideas in sampling design and statistical methods for environmental monitoring and risk assessment, including methods of random sampling, probability distributions, hypothesis tests and confidence intervals, prediction and tolerance intervals, and methods for dealing with Type I left-censored ("below-detection-limit") data. Concepts will be illustrated with data sets taken from current regulatory guidance documents.

Title: Atlas of Cancer Mortality in the United States, 1970-92

Authors: Susan S. Devesa, Ph.D., Dan J. Grauman, M.A., William J. Blot, Ph.D.*, Robert N. Hoover, M.D., and Joseph F. Fraumeni, Jr., M.D., Epidemiology and Biostatistics Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892

*Currently with the International Epidemiology Institute, Ltd., Rockville, MD 20850

Presenter: Susan Devesa, NCI

Abstract

The study of geographic variation in cancer rates may provide clues to the role of environmental or lifestyle factors that may affect cancer risk. The maps themselves cannot provide information about the causes of cancer or its clustering, but they can raise hypotheses about potential causative influences. Earlier atlases showed substantial geographic variations in cancer mortality rates among whites and nonwhites in the United States and stimulated subsequent studies which identified relevant exposures and risk factors. For some cancers mortality rates have not changed greatly over time, whereas substantial increase or decreases have been observed for other cancers. This atlas updates the maps through 1992, presenting for the first time, data specifically for blacks. During the 23-year study period 1970-92, more than 8.5 million whites and 1.0 million blacks died due to cancer. The national annual age-adjusted mortality rate per 100,000 person-years for all cancers combined ranged from 135 among white females to 292 among black males. A total of 40 cancers (including all forms combined) were considered. Some examples of maps from the new atlas will be presented. The patterns of cancer in the United States, some of which have changed over time, may provide additional leads for the evaluation of the determinants of cancer among American men and women.

JNCI, Jan 1995 trends

Title: Evaluating Disease Cluster Alarms

Author/Presenter: Martin Kulldorff, Epidemiology and Biostatistics Program, National Cancer Institute

Abstract

During the last few decades, there have been a considerable number of geographical disease cluster alarms in different parts of the United States. Many are given considerable media attention, and, for natural reasons, there is a considerable amount of worry in the local communities affected. As regards the cause of the clusters, the environment is often a prime suspect.

Before moving into a full-scale epidemiological and environmental investigation, though, it makes sense to find out whether the observed number of cases actually represents a statistically significant excess or not. We cannot simply compare the disease rate inside and outside of the cluster area, since we then have a problem of pre-selection bias. In this talk we will review and illustrate a couple of mutually complimentary methods that can be used to work around that bias, one of which is the spatial scan statistic. A number of applications will be given.

Different Aspects of a Disease Cluster Evaluation

Pre-Selection Bias:
Texas Sharp Shooter Procedure:
- choice of area
- choice of time period
- .. age-group
- .. disease

Three Approaches to Overcome Pre-Selection Bias
- Find Similar Health Hazard Elsewhere + use a Focused Test
- :

Spatial Scan Statistic:

For each circle
- Obtain actual + expected number of cases inside and outside the circle.
- Calculate Likelihood Function

Compare circles:
Pick circle with highest likelihood Function as Most Likely Cluster.

http://www.dcpc.nci.nih.gov/BB/Software.html"

Introduction: Spatial Disease Clusters: Detection and Inference, Statistics in Medicine 1998, 14, 799-810, with Nagwawalla

Title: Representativeness in Statistics and Quality Assurance

Author: John Warren, Quality Assurance Division, Office of Research and Development (ORD)

Presenters: John Warren, ORD, and Malcolm J. Bertoni, Research Triangle Institute (RTI)

Abstract

The concept of "representativeness" is quite clear to a statistician, especially in the context of survey sampling with respect to a well-defined frame. The concept is considerably less clear when the context is environmental sampling because the homogeneity of sampled media and physical environment from which the sample is drawn must be considered.

The session will explore the differing concepts of "representativeness" as used (and possibly abused) by the environmental community, include a discussion of Gy's theory of sampling as a possible solution, and finally engage the attendees in a free and frank discussion of further aspects of the concept.

Title: Toxic Severity for a Useful and Understandable Benchmark Dose

Authors: Linda Teuschler and Richard C. Hertzberg, Ecological Exposure Research Division, Office of Research and Development, Cincinnati, OH

Presenter: Linda Teuschler, ORD

Abstract

Regression on ordered categories of toxic severity is recommended in order to address two criticisms of EPA's risk assessment procedures for noncancer effects. The first criticism is that presenting risk only as probability does not consider the impact of the event. Second, the goal of the benchmark dose is vaguely defined, in part because it focuses on one effect from one study. By including all reported effects into the regression procedure and tracking the toxic severity, one ends up with a benchmark dose that closely follows the definition of the Reference Dose. In addition, by keeping distinct the effects of different severity, categorical regression allows for a definition of a benchmark dose that satisfies both a low specified risk of minor effects and an even lower specified risk of major effects.

Title: Severity Analysis Using Ridits

Author/Presenter: Mary Marion, Health Effects Division, Office of Prevention, Pesticides, and Toxic Substances

Abstract

The United States Environmental Protection Agency, Office of Prevention, Pesticides, and Toxic Substances, Office of Pesticide Programs has been given the task of reviewing chemical registrant data and analyses, some of which use the statistical technique of ridits. The technique of ridit analysis used in severity analysis was studied for its feasibility for use at the Agency.

Two toxicological data sets chosen were that of one study evaluating the severity of glomerulonephropathy in male rat kidneys with dose increments of the chemical being reviewed and another of mononuclear cell leukemia, also in male rats.

The mathematical theory behind this technique will be presented. This is a continuation of a paper presented in 1995 at the poster session of the SUGI 21 Conference held in Chicago, Illinois.

Title: Interpreting Data from a National Survey of Protozoan Pathogens in Drinking-water Sources

Author/Presenter: John F. Fox, Engineering and Analysis Division, Office of Water

Abstract

In 1997-98, EPA and participating water treatment systems will conduct a nationwide sampling program to assess protozoa (Giardia and Cryptosporidium) in drinking-water sources (untreated, raw water) and, at a smaller number of systems, in the treated drinking water. Several hundred participating treatment plants will each submit one sample per month for 12-18 months. The chief objective of the protozoan sampling program is to characterize the nationwide distribution of protozoan concentrations in source water, with the treatment plant as the unit of sampling, in particular the distribution of plant mean, median, and 90th percentile concentrations. A related problem is to characterize the variability and distribution over time of concentrations at one plant. This presentation will discuss opportunities and challenges in developing appropriate point and interval estimates from these data to achieve national-level characterizations of protozoan concentrations in raw and treated water. About one year remains before interim analysis of data. We welcome suggestions regarding data analysis and interpretation!

Title: Relationships Between Dioxins in Soil, Air, Ash, and Emissions from a Municipal Solid Waste Incinerator Emitting Large Amounts of Dioxins

Author: Matthew Lorber, National Center for Environmental Assessment (NCEA), Office of Research and Development (ORD)

Presenters: Matthew Lorber and Paul Pinsky, NCEA, ORD

Abstract

Environmental measurements including air concentrations and soil concentrations of dioxins were taken in the vicinity of a municipal solid waste incinerator emitting large amounts of dioxins. Also available were two separate stack tests measuring concentrations and amounts of dioxins being emitted, and concentrations in combuster ash. An "incinerator signature," defined as the profile of the 17 toxic dioxin and furan congeners where each is described in proportion to total dioxins, was found in the ash and in subsets of the other two matrices. The profiles in all media were also examined using principal component analysis to determine what features best distinguished the profiles in each media. This study also investigated the relationship of dioxin soil concentration as a function of distance from the incinerator, and determined an urban background soil concentration, further from the incinerator, as compared to elevated soil concentrations near the incinerator. A background urban air concentration was determined and compared to measurements of elevated air concentrations, which also had the signature profile.

Title: Statistical Modeling of Dioxin Concentration Data from Sediment Cores

Authors: Paul F. Pinsky and David Cleverly, National Center for Environmental Assessment, Office of Research and Development (ORD)

Presenter: Paul Pinsky, ORD

Abstract

Evidence from several sources suggests that emissions of dioxins into the environment began to stabilize in the 60's or 70's and have been declining since the 70's or 80's. One of the most important of these sources is the historical record from sediment cores in U.S. lakes. Recently, a joint EPA and DOE study measured levels of dioxins and coplanar PCB's in the sediment core of 11 U.S. lakes. Samples from different sediment layers were dated, effectively transforming the data from each lake from a spatial series to a time series. The resulting data base consists of a large number of time series (11 lakes times 30 concentrations of related chemicals) with each time series being relatively short (5 to 11 time points). In this session, we will describe a modeling strategy for these data and interpret the modeling results with the aim of summarizing overall trends as well as identifying any trends specific to certain lakes or chemicals.

Title: EPA Cooperative Agreements

Author/Chair: Barry Nussbaum, Office of Policy, Planning and Evaluation

Participants: Larry Cox, Office of Research and Development, Peter Guttorp, University of Washington, and G.P. Patil, Penn State University

Abstract

This panel discussion will feature investigators from two of the major cooperative agreements on environmental statistics. The panel will discuss the use of cooperative agreements such as these to encourage statistical research on theoretical and applied environmental topics. There will be general comments by EPA on how to get tasks funded and work initiated. Then professors from two of the universities with such agreements will discuss their side of the equation: how they operate under the agreement and what they do. Included will be the vision for future work and applications. The panel will also have time for a hopefully lively question and answer period.

Title: Spatial Sample Design

Author/Presenter: Evan Englund, National Exposure Research Laboratory, Office of Research and Development, Las Vegas

Abstract

Spatial samples, in addition to having number, referred to by classical statisticians as sample size, also have sample support or sample volume or mass. QUAMS, thanks to Dean Neptune, represents this concept by sample unit, remediation unit, and exposure unit. The support, since it cannot be analyzed chemically in total, must be represented by a composite sample in which the subsamples survey the *in situ* sample unit. The definitions and methods of obtaining spatial representativeness will be presented verbally (many "real world" examples and few equations). The relationships of support size and change of support to spatial variance and regularization of semivariograms for correct varicography will be explained. The methodological "rules of thumb" for spatial sample design will be enumerated, clarified, and organized.

Title: Skewed Frequency Distributions

Author/Presenter: George Flatman, National Exposure Research Laboratory, Office of Research and Development, Las Vegas

Abstract

The frequency distribution of both random variables and spatial variables has the ubiquitous problem of skewness for data interpreters and decision makers. Presenting the mean of a skewed distribution is disinformation to all data interpreters or managers (RPM or OSC) if they assume normality. The appropriate model for skewed frequency distributions may be a mixture (plume mixed with background) model rather than one lognormal model. When does a simplifying model become an over simplification? The mixture model does a better job at explaining most waste sites. Methods of separation, such as QQ-plots and robust methods, will be discussed. The various methods of evaluating a lognormal mean will be evaluated and illustrated by real world data and by virtual (simulated) data. The number of questions will exceed the number of answers.

Title: Confidence Regions and Tests in a Calibration Problem

Author/Presenter: Thomas Mathew, Department of Mathematics and Statistics, University of Maryland Baltimore County

Abstract

Consider a univariate normally distributed response variable related to a univariate explanatory variable through the usual linear regression model. Suppose independent observations are available on the response variable corresponding to known values of the explanatory variable. Now consider another observation on the response variable, corresponding to an unknown value of the explanatory variable. The problem of calibration or inverse regression deals with statistical inference on this unknown parameter. The data on the response variable, corresponding to known values of the explanatory variable is referred to as calibration data. We will address the problem of constructing confidence regions and hypotheses tests for the unknown value of the explanatory variable. Two types of problems will be studied: the calibration data is used to construct confidence regions and to test for a single unknown value of the explanatory variable, or for a sequence of unknown values of the explanatory variable. The computational aspects and the practical implementation of our procedures will be illustrated in detail by applying them to some chemical and environmental data.

Title: Recent Developments in the Estimation of U.S. Fish Consumption

Authors: Henry D. Kahn and Helen Jacobs, Environmental Analysis Division, Office of Water, Kathleen Stralka, Science Applications International Corporation

Presenter: Henry Kahn, OW

Abstract

Estimates of U.S. per capita fish consumption play a key role in a number of Environmental Protection Agency program decisions. In particular, exposure estimates used in determining water quality criteria and related standards are based, in part, on estimates of the amount of fish consumed and contamination levels in the fish. This presentation will report on estimates of fish consumption based on recent work with the USDA's combined 1989, 1990, and 1991 Continuing Survey of Food Intake by Individuals (CSFII). These estimates reflect adjustments based on USDA's Recipe file which provides the amount of fish in combination foods and changes in the habitat designations (freshwater/estuarine and marine) for certain species of fish.

Title: Using the Web and Other Networking Technologies in Supporting SAS for the Enterprise

Authors: Chapman Gleason, Center for Environmental Statistics, Office of Policy, Planning and Evaluation, and John Shirey, Enterprise Technology Services Division, Office of Administration and Resource Management

Presenter: Chapman Gleason, CES, OPPE

Abstract

The Environmental Protection Agency (EPA) has just begun an Enterprise Computing Offer (ECO) with SAS Institute. The EPA SAS ECO provides 21 SAS products (base, AF, Assist, ETS, Connect, FSP, Graph, Share, Tutor, Stat, IML, Insight, Lab, Access for Oracle, Access for ODBC, CPE, GIS, QC, Toolkit) on several desktop operating systems (Windows, Windows 95, Windows NT, MacOS, SunOS, Digital Unix, OSF1, HP/UX, DG/UX) in EPA. This product mix will allow SAS users to design and develop client/server SAS applications and provide EPA scientists and policy analysts with better desktop scientific, data management, and statistical software. This session describes EPA's implementation strategy to support SAS across a heterogeneous LAN/WAN computing environment consisting of more than 300 Novell servers and LANs running IPX protocol, Windows PCs on Novell LANs running TCP/IP and IPX protocols, Unix workstations and servers (running TCP/IP protocol), and an IBM mainframe housed at the National Computer Center located in Research Triangle Park, North Carolina. All the computers are accessible via SAS from the Desktop using TCP/IP protocol. The session will include discussion of how EPA:

1) Prepared custom installation instructions for SAS on EPA's Novell LANs which run Networked MS Windows.
2) Pkziped the SAS Windows Installation CD-ROM and set up an FTP Server for SAS to distribute SAS to users on Novell's LANs.
3) Designed and implemented a Lotus Notes Mail-In Data Base and billing strategy to keep track of the user population.
4) Implemented a SAS Listserver, called EPASAS-L, to allow users to share SAS technical problems and solutions.
5) Designed an Internal SAS Web using a Lotus Notes InterNotes server and Data Base which replicates and publishes to the Web each hour. This Lotus Notes Data Base is replicated to each EPA Region allowing SAS users at remote sites to document their implementation of SAS products, SAS applications, and SAS code and share it with other EPA SAS users.
6) Implemented a mail user-ID for the SAS Notes DB, so that users without Notes Clients can mail a document (including Graphics) to a user-ID called epasas Web@epamail.epa.gov was, and the document will automatically be published to the EPA SAS Web.
7) Implemented the SAS and Lotus Notes Interface allowing SAS programs to write to the SAS/Web via SAS clients on remote Systems.

Abstract (continued)

One of the benefits of client/server computing and the popularization of Internet protocols has been the rapid development of the World Wide Web (WWW). However, HTML development has languished because of single file names being required in HTML "home pages." One product that has overcome that barrier and EPA has used to implement its SAS Web is a Lotus Notes InterNotes Server. An InterNotes Server is a Notes server that runs under Windows NT Advanced Server and has the HTTP demon running as an NT service. The InterNotes Server takes a Notes Data Base and converts the Notes Documents into HTML documents and publishes the Notes Views as HTML links to the Notes Documents. EPA has used this capability to save SAS users and developers the learning curve while learning HTML, which is both tedious and time consuming. InterNotes also allows a "macro" level of integration of keeping track of hundreds of HTML file names which are prevalent on Unix systems.

Title: Pesticide Residue Monitoring Data

Author/Presenter: Ed Brandt, Economic Analysis Branch, Office of Pesticide Programs

Abstract

The Government Performance and Results Act requires all government agencies to connect the process of planning, budgeting, and accountability. This paper addresses the issues concerning pesticide residue monitoring data. Using National residue data from 1992 to 1995, the paper analyzes the consistency between different residue monitoring programs, identifies gaps in the development of national estimates of dietary exposure, and suggests approaches to better sampling strategies in the future to improve overall dietary exposure estimates.

Title: A Master Sampling Frame for the Collection of Non-Agricultural Pesticide Usage Data

Author/Presenter: Alan R. Goozner, Economic Analysis Branch, Biological and Economic Analysis Division, Office of Pesticides Programs

Abstract

The EPA recently conducted the 1993 Certified Commercial Pesticide Applicator Survey. The survey was conducted at considerable cost. Much of the time involvement was the construction of a sampling frame. As a follow-on to this experience, several questions arose: Could a master sampling frame be constructed that would allow quicker, more efficient replication of a similar survey? Would it allow surveying more specialized aspects of the applicator population? EPA statisticians are encouraged to offer their insights and opinions as to the feasibility of the idea.

Should the EPA offer seed money to have this frame constructed in the Private sector? Would private sector research companies use such a frame? Would they pay for samples drawn from such a frame? Would the frame facilitate more research into the aspects of non-agricultural pesticide usage that would otherwise not be done? At a minimum, should the EPA more fully investigate the feasibility of frame construction and usability?

Title: The National Air Quality and Emissions Trends Report, 1995

Author/Presenter: David Mintz, Office of Air Quality Planning and Standards, Office of Air and Radiation

Abstract

This twenty-third annual report documenting air pollution trends in the United States was released by Administrator Carol Browner at a major press conference on December 17, 1996. The report provides information on those pollutants for which National Ambient Air Quality Standards have been established. These pollutants are carbon monoxide (CO), lead (Pb), nitrogen dioxide (NO_2), ozone (O_3), particulate matter whose aerodynamic size is less than or equal to 10 microns (PM-10), and sulfur dioxide (SO_2).

While the report focuses on national trends in air quality concentrations and emissions for these criteria pollutants, it also features information on related topics. These include visibility, air toxics, nonattainment areas, urban area trends, reformulated gasoline, and Photochemical Assessment Monitoring Stations (PAMS).

Title: Estimation of Precision of Low Concentration Chemical Analytical Measurements and Establishment of Detection and Quantification Limits

Authors: Henry D. Kahn, Chief, Statistical Analysis Section, Office of Water, and Kathleen Stralka, Statistician, Science Applications International Corporation (SAIC)

Presenter: Henry Kahn, EPA, and Kathleen Stralka, SAIC

Abstract

Estimates of precision of low concentration chemical analytical measurements are critical to establishing detection and quantification levels. This presentation will consider estimates of precision based on the EPA procedure for determining a "Method Detection Limit" and the Rocke-Lorenzato model. The methods will be illustrated using some inductively coupled plasma - mass spectroscopy data and applications to establishing detection and quantification levels will be discussed.

Title: A Two Component Model for Error in Analytical Chemistry and Issues of Detection and Quantification

Author/Presenter: David M. Rocke, Director, Center for Statistics in Science and Technology, University of California - Davis

Abstract

A new model for measurement error in analytical chemistry will be presented. A commonly used model that assumes the standard deviation of analytical error increases proportionally with the concentration of the analyte cannot be used for very low concentrations. For measurements of near zero amounts, the standard deviation is often assumed to be constant, which does not apply to larger quantities. Neither model applies across the full range of concentrations of an analyte. The new model contains two error components, one additive and one multiplicative, and exhibits sensible behavior at both low and high concentration levels. The use of the model with maximum likelihood estimation and application to some gas chromatography/mass-spectrometry and atomic absorption spectroscopy data will be described. Implications for detection and quantification will be discussed.

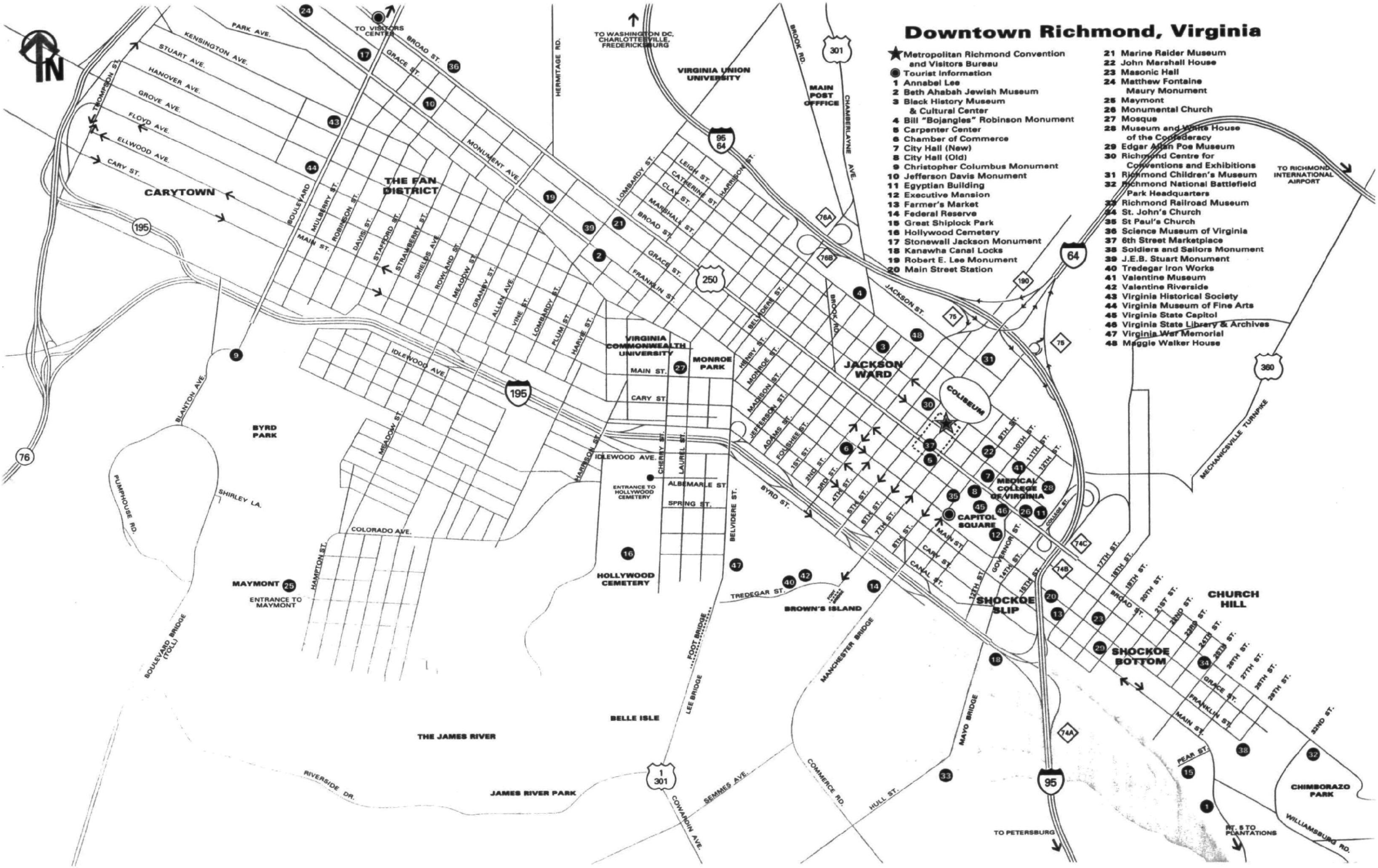
Title: Statistical Graphics for Environmental Applications: Developments and Challenges

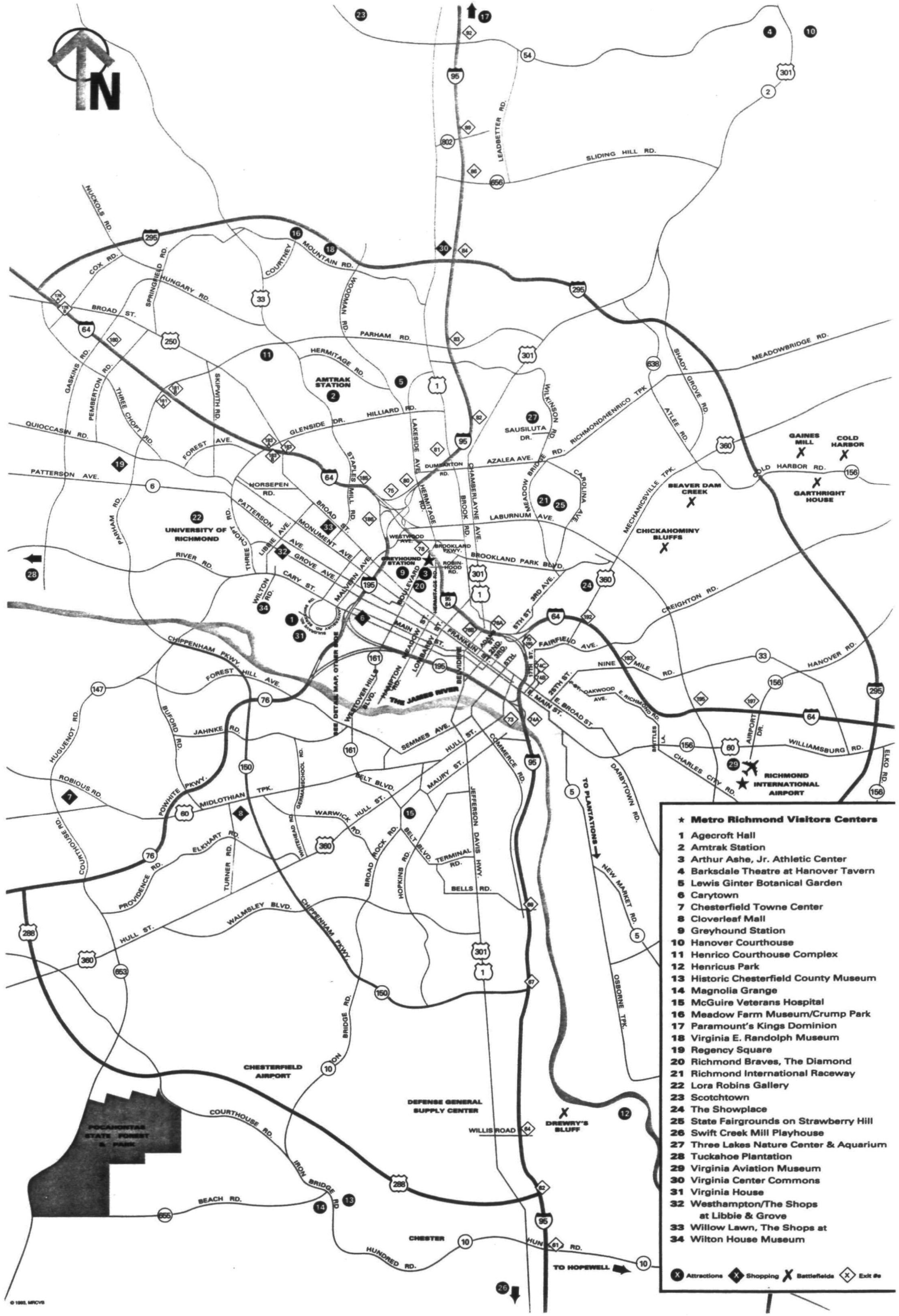
Author/Presenter: Daniel B. Carr, School of Information Technology and Engineering, George Mason University

Abstract

Development of statistical graphics for environmental applications is a many faceted challenge. In the first part of this session, recently developed templates for communicating environmental summaries to broad audiences will be presented. The templates address issues such as converting tables to plots, linking statistical summaries and maps, and representing metadata to provide an appropriate basis for interpretation. Also, a JAVA implementation that enables user manipulation in a low-resolution dynamic web environment will be described. The second part of the session will focus on graphics challenge areas. For example, one challenge area involves working with massive data sets. An example uses the gridding of breeding bird prevalence data to a continental U.S. EMAP grid to raise issues about global gridding of satellite imagery. The second challenge area concerns visualizing statistical and ecological models and their impact on a specific analysis. Recent developments in environmental graphics provide important new capabilities, but some deep challenges remain.

EVALUATION


Downtown Richmond, Virginia
★ Metropolitan Richmond Convention and Visitors Bureau
◉ Tourist Information
1 Annabel Lee
2 Beth Ahabah Jewish Museum
3 Black History Museum & Cultural Center
4 Bill "Bojangles" Robinson Monument
5 Carpenter Center
6 Chamber of Commerce
7 City Hall (New)
8 City Hall (Old)
9 Christopher Columbus Monument
10 Jefferson Davis Monument
11 Egyptian Building
12 Executive Mansion
13 Farmer's Market
14 Federal Reserve
15 Great Shiplock Park
16 Hollywood Cemetery
17 Stonewall Jackson Monument
18 Kanawha Canal Locks
19 Robert E. Lee Monument
20 Main Street Station
21 Marine Raider Museum
22 John Marshall House
23 Masonic Hall
24 Matthew Fontaine Maury Monument
25 Maymont
26 Monumental Church
27 Mosque
28 Museum and White House of the Confederacy
29 Edgar Allan Poe Museum
30 Richmond Centre for Conventions and Exhibitions
31 Richmond Children's Museum
32 Richmond National Battlefield Park Headquarters
33 Richmond Railroad Museum
34 St. John's Church
35 St Paul's Church
36 Science Museum of Virginia
37 6th Street Marketplace
38 Soldiers and Sailors Monument
39 J.E.B. Stuart Monument
40 Tredegar Iron Works
41 Valentine Museum
42 Valentine Riverside
43 Virginia Historical Society
44 Virginia Museum of Fine Arts
45 Virginia State Capitol
46 Virginia State Library & Archives
47 Virginia War Memorial
48 Maggie Walker House
TO VISITORS CENTER
TO WASHINGTON DC, CHARLOTTESVILLE, FREDERICKSBURG
TO RICHMOND INTERNATIONAL AIRPORT
TO PETERSBURG
RT. 5 TO PLANTATIONS
CARYTOWN
THE FAN DISTRICT
VIRGINIA UNION UNIVERSITY
MAIN POST OFFFICE
VIRGINIA COMMONWEALTH UNIVERSITY
MONROE PARK
JACKSON WARD
COLISEUM
MEDICAL COLLEGE OF VIRGINIA
CAPITOL SQUARE
SHOCKOE SLIP
SHOCKOE BOTTOM
CHURCH HILL
CHIMBORAZO PARK
BYRD PARK
MAYMONT
ENTRANCE TO MAYMONT
ENTRANCE TO HOLLYWOOD CEMETERY
HOLLYWOOD CEMETERY
BROWN'S ISLAND
BELLE ISLE
THE JAMES RIVER
JAMES RIVER PARK
PARK AVE.
KENSINGTON AVE.
STUART AVE.
HANOVER AVE.
GROVE AVE.
FLOYD AVE.
ELLWOOD AVE.
CARY ST.
MAIN ST.
BROAD ST.
GRACE ST.
FRANKLIN ST.
MONUMENT AVE.
THOMPSON ST.
BOULEVARD
MULBERRY ST.
ROBINSON ST.
DAVIS ST.
STAFFORD ST.
STRAWBERRY ST.
SHIELDS AVE.
ROWLAND ST.
MEADOW ST.
GRANBY ST.
ALLEN AVE.
VINE ST.
LOMBARDY ST.
PLUM ST.
HARVIE ST.
HERMITAGE RD.
BROOK RD.
CHAMBERLAYNE AVE.
LEIGH ST.
CATHERINE ST.
CLAY ST.
MARSHALL ST.
HARRISON ST.
BELVIDERE ST.
HENRY ST.
MONROE ST.
MADISON ST.
JEFFERSON ST.
ADAMS ST.
FOUSHEE ST.
1ST ST.
2ND ST.
3RD ST.
4TH ST.
5TH ST.
6TH ST.
7TH ST.
8TH ST.
9TH ST.
10TH ST.
11TH ST.
12TH ST.
14TH ST.
15TH ST.
17TH ST.
18TH ST.
19TH ST.
20TH ST.
21ST ST.
22ND ST.
23RD ST.
24TH ST.
25TH ST.
26TH ST.
27TH ST.
28TH ST.
29TH ST.
32ND ST.
GOVERNOR ST.
COLLEGE ST.
JACKSON ST.
CANAL ST.
BYRD ST.
IDLEWOOD AVE.
BLANTON AVE.
PUMPHOUSE RD.
SHIRLEY LA.
COLORADO AVE.
HAMPTON ST.
CHERRY ST.
LAUREL ST.
ALBEMARLE ST.
SPRING ST.
TREDEGAR ST.
FOOT BRIDGE
LEE BRIDGE
MANCHESTER BRIDGE
MAYO BRIDGE
BOULEVARD BRIDGE (TOLL)
RIVERSIDE DR.
COWARDIN AVE.
SEMMES AVE.
COMMERCE RD.
HULL ST.
PEAR ST.
WILLIAMSBURG RD.
MECHANICSVILLE TURNPIKE
76
195
250
301
95 64
64
95
360
76A
76B
74A
74B
74C
75
190



N
★ Metro Richmond Visitors Centers
1 Agecroft Hall
2 Amtrak Station
3 Arthur Ashe, Jr. Athletic Center
4 Barksdale Theatre at Hanover Tavern
5 Lewis Ginter Botanical Garden
6 Carytown
7 Chesterfield Towne Center
8 Cloverleaf Mall
9 Greyhound Station
10 Hanover Courthouse
11 Henrico Courthouse Complex
12 Henricus Park
13 Historic Chesterfield County Museum
14 Magnolia Grange
15 McGuire Veterans Hospital
16 Meadow Farm Museum/Crump Park
17 Paramount's Kings Dominion
18 Virginia E. Randolph Museum
19 Regency Square
20 Richmond Braves, The Diamond
21 Richmond International Raceway
22 Lora Robins Gallery
23 Scotchtown
24 The Showplace
25 State Fairgrounds on Strawberry Hill
26 Swift Creek Mill Playhouse
27 Three Lakes Nature Center & Aquarium
28 Tuckahoe Plantation
29 Virginia Aviation Museum
30 Virginia Center Commons
31 Virginia House
32 Westhampton/The Shops at Libbie & Grove
33 Willow Lawn, The Shops at
34 Wilton House Museum
Attractions
Shopping
Battlefields
Exit #s
AMTRAK STATION
UNIVERSITY OF RICHMOND
GREYHOUND STATION
THE JAMES RIVER
GAINES MILL
COLD HARBOR
BEAVER DAM CREEK
GARTHRIGHT HOUSE
CHICKAHOMINY BLUFFS
RICHMOND INTERNATIONAL AIRPORT
CHESTERFIELD AIRPORT
DEFENSE GENERAL SUPPLY CENTER
DREWRY'S BLUFF
POCAHONTAS STATE FOREST & PARK
CHESTER
TO PLANTATIONS
TO HOPEWELL
SEE DETAIL MAP, OTHER SIDE

DINING downtown

A Guide to Restaurants in Metro Richmond

downtown, including historic church hill, shockoe slip and shockoe bottom

DOWNTOWN

• **Allies American Grill.** The Richmond Marriott, 500 East Broad Street, 643-3400. **B/L/D/WB.** Steaks, salads, pasta, sandwiches. Open 7 days a week: Breakfast: 6:30a.m.-11a.m Lunch: 11 a.m.-2p.m. Dinner: 5 p.m.-10p.m Sunday Brunch Buffet: 7a.m.-noon. $3.50-$16.95.

• **Apollo Italian Restaurant:** 703 East Broad Street, 649-7070. *Italian* **L/D.** Take out. Lunch, dinner, Monday-Friday: 9a.m.-8p.m.; Saturday: 10a.m.-8p.m. Closed Sunday. $2-$16. No max group size.

• **Becky's:** 100 E. Cary Street, 643-9736. **B/L+.** Southern style cooking including homemade soups, sandwiches. Breakfast and lunch: 7a.m.-3p.m. Drinks and lite fare only from 3p.m.-10:30p m. Take out available. $3 50-$4.95.

• **Bill's Barbecue:** 700 E. Main Street, 643-9857. **B/L/D.** Famous pork barbecue, strawberry pie, limeades. Soup and salad bar. Open 7 days a week. Breakfast: 7a.m.-10.30a.m.; lunch/dinner: till 9p.m. $1.50-$5. Maximum group size: 180 (call in advance).

• **Blue Point Seafood Restaurant:** 550 East Grace Street, 783-8138. *Seafood.* **L/D.** Fresh seafood flown from Boston, pasta, mixed grill. Lunch: $4.95-$11.25; dinner: $8 95-$17. Catering for private parties, receptions; max. group size. 60; 1,000 with use of attached Sixth Street Marketplace. Lunch and dinner Monday-Saturday, 11:30a.m.-4p.m. and 4p.m.-9p.m. Open Sundays for adjacent Carpenter Center performances or other large events downtown. Children's menu.

• **Cafe Olé:** - 2 North 6th Street, 225-8226. *Mexican.* **B/L/D.** California-style burritos, quesadillas & taco salads. Breakfast. Monday-Friday, 8a.m-10·30a.m. Lunch. Monday-Tuesday, 11:30a.m.-3:30p.m. Lunch/Dinner. Wednesday-Thursday, 11:30a.m.-7.30p m. Closed Saturday-Sunday. Breakfast from $3, Lunch/Dinner from $5.

Casablanca: 6 East Grace Street, 648-2040. *Standard American Fare* **L/D.** Sandwiches, salads, burgers. Open 7 days a week. Lunch: 11a.m.-2a m. Ample portions, pool table. $5-$8.

• **Chez Foushee** 203 North Foushee Street, 648-3225 *Eclectic* **B/L/D.** Monday-Friday. Continental breakfast· 9a m -11a m. Lunch: 11a.m -3p.m Dinner Tapas bar· 4p m -9p.m. Soups, sandwiches, prepared salads and tapas menus, also boxed lunch, full-service catering & private parties, weddings, corporate meetings. $3-$10

• **China Gourmet.** 204 East Grace Street, 788-8888 *Chinese* **Lunch only.** Monday-Friday, 11a.m -6p m., Saturday, Noon-5p.m. Closed Sun. Average check $5. Maximum group size: 25. Reservations required for party of 10 or more.

• **Cross Roads Restaurant & Lounge:** 217 West Clay Street, 643-2060. *French/Cajun* **D/WB.** Deep Southern/New Orleans Jazz & Blues. Dinner: Wednesday-Sunday, 5p.m.-9p.m. Sunday Brunch: 11a.m.-2p.m. Closed Monday-Tuesday. Dinner $5-$15; children & senior specials. Maximum group size. 85. Live Jazz Show 9p.m.-2a.m. Tuesday-Saturday Christian Jazz, 2p.m -6p.m. Sunday. Reservations recommended.

• **DJ's Fresh Garden Cafe:** 701 East Franklin Street, 643-6592. *Deli/Bakery.* **B/L.** Hot lunch specials, catering for parties, cookies, cakes. Monday-Friday, 7a.m.-4p.m. Closed Saturday-Sunday. Breakfast $1-$3, Lunch $3.89-$4.65. Max group size: 42.

• **The French Quarters Restaurant:** 421 East Franklin Street, 643-1268. *French* **L/D/WB.** Continental French cuisine Open 7 days a week. Lunch· Monday-Friday, 11:30a.m.-2p.m. Dinner: 5:30p.m.-10p.m., Saturday, till 11p.m. Sunday Brunch: 11:30a.m.-2p.m. Lunch $3.95-$11.50, Dinner $12.95-$24.95. Max group size 175. Reservations recommended.

• **Fu Kim:** 515 East Main Street, 780-2999. *Chinese/Vietnamese.* **Lunch only.** Monday-Saturday, 11a.m.-2p.m. Closed Sunday. $1-$5.

• **Homemades by Suzanne:** 10 South 6th Street, 775-2117. *Boxed Lunches/Bakery/Deli.* **B/L.** Homemade breads, salads, soups & desserts, delivery available. Continental Breakfast, Lunch: Monday-Friday 9a.m.-3p.m. Closed Saturday-Sunday. Breakfast $1.50-$3.50. Lunch $6.25-$8.50.

• **J.P. Crowder's Deli:** 305 Brook Road, 648-2565. *BBQ/Deli/Home Cooking* **B/L.** Take out, country Smithfield hams, sandwiches. Monday-Saturday, 6:30a.m.-4p.m.

• **Just Willie's Cafe:** 6 North 6th Street, 643-9330. *Home Cooking.* Fresh turkey, baked ham, homemade chicken salad, soup. **Lunch only.** Monday-Friday, 11a.m.-3p.m. Closed Saturday-Sunday. $2.50-$4.95 Max group size: 10.

• **Lemaire Restaurant:** - The Jefferson Hotel, Franklin & Adams Streets, 788-8000 ext. 6366 - *Regional/VA.* **B/L/D/WB.** Richmond's only AAA 5-Diamond Restaurant. Upscale, 7 private dining rooms, extensive wines. Breakfast/buffet: Monday-Friday 6:30a.m.-10a.m., Saturday-Sunday till 11a.m. Lunch: Monday-Friday, Noon-2p.m. Dinner: Monday-Saturday, 5.30a.m.-10p.m. Breakfast $10, children $7.95, Lunch $14, children $9.95, Dinner $34, children $18. Maximum group size: 75. Reservations recommended.

• **Linden Row Inn:** 100 East Franklin Street, 783-7000. *Southern Cusine* **B/L/D/WB.** Open 7 days a week. Chef's specials, steaks, pastas, fish, exceptional crabcakes; also patio dining at this antebellum landmark, a property of the National Trust for Historic Preservation. Continental breakfast. Monday-Friday, 7a.m.-10:30a.m., Saturday, 7:30a.m.-10.30a.m. Lunch: Monday-Sunday, 11:30a.m.-2:30p.m. Dinner: Monday-Sunday, 5:30p.m.-10p.m. Sunday Brunch: 11:30a.m.-2:30p.m .Breakfast: $4.25-$5.95, Lunch $6.25-$12.95, Dinner $14.95-$22.95. Reservations recommended.

• **Nick's:** 707 East Main Street, 644-1212. *Home Cooking* **B/L.** Boxed lunches, catering, lowfat/fat free menu available. Breakfast: Monday-Friday 6:30a.m.-2:30p.m. Lunch: Monday-Friday, 10a.m.-2:30p.m. Closed Saturday-Sunday. $2.50-$4.50.

• **Ocean Restaurant:** 414 East Main Street, 649-3456. *Home Cooking.* **B/L.** Breakfast: Monday-Friday, 7a.m.-11a.m. Lunch: Monday-Friday, 11a.m.-2:30p.m. Closed Saturday-Sunday. Breakfast $1.25-$2.99, Lunch $1.25-$3.95. Maximum group size: 25.

• **Padow's Hams & Deli:** - 1110 East Main Street, 648-4267. *Deli.* **B/L.** Specializing in Smithfield country, honey glazed & spiral sliced hams in-store & mail order, plus sandwiches, prepared salads, soups, take-out and eat-in. Monday-Friday, 7a.m.-5p.m. Closed Saturday-Sunday. Breakfast from $1; lunch from $2.50.

• **The Pavilion:** Crowne Plaza Hotel, 555 Canal Street, 788-0900. **B/L/D/WB.** Steaks, pasta, crab cake platter ($17.95) Open 7 days a week: Breakfast, 6a.m.-11a.m., Sunday Buffet till 10:30a.m. Lunch, 11a.m.-2p.m. Dinner, 5p.m -10p.m. Breakfast to $10 Lunch $3.95-$11.95. Dinner $7.95-$18.95. Banquet facilities available.

• **Penny Lane Pub & Restaurant** - 207 North Seventh Street, 780-1682. **L/D.** Authentic British style pub known for fish and chips, steak and kidney pie, hearty fare. Lunch: Monday-Friday, 11a.m.-2p.m. Dinner: Monday-Saturdy from 5 p.m. Lunch to $10. Dinner: $6.95-$17.95.

• **Perly's Restaurant:** 111 East Grace Street, 649-2779. *Home Cooking/Deli* **B/L.** Breakfast: Monday-Friday 7a.m.-11a.m. Lunch: 11a.m -3p m Closed Saturday-Sunday. Breakfast $1.50-$4.50. Lunch $2.75-$6.85. Max group size 8.

• **Mr. Beauregard's Thai Room:** 103 East Cary Street, 644-2328. *Thai/American cuisine.* **L/D.** Formal/casual dining. Lunch (Thai & American fare): Monday-Saturday, 11a.m.-2:30p.m. Dinner (Thai cuisine): Monday-Thursday, 4:30p.m.-11p.m., Friday-Saturday, 4:30p.m.-12p.m. Sunday, 4:30p.m.-9p.m. Lunch: $5-$10. Dinner: $9-$15. Max group size 190.

• **Pierces Pitt Bar-B-Que** - 1116 East Main Street, 643-0427. *BBQ* **Lunch only.** Boxed lunches, catering, outdoor open pitt BBQ, featuring hand-chopped pork, nbs, chicken; also sanwiches, salads Available for morning, afternoon or evening meetings, parties Lunch: Monday-Friday, 10:30a m.-4:30p m. Closed Saturday-Sunday. $2-$8. Max group size 75

• **The Red Door:** 314 East Grace Street, 649-1588. *Greek/Italian/American* **L/D.** Daily homemade foods & bread, daily specials. Lunch: Monday-Saturday, 10a.m.-5p.m. Dinner, Monday-Saturday, 5p.m.-8p.m Closed Sunday. $1.50-$7.25. Maximum group size 75.

• **Saigon Restaurant:** 903 West Grace Street, 355-6633. *Vietnamese* **L/D.** Lunch. Monday-Friday, 11a.m.-2p.m. Dinner, Monday-Friday,5p.m.-10:30p.m. Saturday, Noon-10p.m. Closed Sunday Lunch $3.75, Dinner $4.95-$10.95. Maximum group size 48.

• **Steve's Restaurant.** 110 North 5th Street, 649-3460. **B/L.** Homemade specials, Italian dishes, corned beef and cabbage. Monday-Friday. Breakfast and lunch: 7a.m.-2.45p.m. $2.99-$5.50.

• **3rd Street Diner:** 218 East Main Street, 788-4750 **B/L/D.** Open 24 hours Basic fare in double-decker diner: burgers, fries, daily specials and greek specials. Breakfast served all day. Prices from about $2.

• **T.J.'s:** The Jefferson Hotel, East Franklin and Adams Streets, 788-2000. **L/D/WB.** Lunch and dinner: Mon-Sat 11a.m.-2a.m. Dinner from 5 p m Lite menu, entrees, salads, sandwiches, pasta, chicken, steaks. $6-$17. Sunday Champagne Brunch (reservations required) at $28.95 per person is from 10:30a.m.-2p.m.

• **Tony's Bar-Be-Que** - 207 North Third Street, 644-8544. **B/L.** All homemade fare, sandwiches, chicken fillet, BBQ. Breakfast and lunch: Monday-Saturday, 6a m -4p m $1 99-$3 99 Closed Sunday

From eclectic and southern cuisine to Indian and Thai restaurants ... from Antebellum homes to converted tobacco warehouses ... Richmond's downtown restaurants appeal to everyone's palette!

• **Ukrop's Fresh Express:** 10th & Main Street Streets, 648-5633. **B/L.** *Deli* Sandwiches, salads, soups, chips, plus "heart healthy" items. Eat in or carry out. Monday-Friday. Breakfast: 7a.m.-10a.m. Lunch. 10a.m - 3:30p.m. $.99-$4.99. Maximum group size 30.

• **Vie de France:** James Center, 1051 East Cary Street, 780-0748. **B/L.** 7:30a.m.-4:30p.m. Monday-Friday. Sandwiches, soups, salads, muffins, bagels. Cafeteria style/self-serve. Sit in or carry out. $.99-$5.99.

• **Wall Street Deli:** 100 North 8th Street, 643-3354 *Deli.* **B/L.** Classic deli serving subs & sandwiches, corned beef, pastrami prepared in house, New York bagels. Monday-Friday, 7a.m.-3p.m. Closed Saturday-Sunday. $3-$7. Maximum group size 38.

• **Winnie's Caribbean Cuisine** - 200 East Main Street, 649-4974. *Carribean.* **L/D.** St. Lucian owner/chef is known for her hot and spicy jerk chicken, conk fritters, crab cakes and tasty Caribbean lemonade, plus lunch and dinner specials. Tropical decor, bright colors, reggae music. Lunch: Monday-Friday, 11a.m.-3p.m. Dinner: 5p.m.-9:30p.m. Saturday (dinner only), 6p.m.-10:30p.m. Lunch $4.50-$7.99. Dinner $5.50-$12.99. Closed Sunday, but will open for large groups with prior arrangement.

CHURCH HILL

• **Annabel Lee Riverboat:** 4400 East Main Street, 644-5700. *Variety.* Riverboat cruise & buffet-style dining with live entertainment & commentary; lunch/brunch/dinner & plantation cruises available. Lunch, Tuesday Plantation Cruise 10a.m.-5:30p.m., Wednesday-Friday, Noon-2p m , Saturday 11a m -1p.m. Dinner Wednesday-Thursday, 7p.m -9:30p.m., Friday-Saturday, 7:30p.m -10:30p m. Sunday, 6p.m -8:30p.m. Closed Monday. Lunch $17.95, children $9.95, Dinner $24.95, children $11.95. Max group size 350.

• **The Hill Cafe.** 2800 East Broad Street, 648-0360 **B/L/D/WB.** Diverse menu including lobster, prime rib, salads, burgers, burritos and quesadillas Breakfast: Monday-Friday, 7a m.-3p.m. Lunch: Monday-Friday, 11a m.-3p.m. Dinner: Tuesday-Sunday, 5:30p.m.-2a m Saturday and Sunday Brunch: 10:30a.m.-3:30p.m. Lunch $4.95-$6.95. Dinner: $5.95-$15.95. Can accommodate large groups.

• **Millie's Diner:** 2603 East Main Street, 643-5512. **L/D/WB.** Globally inspired eclectic menu featuring "fusion" cuisine Menu changes monthly. Lunch: Tuesday-Friday, 11a.m.-2:30p.m. Dinner: Tuesday-Saturday, 5:30p.m -10:30p.m. Sunday dinner till 9p m. Saturday brunch: 10 a.m.-3p m., Sunday brunch 9a.m - 3p.m. Lunch $6-$10. Dinner $14.95-$21.95.

• **Mr. Patrick Henry's Inn:** 2300 East Broad Street, 644-1322. *Continental.* **L/D.** Warm and woody inside of this 1850s row house converted to an inn and restaurant. Garden dining, soups, salads, entrees, chef's specials. Lunch: Monday-Friday, 11:30a.m.-2:30p.m. Dinner: Monday-Saturday, 5:30p.m.-9:30p.m. Lunch: $6-$12. Dinner: $18-$23.

• **Poe's Pub:** 2706 East Main Street, 648-2120. **L/D.** Irish pub atmosphere, best known for its catfish and ribs. Casual dining. $2.95-$15.95. Open 7 days a week for lunch and dinner from 11 a.m.-2 a.m.

SHOCKOE SLIP

• **The Berkeley Hotel Dining Room:** 1200 East Cary Street, 780-1300 **B/L/D.** American, European, voted one of Richmond's best restaurant experiences, extensive wine list, many from Virginia. Breakfast:7a.m.-10:30a.m. Lunch: 11:30a.m-2p.m. Dinner: 6p.m.-10 p.m. Breakfast and lunch: $2.50-$13.50 Dinner:$17. Maximum group size: 20.

• **LaGrotta Restaurant:** 12th & East Cary Streets, 644-2466. *Italian* **L/D.** Voted Best Lunch Spot by *Style Weekly* & voted ••••1/2 stars by *Richmond Times-Dispatch* Lunch: Monday-Friday, 11:30a.m.-2:30p.m. Dinner, Monday-Thursday, 5:30p.m -10p.m., Friday-Saturday, 5:30p.m.-11p.m., Sunday 5p.m.-9p.m. Lunch $6.95-$8.95, Dinner $9.95-$18.95. Max group size: 130.

• **Nana Zushi:** 1309 East Cary Street, 225-8801. *Japanese* **L/D.** Sushi bar and a la carte, terriyaki, tempura dishes. Lunch: Monday-Friday, 11:30a.m.-2p.m. Dinner: Monday-Saturday, 5:30p.m.-10p.m. Lunch $4.95-$9.50. Dinner $6-$12.

• **Peking Pavilion:** 1302 East Cary Street, 649-8888. *Chinese* **L/D/WB.** Northern Chinese cuisine. Lunch: Sunday-Friday, 11:30a.m.-2:15p.m. Dinner, Sunday-Friday, 5p.m -9:45p.m., Saturday, 5p.m.-10:45p.m. Sunday Brunch. Lunch $4-$7, Dinner $7-$14. Max group size 200.

• **Richbrau Brewing Co. and Restaurant:** 1214 East Cary Street, 644-3018. **L/D.** Virginia's original microbrewery and only microbrewery restuarant. All beer is made on premesis. Open 7 days a week. Two floors featuring full service restaurant downstairs and bar with pool and darts upstairs. Menu includes fish and chips, pastas, chef's specials, catch of the day, soups, salads, sanwiches. $3.95-$15. Lunch: 11:30a.m.-4p.m. Dinner from 4 p.m. Children's menu. Large groups accomodated.

• **Sam Miller's Warehouse:** 1210 East Cary Street, 644-5465. *Seafood/Regional VA* **B/L/D.** Breakfast: 10a.m.-5p.m. Lunch: 11a.m.-5p.m. Dinner: 5p.m.-11p.m. Breakfast $5.95-$12.95, Lunch $5-$12, Dinner $11.95-$22.95. Bus parking available. Max group size: 175.

• **Skipjack Tavern & Comedy Club:** 109 South 12th Street, 644-0848. **L/D.** Open 7 days a week. Restaurant features raw bar with clams, oysters and crab legs from owner's Chicoteague oyster farm; plus fish and chips, sandwiches and traditional entrees. Lunch/dinner: 11:30a.m.-2a.m. $5.95-$16.95. Two seatings for weekend Comedy Club. Call for reservations and information.

• **The Slip at Shockoe:** 11 South 12th Street, 643-3313. *Home Cooking/Soul Food.* **B/L/D.** Lunch & dinner buffet, salad bar, sandwiches, beer, wine, mixed beverages. Breakfast: Monday-Friday, 7a.m.-11a.m. Lunch: Monday-Friday, 11a.m.-3:30p.m. Dinner buffet: Friday, 5p.m -9:30p.m., Sunday, 6p.m.-9:30p m. Closed Saturday. Breakfast $ 99-$3.99, Lunch $1.49-$4 99, Dinner $3.95-$8.95. Max group size 125 Dancing Thursday-Saturday, 9p m.-2a.m., Sunday, 9p.m.-1a.m Happy Hour Friday, 5p.m.-9p.m.

The Tobacco Company Restaurant: 1201 East Cary Street, 782-9431. *Continental* **L/D/WB.** Seafood & VA specialties, historic landmark known for prime rib with seconds on the house, live entertainment nightly. Lunch: Monday-Saturday 11:30a.m.-2:30p.m. Dinner: Monday-Friday 5:30p.m.-10:30p.m., Saturday, 5p.m.-11p.m., Sunday, 5:30p.m.-10p.m. Sunday Brunch, 10:30a.m.-2:30p.m. Lunch $2.99-$9.95, children $2.99-$6, Dinner $13.95-$26.95, children $2.99-$10. Max group size 100. Space for catered receptions up to 300.

SHOCKOE BOTTOM

• **Awful Arthur's Seafood Co.:** 101 North 18th Street, 643-1700. *Seafood/Regional/VA* **L/D.** Fresh seafood, rawbar with oysters, clams, crablegs, shrimp & crawfish, daily specials, theme nights. Lunch-Dinner: Monday-Friday, 11:30a.m.-2a.m., Saturday, Noon-2a.m., Sunday, 4p.m.-2a.m. Lunch $5-$8, Dinner $9-$15.

• **Sea Breeze Cafe:** 3 South 15th Street, 649-8516. *Caribbean.* **L/D.** Hot and spicy island food; known for conk fritters and mango shrimp. Lunch: Tuesday-Friday, 11:30-2:30p.m. Dinner: Tuesday-Friday, 5:30p.m.-2a.m. Saturday and Sunday 5p.m.-2a.m. Lunch $3.75-$7.50. Dinner $5-$14. Can accomodate large groups with advance notice.

• **The Bottom Line Tap & Grill:** 1800 East Main Street, 644-5944. *American* **L/D.** Sandwiches/*Pub Grub*, best selection of bottled beer in Bottom. Lunch: Monday-Saturday, Noon-2p.m. Dinner: Monday-Saturday, 5p.m.-2a.m. Closed Sunday. Lunch-Dinner $5-$10. Max group size 6.

• **Bottoms Up Pizza:** 1700 Dock Street, 644-4400 *Pizza* **L/D.** Gourmet pizza, 2 open-air decks, voted best pizza 7 years running. Monday-Wednesday, 11:30a.m.-11pm, Thursday 11:30am-midnight, Fri 11:30a.m.-2a.m., Saturday, Noon-2a.m., Sunday, Noon-midnight. $5-12. Max group size 300.

• **Calypso Cafe:** 1718 East Franklin Street, 225-9776. *Caribbean.* **L/D/WB.** Also seafood, steaks & vegetarian, Caribbean theme (Jimmy Buffet & rum runners), catering, parties, one of city's largest rooftop open air decks. Lunch: Monday-Sunday, 11a.m.-3p.m. Dinner: Monday-Sunday, 4p.m.-10p.m. Plus, Sunday Brunch. Lunch $5-$8, Dinner $6-$12. Max group size 12.

• **Castle Thunder:** 1726 East Main Street, 648-3038. **L/D.** Extensive sandwich menu, new outdoor dining deck on Main Street. Open seven days a week, 11:30 a.m.-2a.m. $4.95-$6.95. Maximum group size: 150.

• **Chaplin's Grill:** 2001 East Franklin Street, 643-7520. Pasta, steaks, Cajun shrimp. Friday-Saturday, 10p.m.-2a.m.

• **Chetti's Cow and Clam Tavern:** 21 N. 17th St. 644-4310. *Seafood.* **Dinner only.** Oldest bar in the Bottom. Shellfish, pasta and steaks served informally. Home of the famous "Moister Oyster" -- shucked oyster, cocktail sauce with a shooter of beer on the side. Dinner: Tuesday-Saturday, 5p.m.-2a.m. Closed Sunday and Monday. $3.25-$15.95. 1/2 price specials on Tuesday. Large groups OK with notice.

• **Cobblestone Brewery & Pub:** 110 North 18th Street, 644-2739. Cajun. New Orleans. Jamaican, steaks.

• **The Frog and the Redneck:** 1423 East Cary Street, 648-3764. *Modern American Regional* **Dinner only.** Consistently rated as one of Southeast's finest restaurants and winner of many awards for excellence. Features great local products including seafood, meats and veggies Celebrity chef Jimmy Snead cooked with Julia Childs on "In Julia's Kitchen with Master Chefs." Dinner: Monday-Friday, 5:30p.m.-10p.m. Saturday, 5pm.-10:30p.m. Will accomodate large groups with advance notice.

• **Goodfellas:** 1722 East Main Street, 643-5022. Progressive rocknroll with state of the art sound system and house D.J. Paul. Wednesday-Saturday: 5p.m.-2a.m.

• **The Hard Shell:** 1411 East Cary Street, 644-5341. *Seafood.* **L/D.** Seafood spot with lobster bar, steak, diverse menu. Lunch: Monday-Saturday, 11:30-2:30p.m. Dinner: Monday-Saturday, 5:30-10:30. Closed sundays. Lunch $4.25-$6.95. Dinner $12.95-$21.95.

• **Havana '59:** 16 North 17th Street, 649-2822 *Cuban/Caribbean.* **Dinner only.** The ultimate in theatrical dining. Cuban cuisine in a re-created 1950's Havana streetscape. Cigar smoking, rooftop dance floor, great fresh juices. 4:30p.m -closing $8-20

• **Homer's Real Sports Grill:** 14 North 18th Street, 643-2222. *American.* Two laser disc video screens Hearty food including fried chicken, Buffalo wings, meatloaf. Dinner: Monday-Friday, 4p.m.-2a m Lunch/Dinner: Saturday and Sunday, 12p m.-2a m.

• **Johnson's Grill:** 1802 E. Franklin St., 648-9788 *Soul Food* No smoking or alcohol Open Monday-Friday. Breakfast 6a.m.-11a.m. Lunch 11a.m -1p.m Closed Saturday-Sunday Breakfast $3.95, Lunch $4-$6.50. Max group size 55.

• **Main Street Grill:** 1700 East Main Street, 644-3969 *Vegetarian/American Grill* **L/D/WB.** Grill by day, vegetarian by night. Open Tuesday-Sunday. Breakfast· 7 a.m -11a.m., Lunch 11a.m.-2.30p.m. Dinner: 6p.m -12p.m. (vegetarian cuisine). Casual dining Breakfast to $5; lunch and dinner around $7.95.

• **Marks:** 1707 East Franklin Street, 649-1079. *American* **L/D.** Sandwiches, homemade chips, pool table, live music on weekends. Lunch/Dinner· Monday, 11:30a.m.-8p.m. Tuesday-Friday, 11:30a.m.-2a.m. Saturday, 5p.m.-2a.m.

• **Medley's:** 1701 East Main Street, 648-2313 *Cajun/Creole.* **L/D/WB.** News Orleans-style Cajun & Blues Bar. Lunch: Monday-Saturday, Noon-3p.m Brunch Sunday, Noon-3p.m. Dinner: Monday-Saturday, 6p.m-11p.m. Appetizers only Monday-Saturday, 3p.m.-6p.m., 11p.m.-2a.m. Lunch $5.95, Dinner $12.95, Appetizers $4.50-$7.50.

• **Moondance Saloon and Restaurant:** 9 North 17th Street, 788-6666. *Southwestern.* **Dinner only.** Cuban chef, blackboard menu, great drinks. Dinner· Tuesday-Saturday, 6p.m.-11p.m. plus afterhours sandwiches till 1a.m. Alternative music night on Monday from 9p.m. featuring college bands and a limited menu $6-$14.95.

• **None Such Place:** 1721 East Franklin Street, 644-0832. *Regional/VA* Traditional VA cuisine using fresh ingredients & classic culinary techniques, housed in oldest commercial building in Richmond. Lunch· Monday-Saturday, 11:30am-3pm, Dinner: Monday-Saturday, from 5:30pm. Closed Sunday Lunch $5.95, Dinner entrees $11.50-$20.95 Max group size 80-100.

• **Rack-n-Roll Cafe:** 1713 East Main Street, 644-1204 *American grill.* Sports bar atmosphere with pool tables, darts, foozeball. Lunch-Dinner: Monday-Wednesday, 11:30a.m.-midnight. Thursday-Friday, 11:30a m -2a.m. Dinner: Saturday-Sunday, 6p.m -2a m Lunch $4.50-$6.50. Dinner $6.50. Maximum group size 300.

• **River City Diner:** 1712 East Main Street, 644-9418. *American* Diner Food with flair, breakfast anytime. Tuesday-Wednesday, 8a.m.-2a.m., Thursday-Friday, 8a.m.-4a.m., Saturday, 24 hours until Sunday 3p.m Closed Monday. Average check $6.25.

• **Rock Bottom Pizza·** 13 North 17th Street, 225-1382. *Pizza.* 70s atmosphere. Wednesday, 9p m -2a m., Thursday-Saturday, 6p.m.-2a.m. Closed Sunday-Tuesday. Max group size 30.

• **Shotz:** 4 North 18th Street, 649-7468. *Deli/Pizza* Fresh cooked pizza & subs, bar crowd after 10p.m., private parties, 21 & over only. Dinner Monday-Saturday, 5p.m.-2a.m. Closed Sunday. Dinner $5. Maximum group size 20.

• **Southern Sugar & Spices:** 2116 East Main Street, 788-4566. *Southern* **B/L/D.** Real, down-home southern cooking including fried chicken, liver and onions, meatloaf, fish, pork chops. Monday-Saturday. Breakfast: 9a.m.-11a.m. Lunch: 11a.m -4p.m. Dinner· 4p.m -9p.m. Breakfast: to $5. Lunch $5-$8. Dinner. $9-$16.

• **Star of India:** 1703 East Franklin Street, 648-5470 *Indian.* **L/D/WB.** Lunch· Monday-Saturday, 11:30a.m.-2:30p.m. Dinner: Monday-Thursday 5p.m -10p.m., Saturday, 5p.m.-11p.m. Lunch buffet weekdays, $5.95. Dinner $7.50-$13. Sunday brunch $7.95

• **Sunset Grill·** 1814 East Main Street, 643-2926. *Surf & Turf* **L/D.** Capitalizes on neighborhood meat market. Chicken, burgers, steaks every which way. Lunch· only in spring & summer months on large outdoor patio. Dinner: Thursday-Saturday, 8p.m-1a.m Lunch $4-6. Dinner: $5-$12 Closed Sunday-Wednesday.

• **Surf Side Grill:** 1714 East Franklin Street, 644-8704 *Seafood* Beach-fresh seafood Lunch-Dinner: Monday-Friday, 11a.m.-10p.m. Dinner: Saturday, 4p m -2a.m. Closed Sunday. Lunch $4.95-$8.95. Dinner $8.99-$17.99. Lunch-Dinner children $4.25.

Publisher assumes no responsibility for the accuracy of information contained herein and advises contacting restaurants directly to confirm information about hours of operation, prices, location, menus, etc. For more information on RICHMOND call (804) 782-2777

In Search of . . .

Environmental Statistician

The United Nations has an opening for an environmental statistician. Salary is $108,000.

For information contact the UN web site: http:\\www.un.org
Press "general information" and then
Enter "UN employment"

United Nations contact person is Patricia Nicolos, (212) 963-5783.

This information was provided by EPA contact, Kathleen Hogan, (202) 260-9349.

United States
Environmental Protection
Agency EPA
Policy, Planning, and Evaluation (2163)

THE TWELFTH ANNUAL EPA CONFERENCE ON ENVIRONMENTAL STATISTICS

Richmond, VA *April 1-3, 1997*

I D E A S ARE NEEDED

?

TO FILL THIS SPACE

Wanted: *Conference Logo*

Theme: *Statistics for the Future*

Reward: *Contact Barry Nussbaum*

EPA TWELFTH ANNUAL CONFERENCE ON ENVIRONMENTAL STATISTICS

SPECIAL *Preview* EDITION

STATISTICS FOR THE FUTURE April 1-3, 1997

RICHMOND, VA *Site of the Twelfth Annual EPA Conference on Environmental Statistics.* "Thought the EPA Conference was supposed to come to town last year," mused a Richmond resident. Well, we didn't make it then, but we're back and looking for a big turnout at this year's conference. Personnel from EPA and other Federal and state agencies will gather south of the Mason-Dixon Line for a two-an- a-half-day conference. The theme is STATISTICS FOR THE FUTURE. The program, focusing on relevant *applications of statistics* in government programs and how to enhance statistical support, will feature hands-on *training sessions* and opportunities to learn about new *statistical techniques and software.* There will be sessions on health statistics, detection limits, water quality, and the use of statistics in Quality Assurance.

The conference's real, underlying *benefit to you* is the opportunity to exchange with others involved in similar programs, with related problems, and on a one-to-one level. Informal sessions, such as the Poster/Technology Session and Roundtable Discussions, provide an atmosphere for sharing information, solving problems, and building a network.

There is plenty of opportunity to *get involved.* Check out the *"Call for Your PARTICIPATION"* ad in this Special Preview Edition. There is no limit to how much involvement and fun you can have. And, from winter weather predictions, you'll want to cut loose and enjoy springtime in the old South at the EPA statistics conference.

THE CONFERENCE IS BACK. Y'ALL COME.

SPECIAL FEATURES

****
No Registration Fee
****
Transportation Provided from EPA Headquarters
****
Costs within Government Per Diem
****
Fulfills Qualifications as Training
****

A Message from the Chairman

WOW, what a year it has been. I'm sure I'm not alone in saying that I've never seen a set of furloughs and travel restrictions that affected us as severely as last year. But a funny thing happened on the way to no forum. You may recall that despite all our money saving techniques, we had to forgo our annual conference on statistics. In order to capitilize on the plans already in progress by some of the professorial types who were developing tutorial sesisons, we decided to hold these training sessions in Washington and RTP. This avoided travel costs and travel restrictions for the attendees from our two major locations. We didn't intend to shut out regional and laboratory folks at other locations, but we had to do the best we could under unusual circumstances. So what happened? We didn't just salvage some sessions, we actually learned that there was a real demand for this training, and a good bit of response came from people who normally didn't attend the annual conference. Imagine my surprise to hear "new" participants asking why they weren't on the list. They had heard about the conference from a colleague down the hall.

So we are applying what we learned. *FIRST,* I have personally arranged that the government will not stop this year. *SECOND,* we are still employing our cost reduction methods to make the travel more palatable to attendees. *THIRD,* and most importantly, we are combining the conference with enriched training in Richmond on April 1-3, 1997 (no fooling!). *FOURTH,* we are adding separate training sessions in the late spring. We think we may have hit on the best of both worlds with this scheme. But it really depends on your participation to make it a real success. So, jump on the band wagon, and participate! Write a paper, present a poster, serve on a panel, and be active. I look forward to seeing you in Richmond.

One last dilemma: If we had to postpone last year's conference, is this the 12th annual conference on statistics, the 13th annual conference on statistics, or the 12th almost annual conference on statistics?; and was Grover Cleveland really the 22nd and the 24th President all by himself? If you can help me with any of this, please call, write, fax, e-mail, etc. Thanks. *BARRY NUSSBAUM*

Emphasis on Training

Response to the series of statistical training programs offered last spring in DC and RTP was tremendous. Courses in *Regression Diagnostics, Information Visualization,* and *SAS Applications* attracted a large and varied audience. Positive feedback on the training programs has led to a greater emphasis on training opportunities at this year's conference as well as training courses to be offered in the late spring and/or early summer of next year.

The Conference offers a variety of *training features,* such as:

✔✔ *Abstracts of all Papers Presented at the Conference*
✔✔ *Training Programs Designed Specifically for EPA Statistical Needs*
✔✔ *Information from Current Publications in Environmental Statistics and Information Science*
✔✔ *Informal Discussions with Other Statistiticians to Focus on Specific Problems and Probable Solutions*

Train for the Future in Statistics

CALL FOR YOUR

PARTICIPATION

(YOU are the conference)

UNCLE SAM and your EPA co-workers can benefit from your experience. . . be a participant in this year's conference. We invite you to:

- ⇨ *Make a Presentation*
- ⇨ *Chair a Session*
- ⇨ *Present a Poster*
- ⇨ *Moderate a Roundtable Discussion*
- ⇨ *Become a Member of the Conference Planning Committe*

OR. . . you may have another idea!

Whatever you would like to do, name it, and contact BARRY NUSSBAUM NOW! by phone at (202) 260-1493 or by fax at (202) 260-4968 or by e-mail at Nussbaum.Barry@epamail.epa.gov

WHY ATTEND

- **Learn latest developments in environmental statistics**
- **Share what YOU are doing**
- **Meet other colleagues**
- **Present a poster; make a presentation**
- **See demonstrations of the latest statistical programs**
- **Get answers to statistical problems**
- **Build team spirit**
- **Receive training in new software, statistical methods, computers**
- **Build a network of statistical and information specialists for the FUTURE**

WHO WILL BE THERE

- **EPA statisticians and survey specialists**
- **EPA developers and users of environmental information and statistics**
- **EPA policy and decision makers**
- **State and local government environmental information developers and users**
- **University experts and students**
- **Special Guest Speakers**
- **YOU**

REGISTRATION

FOR THE *TWELFTH ANNUAL EPA CONFERENCE ON ENVIRONMENTAL STATISTICS*

RICHMOND, VA *APRIL 1-3, 1997*

Complete registration packets will be mailed on

JANUARY 31, 1997

Is your mailing information correct?
Did we miss someone? Do you want to add a colleague to the list?

Contact MARCIA GARDNER
SRA TECHNOLOGIES, INC.
Phone (703) 205-8547, fax (703) 205-6260 or E-mail: MARCIA.GARDNER@sratech.com

EPA
United State
Environmental Protection Agency
(2163)
401 M Street SW.
Washington. DC 20460

Official Business
Penalty for Private Use $300

Margaret G. Conomos
(2163)

Question 1. How large does a group have to be to show health effects from arsenic exposure between 10 and 50 μg/l?

The 1960's Taiwan Epidemiological study studied people exposed to arsenic in drinking water beginning in 1900. Wells ranged from 0.01 to 1.82 ppm (10-1,820 ppb or μg/l).

Doctors physically examined 40,421 people out of 103,154 in 37 villages.
728 cases of skin cancer, 153 histologically confirmed.
72% had hyperkeratosis and 90% had hyperpigmentation.

The control group of 7,500, had an age distribution similar to the study population. Arsenic ranged from non-detect to 0.017 mg/l (17 ppb or μg/l). No skin cancer, hyperkeratosis or hyperpigmentation in the control population. The expected number of skin cancer cases, using the skin cancer rate for Singapore Chinese from 1968-1977 is a little less than 3. Using this as the expected prevalence, the probability of observing no cancer cases is 0.07.

EPA's drinking water criteria is 50 μg/l or 50 ppb. The Taiwan study identified a NOAEL (No observed adverse effects level) of 0.8 μg/kg/day, and a corresponding concentration in drinking water of 10 μg/l.

Question 2: How many infants should be in each concentration range, for a study of sulfates?

The Center for Disease Control (CDC) is proposing to study 1,000 babies exposed to sulfate in their drinking water and compare them against 250 babies not exposed to sulfate. They haven't identified the babies nor the exposure concentration ranges yet. Sulfates cause a laxative effect above 1,000 mg/l, and EPA's proposed drinking water criteria is 500 mg/l, a level at which sulfates aren't expected to be a problem.

CDC's Sample Size Calculations for the planned study are attached.

In 1995 CDC studied 276 infants, and found 39 cases of diarrhea, with a median of 264 mg/l, and a range of 0-1271 mg/l. Non-cases had a median of 260 mg/l, and a range of 0 to 2787 mg/l. However, as seen by the attached graph, there were very few infants being exposed to 500 mg/l or higher.

Question 3: Are 100 participants, divided into 0, 500, 800, and 1200 mg/l (40 per group) enough to establish a dose that causes diarrhea?

In 1994, 4 volunteers drank water with 0, 400, 600, 800, 1000, and 1200 mg/l sulfate at 48 hours. In a follow up study six people drank 1200 mg/l sulfate for six days and didn't report diarrhea.

From Irene Dooley 202/260-9531

Sample Size Calculations[1]

Confidence	Power	Unexposed: Exposed	Disease in Exposed	Risk Ratio	Sample Size		
					Unexposed	Exposed	Total
95%	80%	1 4	13%	1 5	345	1,381	1,726
"	"	"	"	1 6	250	1,001	1,251
"	"	1 5	"	1.5	332	1,662	1,994
"	"	"	"	1.6	241	1,205	1,446
"	"	1:6	"	1.5	324	1,943	2,267
"	"	"	"	1.6	235	1,409	1,644

[1] Using EpiInfo (Version 5 01b DOS) sample size calculations for unmatched cohort and cross-sectional studies (Exposed and Nonexposed).

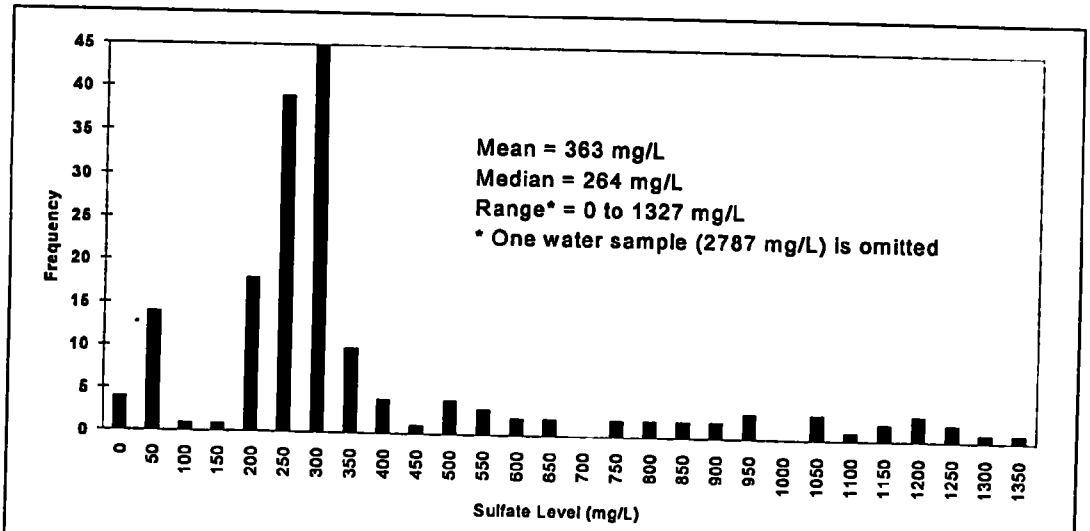


Figure 3. Frequency Distribution of Sulfate Concentration for All Water Samples Submitted: June-October, 1995 (N=172)

SEER*Stat

The SEER*Stat system is a statistical package for the analysis of SEER and other cancer databases. SEER*Stat provides a graphical user interface for the production of the following statistics and statistical tests.

- Frequencies
- Percentages
- Crude (non-adjusted) rates with standard errors and confidence intervals
- Age-adjusted rates with standard errors and confidence intervals
- Trends over time as percent changes, from crude or age-adjusted rates
- Trends over time as estimated annual percent changes, from crude or age-adjusted rates, with confidence intervals
- Comparison of estimated annual percent changes with zero
- Comparison of two estimated annual percent changes with one another

SEER Web Site

Home Page URL: ***http://www-seer.ims.nci.nih.gov/***

The SEER web site contains a variety of information about the SEER program.

Topics areas include:

- News
- About SEER
- Publications
- Online Systems
- Online Data
- Scientific Systems
- Registries
- Other Links

Online Systems

Cancer Query System (CANQUES) on the Web

CANQUES on the Web is an interactive system with a Java interface that allows the user to access a variety of pre-calculated cancer statistics. There are currently in excess of 7.8 million pre-calculated statistics available. CANQUES performs no calculations and contains statistics that were created by the SEER Program for their routine reporting and the Cancer Statistics Review, 1973-1993 You must have a Java enabled browser to use the system and the most recent release of that browser is recommended.

Type of statistics include:

SEER Incidence Rates
SEER Incidence Trends
U.S. Mortality Trends
SEER Median Age at Diagnosis
U.S. Mortality Median Age at Death
NHL and Kaposi's Sarcoma in San Francisco
SEER Relative Survival

cdc homepage for cdc data.

Online Data

SEER Incidence Data - The February 1996 submission of the SEER Incidence database is available in public use text format as self-extracting DOS executables. This data is for the nine standard registries and it covers diagnosis years 1973-1993. (Password encrypted, requires completion of Public Use Data Agreement to extract data. Public Use Data Agreement is available via internet.)

Population Data for the SEER Registries - The populations for the nine standard SEER registries, to be used in conjunction with the above data, are available as self-extracting DOS executables. This data is stored in text format and contains populations for 1973-1993 by individual registry and also by the counties defining each registry.

United States Population Data - County level populations for each state in the U.S. are available as self-extracting DOS executables. Each state file contains county populations by year, 1973-1993 A file containing total United States populations is also available All files are stored in text format.

Scientific Systems

Portable Survival System

The analysis of patient survival plays an integral part in determining many aspects of cancer prevention, control and treatment and is an important part in the interpretation of cancer statistics. Since survival statistics play such an important role in the analysis of cancer data, the NCI previously developed a system which generated survival statistics for researchers. This system is the NCI's Mainframe Survival System which has been in use for over 25 years. A researcher must have access to and a working knowledge of the NIH IBM mainframe system. This places a limitation on the accessibility of the system. Also, repetitive mainframe usage costs are an issue where a single analysis may cost hundreds of dollars depending on the requested parameters.

Information Management Services, Inc., in consultation with the Cancer Statistics Branch of the National Cancer Institute, has developed a new, expanded and portable version of the Mainframe Survival System called the Portable Survival System (PSS). The PSS is a Microsoft Windows-based application which provides more access and greater ease in generating survival statistics than its mainframe counterpart. The PSS retains all the features of the Mainframe Survival System with several additional features The PSS can be installed on most PCs with access to a CD-ROM drive

The NCI and IMS are currently in the process of integrating the PSS with the SEER*Stat system to provide a single application for calculating a wide variety of cancer-related statistics.

The PSS is available on CD-ROM and may be ordered by mailing or faxing a completed Public Use Data Agreement form (available from the SEER Web site) to the NCI.

Applying Gy's Theory of Sampling to Problems of Representativeness in Environmental Field Investigations

Malcolm J. Bertoni

Center for Environmental Measurements and Quality Assurance
Research Triangle Institute

Some Questions to be Answered

- Why consider Gy's Theory of Sampling?
- What are the main concepts of Gy's Theory?
- How does Gy's Theory help address representativeness?
- What are some limitations and questions when applying Gy's theory to environmental field investigations?
- How can I use this information to improve my lot?

Why consider Gy's Theory of Sampling?

- Provides a theoretical and practical link between statistical sampling concepts and physical sample collection protocols
- Helps clarify the relationships between sampling units, sample support, and the scale of inference
- Provides a more sound scientific basis for making measurements/observations of sampling units

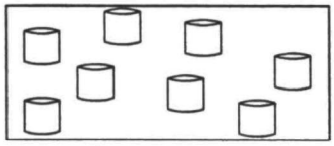
What's the origin of Gy's Theory?

- Pierre Gy, a French mining engineer, developed the theory in the late 1950s through 1970s
- Addresses the estimation of mineral content in ore
- Combines concepts from statistics, physics, geology
- Has been applied to environmental sampling by Pitard, Ramsey, others

Main Concepts from Gy's Theory

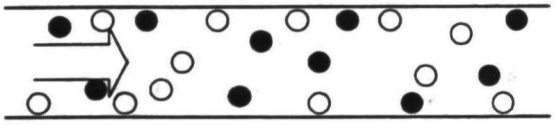
- Types of sampling lots
- Types of heterogeneity
- Classification of errors
- Principles of correct sampling
- Methods for reducing errors

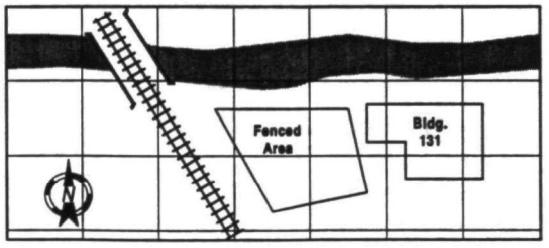
Types of Sampling Lots



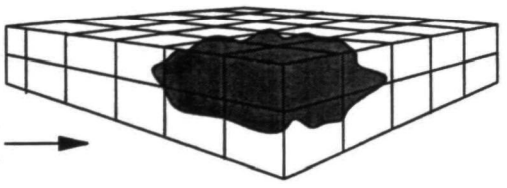
Zero-dimensional

One-dimensional





Two-dimensional



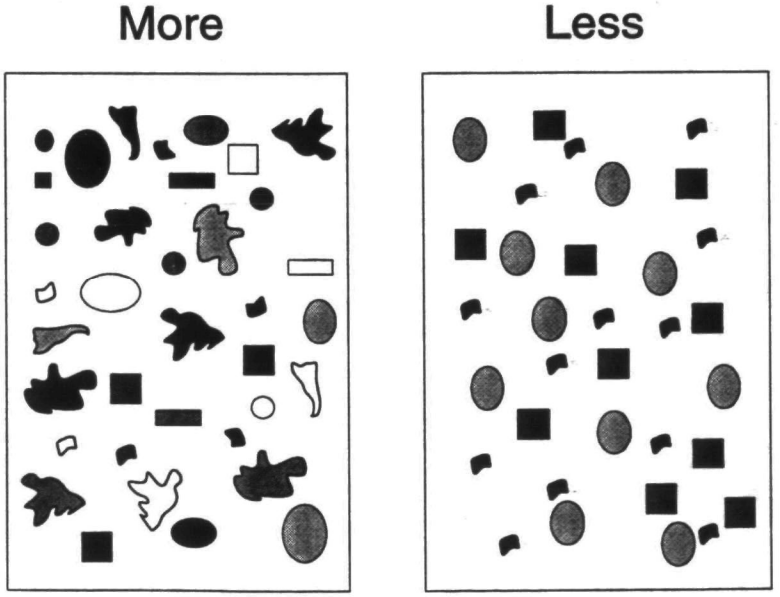
Three-dimensional

Types of Heterogeneity

- Short-Range (random fluctuations)
  - **Constitution heterogeneity**
    - *How many constituents are in the material?*
  - **Distribution heterogeneity**
    - *How are the constituents distributed?*
- Long-Range
  - *Non-random trends, patterns*
- Periodic
  - *Cyclic changes*

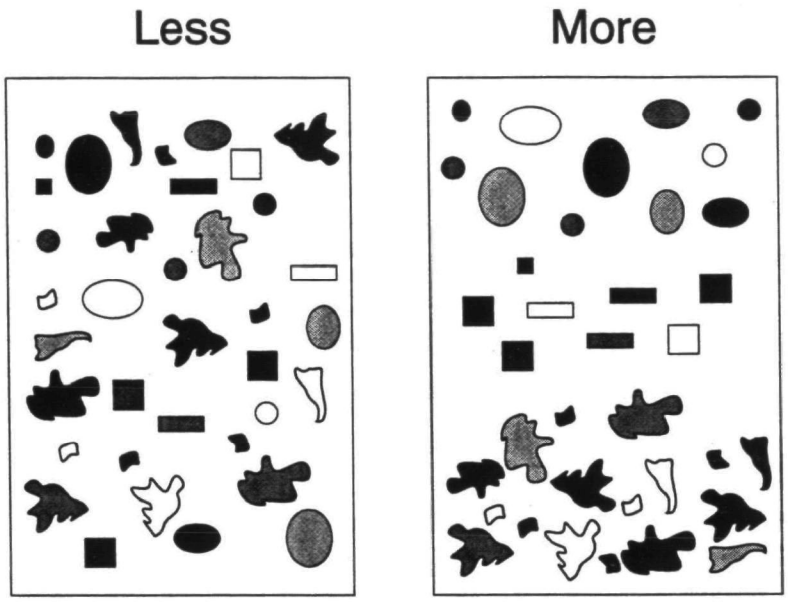
What's in it

Constitution Heterogeneity



Where it's at

Distribution Heterogeneity



How does Gy measure heterogeneity?

- Based on analysis of particles or fragments; extends to groups of particles of fragments
- Interested in the fraction of material having a particular property of interest ("critical content"), expressed as a percent of mass
- Heterogeneity defined in relation to the critical analyte

Heterogeneity of a Particle

$$h'_i = [a_i - a_L]M_i$$

ɔrmalize with repect to average mass of critical analyte:

$$h_i = \frac{[a_i - a_L]M_i N_F}{a_L M_L}$$

where:
- a_i = concentration of particle
- a_L = average concentration of lot
- M_i = mass of particle
- M_L = mass of entire lot
- N_F = number of particles in entire lot

Heterogeneity of a Group

where:

$$h_n = \frac{[a_n - a_L]M_n N_n}{a_L M_L}$$

- a_n = concentration of a group of particles
- M_n = mass of a group of particles
- N_n = number of groups of particles

Definition of Constitution Heterogeneity (CH)

$$CH_L = s^2(h_i) = \frac{1}{N_F} \sum_{i=1}^{N_F} h_i^2$$

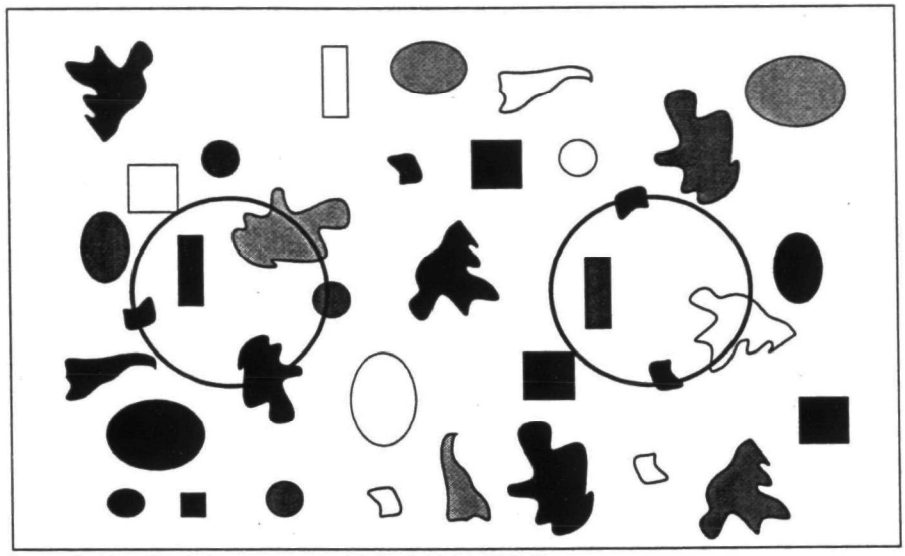
$$CH_L = \frac{N_F \sum_{i=1}^{N_F} [a_i - a_L]^2 M_i^2}{a_L^2 M_L^2}$$

Definition of Distribution Heterogeneity

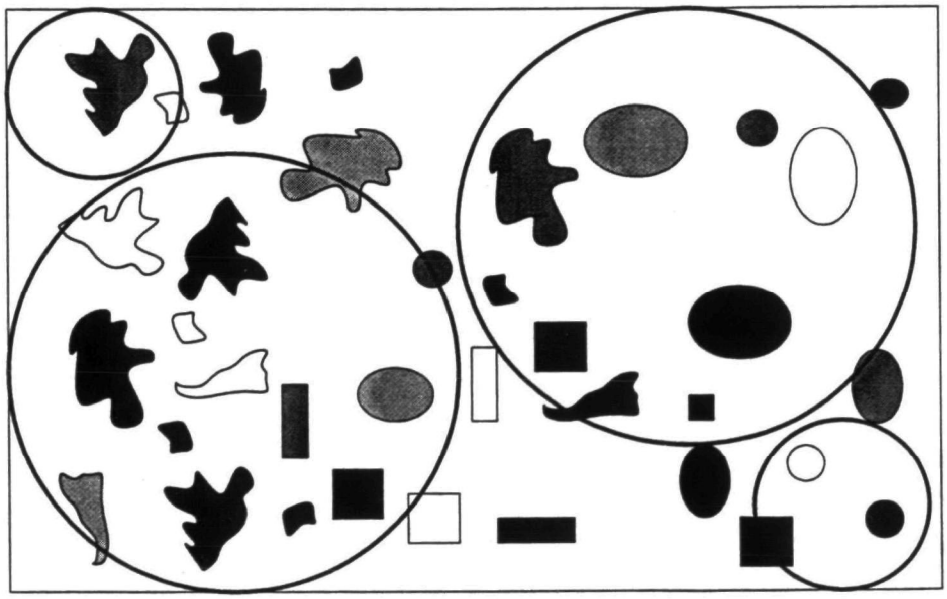
$$DH_L = s^2(h_n) = \frac{1}{N_G} \sum_{n=1}^{N_G} h_n^2$$

$$DH_L = \frac{N_G \sum_{n=1}^{N_G} [a_n - a_L]^2 M_n^2}{a_L^2 M_L^2}$$

DH is defined in terms of possible groupings...



... hence DH is affected by group size



Constant Factor of Constitution Heterogeneity (IH_L)

CH_L is theoretical; it's difficult to estimate, partly due to large N_F term.

Multiplying by the average mass per fragment, $[M_L / N_F]$, eliminates the need to estimate N_F:

$$IH_L = CH_L [M_L / N_F]$$

Constant Factor of Constitution Heterogeneity (IH_L)

IH_L is more practical; can be estimated from observable qualities and measures:

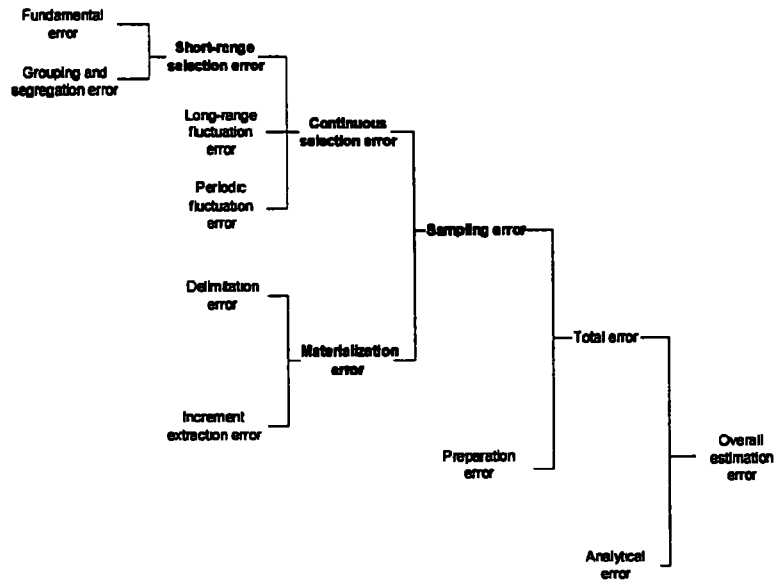
$$IH_L = Cd^3$$

where C = the sampling constant, calculated from several material parameters such as liberation, shape, mineralogical factors;
d = particle diameter.

Why such concern over Heterogeneity?

- Another measure of variability in a population
- Key to understanding and controlling errors in environmental measurements
- Foundation for understanding and applying correct sampling principles

Gy's Classification of Errors



Types of Errors

- Short-range selection error (CE1)
  - **Fundamental error (FE)**
  - **Grouping and segregation error (GE)**
- Long-range fluctuation error (CE2)
- Periodic fluctuation error (CE3)
- Delimitation error (DE)
- Increment extraction error (EE)
- Preparation error (PE)

Fundamental Error (FE)

- Caused by constitution heterogeneity
- Can be estimated *a priori* by studying properties of critical analyte and matrix to be sampled
- Main drivers are:
  - qualities of heterogeneity
  - particle size
  - mass of the sample

Fundamental Error (FE)

$$FE^2 = \left(\frac{1}{M_S} - \frac{1}{M_L} \right) IH_L$$

where M_S = mass of the sample, assuming $M_S << M_L$.

Consequently:

$$FE^2 \approx C \frac{d^3}{M_S}$$

Grouping and Segregation Error (GSE)

- Grouping error introduced when fragments are not selected one at a time (*always!*)
- Segregation error introduced when fragments are not randomly distributed (Distribution heterogeneity)
- Reduce GSE by:
  - generating a sample by taking many increments
  - homogenizing the material when possible
  - selecting random locations for increment extraction

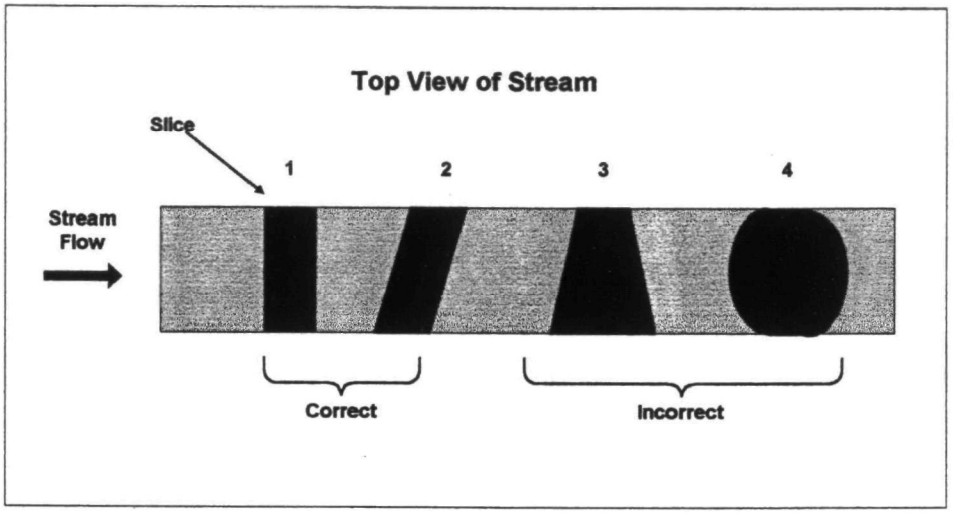
Gy's Theory helps statisticians:

- Choose a sample mass (support) to satisfy FE design constraint for a given particle size and sampling constant
- Reduce FE and/or sample mass through grinding to reduce particle size
- Reduce GSE by specifying, for example, that "10 to 30 increments shall be taken to form a sample"

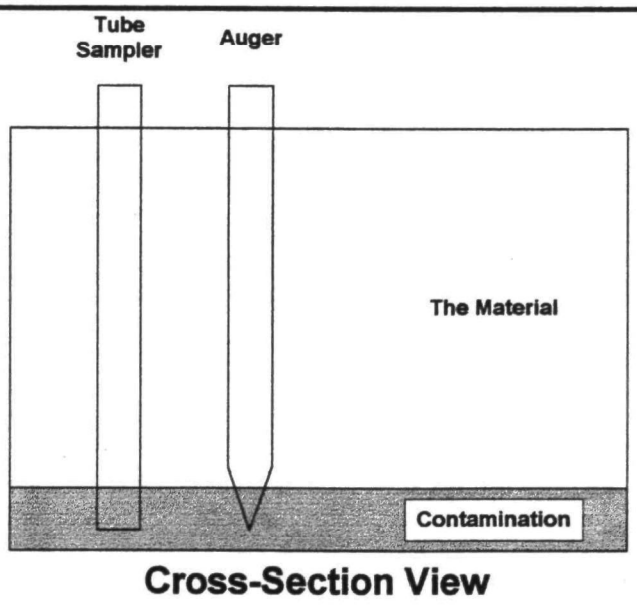
Delimitation Error

- Introduced when incorrect shape and orientation for sample increment is selected
  - design fault
  - equipment selection/specification fault
- Correct shapes:
  - zero dimension -- unit
  - one dimension -- slice
  - two dimensions -- cylinder
  - three dimensions -- sphere or cube

Examples of 1D Delimitations



Examples of 2D Delimitations



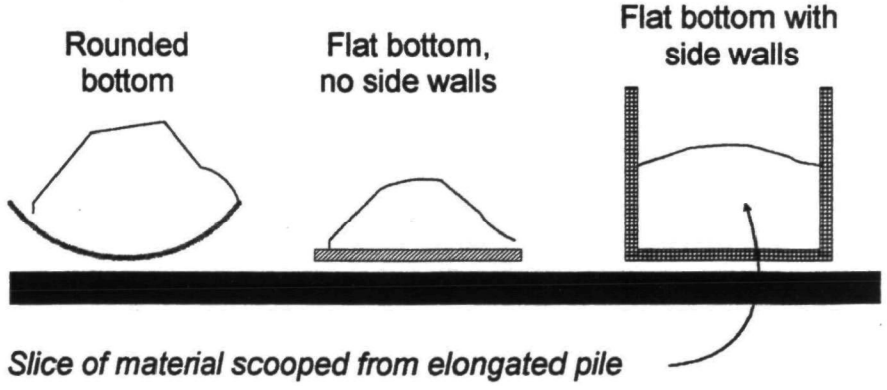
Cross-Section View

Extraction Error

- Introduced when material is imperfectly extracted in relation to the correct delimitation
  - implementation fault
  - equipment selection/specification fault
- Can result in systematic or random error
- Many environmental sampling tools introduce both delimitation and extraction errors

Extraction Error Example

Cross Section Views of Sampling Spoons

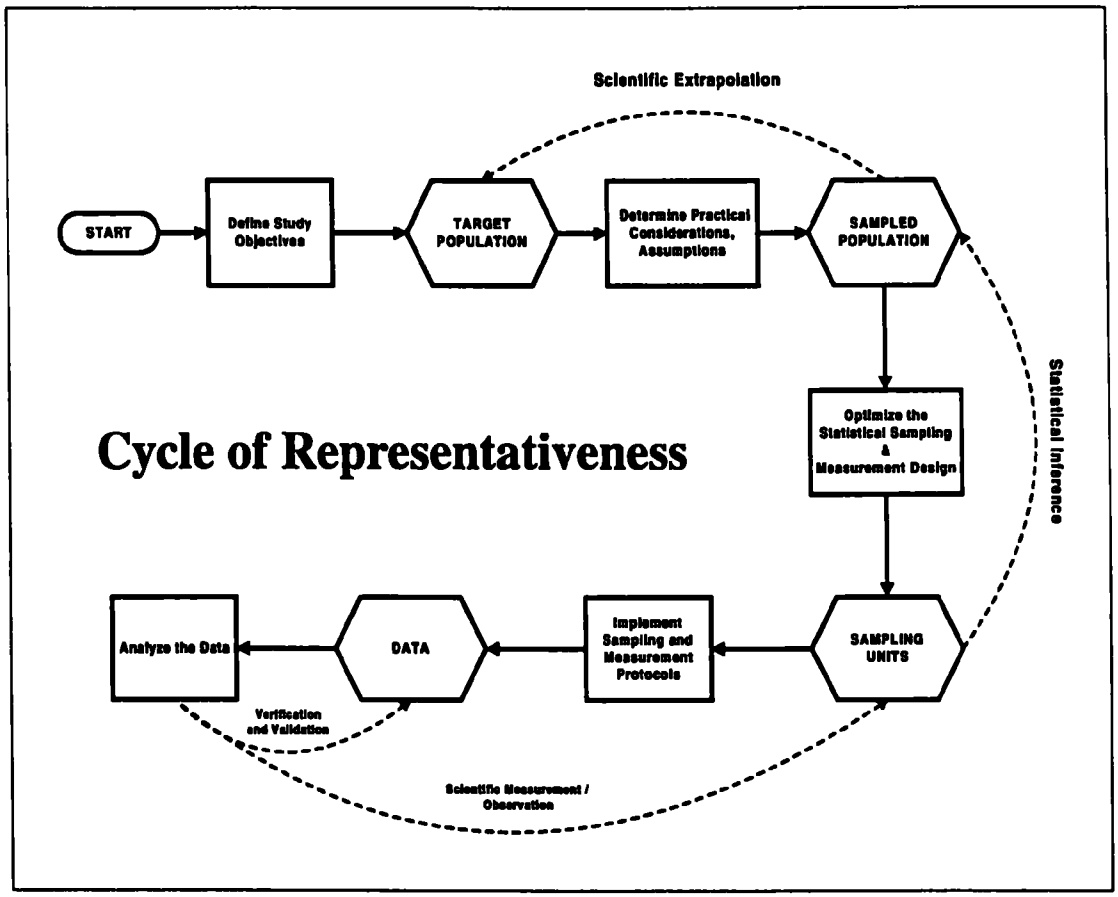


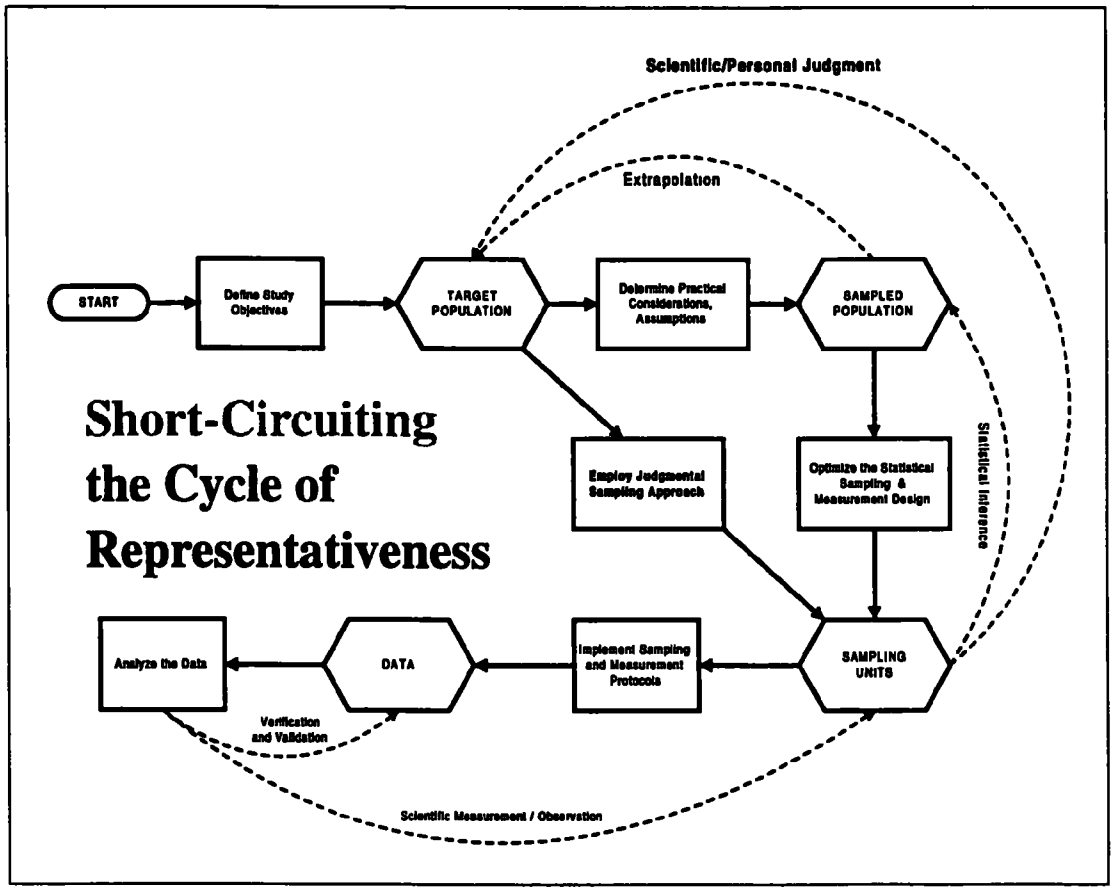
Some limitations and questions

- Gy's Theory based on particle characteristics; environmental sampling often involves media that don't translate well to this model
- Not clear how some chemical contamination applies (e.g., sticky stuff that adheres to particles?)
- Is the average characteristic of the sample always what the investigator wants to know?

How can I use this to improve my lot?

- Design the right measurement protocols (correct delimitation)
- Study the matrix you're sampling
- Increase the mass of the sample
- Take more increments for each sample
- Reduce particle size through grinding (if OK for the material/contaminant)
- Specify correct subsampling protocols


Scientific Extrapolation
START
Define Study Objectives
TARGET POPULATION
Determine Practical Considerations, Assumptions
SAMPLED POPULATION
Statistical Inference
Cycle of Representativeness
Optimize the Statistical Sampling & Measurement Design
Analyze the Data
DATA
Implement Sampling and Measurement Protocols
SAMPLING UNITS
Verification and Validation
Scientific Measurement / Observation


Scientific/Personal Judgment
Extrapolation
START
Define Study Objectives
TARGET POPULATION
Determine Practical Considerations, Assumptions
SAMPLED POPULATION
Statistical Inference
Short-Circuiting the Cycle of Representativeness
Employ Judgmental Sampling Approach
Optimize the Statistical Sampling & Measurement Design
Analyze the Data
DATA
Implement Sampling and Measurement Protocols
SAMPLING UNITS
Verification and Validation
Scientific Measurement / Observation

[1] From: Alan Goozner at DCOPP7 12/27/96 8:55AM (5369 bytes: 85 ln)
To: melko@juno.com at IN
Subject: Master Sampling Frame for Non-Agrcultural Pesticide Research
---------- Forwarded ----------
From: Alan Goozner at DCOPP7 12/19/96 7:57AM (5121 bytes: 85 ln)
To: chlorine-news@igc.apc.org at IN
cc: PEPI LACAYO at X400, BARRY NUSSBAUM at X400, MATTHEW LEOPARD at X400,
Alan Goozner, Rob Esworthy, Edward Brandt
Subject: Master Sampling Frame for Non-Agrcultural Pesticide Research
---------- Message Contents ----------

The EPA and the USDA historically have divided its responsibilities for collection of pesticide usage data where the USDA conducts surveys of farmers for agricultural pesticide usage and the EPA conducts specialized surveys of non-agricultural pesticide usage.

In the past, the EPA conducted the National Home and Garden Pesticide Usage Survey and more recently the Certified Commercial Pesticide Usage Survey. These two surveys were National in scope and cost the Government over a million dollars each to complete.

The EPA is not very well suited for the collection of data. The Office of Pesticide Programs does not have a professional data collection staff and needs to contract out this activity whenever a study is conducted. This requires competing in the private sector for a statistical contractor, the clearance of an information collection request through OMB and preparation of a report that must clear many hurdles before being released to the public. And, by the time the report reaches print, the data can be as much as 2-3 years old.

Needless to say, the private sector can do a much better, more efficient and more timely job in collecting data on . pesticide usage.

In support of this need, the EPA may be in a position to facilitate the collection of more and better pesticide usage data for non-agricultural sites. The idea is to construct a master sampling frame for non-agricultural pesticide usage sample surveys.

If a frame can be constructed and maintained by the EPA, the private sector can request samples from this list to conduct specialized surveys of interest with the intent to share any data with the EPA. The exact consistency of the frame is yet to be determined but it may be composed of two major components of the applicator population: A) Certified Applicators and B) Homeowners.

Experience in conducting the Certified Commercial Pesticide Applicator Survey at the EPA shows that state lists are out of date. Many applicators on state lists have not renewed their license or are no longer actively applying pesticides. If these lists can be cleaned up and screened for certain characteristics that the industry may need to zero in on for future data collection efforts, a highly efficient sampling frame can be constructed. For example, if a National list of pesticide applicators can be

constructed with certain known demographic characteristics and pesticide usage characteristics by types of application work and chemicals used, stratified random samples can zero in to target specific areas of interest for research.

The cost of constructing such a master sampling frame would be prohibitive for any one private organization contemplating a National data collection effort. But, the statistics developed would be more accurate and reliable from a statistical stand point.

The question is: Is this a good idea?
If such a sampling frame was constructed, would your organization use it to collect more/better data on pesticide usage? If used, would it result in a savings in your market research budget? Would it enable better and safer introduction of pesticide products? Would producing more reliable data support the goal of overall pesticide exposure reduction?

You reply and further discussion is encouraged. If there is enough industry support, I am willing to propose this to EPA management in the Pesticides Office as a project. You may want to communicate what specific non-agricultural pesticide usage data collection efforts are underway or being contemplated that may lend itself to using such a master sampling frame. Would use of such a sampling frame result in reduced costs for your organization? How much of a savings would this be on an annual basis?

You may reply directly to:
Goozner.Alan@epamail.epa.gov

Alan R. Goozner, Statistician
USEPA, OPPTS/OPP/BEAD/EAB

Estimating Dietary Exposure to Pesticide Residues

Table of contents

1 Title: Estimating Dietary Exposure to Pesticide Residues

2. Author Ed Brandt, Economist
Economic Analysis Branch
Office of Pesticide Programs 7503W

April 2, 1997 : EPA Statisticians Conference Poster session

3. Abstract: Several new laws have increased the need to estimate aggregate dietary exposures. The Food Protection and Quality Act (FQPA) requires the examination of aggregate exposures for pesticides likely to have additive effects (common modes of action). The Government Performance and Results Act (GPRA) requires all government agencies to reformat the budgeting process to connect measures of program outputs to eventual environmental outcomes . Methodology and results to date are reported concerning the consistency between two major residue monitoring programs, critical data gaps and approaches to future data collection.

4. Statement of problem and approach

a. Increased need for measures of aggregate exposure

i. The importance of a consistent set of residue estimates across pesticides has grown with the passage of FQPA. Previously, decision making for a pesticide focused on whether the residues for the individual pesticide are acceptable.

ii. The need for a national data base on residue data was recommended by the National Academy of Sciences but funding for development has not yet been received.

b. Limitations of existing residue monitoring programs

i. Two major residue monitoring programs are with USDA and FDA The USDA's Pesticide Data Program (PDP) was implemented in May 1991, to provide data on pesticide residues in food to support exposure analyses conducted by EPA in the registration of pesticides.

(1) Principal goal is to measure food safety for vulnerable populations
(2) 1992 to 1995 for selected crops and pesticides -(15 crops and 65 pesticides by 1995) for high consumption to infants and children

and potentially riskier pesticides based on existing tox/exposure data

(3) capture residues most related to actual consumption, i e , oranges include pulp only. The skin is excluded

(4) probability based sample selection at the latest point of distribution.

ii. FDA residue monitoring includes a Surveillance and a Compliance program. Surveillance data not specifically targeted toward known problems of misuse so it tends to be more representative than the Compliance data program which does target producers with past problems.

iii. The Surveillance program has limitations when used to estimate dietary exposure.

(1) Primary role is prevention of illegal residues (over tolerance or no tolerance). The watchdog role limits the flexibility to optimize sampling for estimating dietary exposure alone

(2) Program limited by need to seize shipment in 24 hours if found to be violative. Limits ability to measure residues downstream in the distribution system (post harvest applications) since grower identify is lost.

(3) Monitoring programs designed primarily for enforcement (to ensure the absence of illegal residues) results in small sample sizes on important commodities of high dietary consumption.

(4) Some chemicals not picked up by multi residue methods are omitted altogether because of the incremental costs of inclusion.

c. Government Performance and Results Act of 1993 requires quantitative measures to define goals and objectives

i. Programs must develop plans which connect program outputs to objectives.

5. Suggested measurements for the goal of safe food related to Pesticides

a. Current measures have examined impacts as an indicator of outcomes since so many factors in addition to pesticide exposures influence national health statitics.

b. The following table provides a schematic of the types of measures proposed for each effect level.

Effect level	Items to measure	Measures
Outcomes	cancer(s), neurotoxic effects, endocrine disruption, other toxic effects	national health statistics
Impacts	dietary exposure - residues on food	residue levels percent detects pesticide use
Outputs	New registrations Review of existing registrations	number and type

c. Defining a measure of average annual dietary exposure

i. Limit analysis to variability of an annual national average that is appropriate for lifetime assessments. Not appropriate for an acute or subchronic analysis.

ii. Expected exposure for a residue for chemical x on crop y is a function of the probability of detection multiplied by probability of residue level given detection.

(1) $\sum_{\text{crops}}$ = avg residue per sample*dietary consumption

avg. residue= Prob (any detectable residue on crop x)*Expected residue given detect

iii. These two variables can be combined into a single distribution of the expected residue per sample. Thus, given 1,000 samples of chemical x on crop y, there is an expected residue per sample and a probability distribution of the sample mean.

d. Basis for estimating average residue per sample

i. The mean and variance of the sampling distribution could be determined by knowing the probability of detection (binomial distribution on detects) with a log normal distribution of residues (log normal fits residue data the best, consistent with the constant degradation function modeled by a log normal.

ii. One would expect that percent detect would correlate with percent crop treated, but this is not the case. Other factors, such as time of application,

pesticide formulation with stickers and adherents, degradation rates, weather etc. are thought to be important too. More work is needed on the factors which most affect probability of detection.

iii. There are several alternative ways used to calculate the estimation error of the true residue amount per sample.

(1) From probability of detection and residue distribution given detection. A problem with this approach is that these two variables are not independent. Percent detect is significantly correlated to the residue level and is not correlated with the percent of use

(2) Based on variance of average residue per sample over time. Estimating standard error by sample size and average residue level for each year. Estimated mean is weighted by sample size for each year using a weighted variance estimate.

(3) Calculating percentiles, or in the case of only four years of data analyzed, the range of average residue per sample.

6. Findings of Statistical analyses

a. Descriptive Summaries of residues by pesticide and crop

i. Method 2 and 3 have been calculated but only method 3 is used to construct a table of ranges.

(1) It is easier to understand , does not require assumptions about homogeneity of variances and distribution form, and is closest to existing methods for estimating upper ranges of residue.

(2) Tables are provided in the appendix which summarizes the estimation of average residue per sample per crop.

ii. Analysis of variance indicates that compared to the variance in residue levels among chemicals and crops, there is not a significant difference between years for the same chemical and crop. This makes pooling data across years more appealing to do.

iii. Average residue is further adjusted by a scaler, the average intake per year for infants and children and again for women of childbearing age. Since these tables are rather lengthy, information is summarized again aggregating on either chemical or crop Variance estimates at an aggregate level have not yet been attempted.

b. There is general agreement in the priority ranking between PDP and FDA data for both chemicals and crops, i.e., same chemicals and crops rank the highest with respect to residue exposure.

i. Differences do exist because of food preparation and sampling as well as timing of sampling. for example, FDA residues for citrus are higher than for PDP, because FDA includes the skin. PDP residues are significantly higher for pesticides that are applied during long term storage (root crops for example)

ii. Post harvest treatments account for exposure far in excess of the pounds applied relative to other crops. The majority of post harvest applications are used to treat fungal diseases on tree fruits and vegetables. Insecticides are used post harvest for grain storage. Growth regulators are applied to stored root crops (potatoes) to prevent sprouting.

c. Chemicals not included in either PDP or FDA account for 70% of agricultural pesticide active ingredient use, but much of the poundage is represented by herbicides and fumigants which would normally not be found.

i. The quantity of pesticide use, in lbs active ingredient, has little relation to dietary exposure

ii. Fungicides and Insecticides account for most of residues yet Herbicides have the highest use. Harvest aids and growth regulators also account for high residue levels but the number of pesticides in this category is small.

d. Correlation between PDP and FDA average residue per sample

i. The number of observations (or cases) is defined as pesticides which have PDP and FDA residue data for same crop and sample size exceeds 100 The 100 sample limit is the general rule of thumb used by residue chemistry.

ii. Intercept set to zero to estimate ratio of PDP to FDA. This reduces the loss of one degree of freedom for the intercept estimate as well as more

directly measures the ratio of residues- or multiple between the two.

iii. Key factors affecting estimated ratio of PDP to FDA residue

(a) Portion of product sampled (edible vs. total)

(b) Time of sample collection- including late post harvest applications that occur later in the retail distribution chain

(c) pesticide action, disposition of residues, and systemic activity which results in plant uptake of the pesticide.

Estimated Ratio of Residues between PDP and FDA

Fungicides Only

FUNGICIDES	Estimated ratio PDP/FDA	Signif level	R Square	Cases (obs)	Factors affecting multiple
APPLES	1.65	0.01	82%	6	2 post harvest pesticides- extreme pts
BANANA	0.04	0.02	100%	2	FDA includes peel; PDP does not
CELERY	0.07	0.25	57%	3	extreme points- little correlation
CARROT	17.32	0.27	53%	3	extreme points- little correlation
GREEN BEANS	0.15	0.02	95%	3	
GRAPES	0.32	0.16	43%	5	
LETTUCE	0 14	0.14	95%	2	
ORANGES	0.04	0.01	100%	2	FDA includes skin, PDP pulp only
PEACHES	2.70	0 03	71%	5	post harvest pesticide use
POTATOES	2.36	0.00	100%	3	post harvest pesticide use

Insecticides Only

INSECTICIDES	ratio	Signifi	R Square	Cases	Possible explanations
APPLES	4.45	0.00	64%	19	to be determined
BROCCOLI	0.28	0.02	69%	6	to be determined
CELERY	1.83	0.03	75%	5	to be determined
CARROT	0.53	0.06	55%	6	to be determined
GREEN BEANS	3.12	0.01	65%	8	to be determined
GRAPEFRUIT	0.02	0.17	68%	3	portion of fruit sampled
GRAPES	2.20	0.04	45%	9	to be determined

INSECTICIDES	ratio	Signifi	R Square	Cases	Possible explanations
LETTUCE	0.77	0.00	97%	2	to be determined
ORANGES	0.03	0.00	95%	10	portion of fruit sampled
PEACHES	1.23	0.00	89%	14	to be determined
POTATOES	1.87	0.02	54%	8	to be determined
SPINACH	6.43	0.00	100%	9	to be determined
WHEAT	0.48	0.00	94%	5	to be determined

High residue outliers
Fungicides

Crop	Both PDP and FDA	PDP only	FDA only
Apples	Thiabendazole, Diphenylamine		
Banana	Thiabendazole		
Celery	CHLOROTHALONIL	DICLORAN	
Grapes	Captan	Iprodione and Vincozolin	
Green beans	Chlorthalonil		
Lettuce	Iprodione		
Oranges	Thiabendazole		
Potatoes	Thiabendazole		
Peaches	Iprodione and Dicloran		Captan
Carrots	Iprodione		Pentachlorbiphenyl phenol PCB

Insecticides

Crop	Both PDP and FDA	PDP only	FDA Only
Apples	Propargite		Azinphos methyl and carbaryl

Crop	Both PDP and FDA	PDP only	FDA Only
Oranges	Carbaryl		methidathion and chlorpyrifos
wheat	malathion and chlorpyrifos		
spinach	permethrin		
Potatoes	DDT		Carbofuran
Peaches	Carbaryl Phosmet and Parathion	Azinphos methyl	
Lettuce	Permethrin		
Grapes		Dimethoate, omethoate	Parathion
Grapefruit	Ethion		Dicofol
Green beans		Acephate	Endosulfan
Carrots		Diazinon	DDT
Celery	Acephate and permethrin		
Broccoli	Permethrin		Methamidophos

e. Correlation among crops within PDP and within FDA (correlation matrix is in appendices)

i. Multivariate clustering remains to be done but based on a visual examination of the correlation matrix, the following crops have high correlations and appear to cluster.

(1) apples, grapefruit, oranges, bananas, broccoli
(2) peaches carrots grapes
(3) lettuce spinach
(4) potatoes oranges

(5) Crops that do not correlate with any other crop
   (a) Celery
   (b) wheat
   (c) sweet corn
   (d) processed peas

ii. Crops within FDA based on 20 crops examined

(1) Crops that appear to cluster

   (a) Tomatoes apples string beans peas cantaloupe sweet pepper hot peppers carrots

   (b) apple pear grapes potato orange cantaloupe
   (c) peach cherry

(2) Crops that do not cluster

   (a) Catfish
   (b) wheat
   (c) strawberries

7. Suggestions to improve existing programs to estimate national dietary exposure

   a. Decrease sample sizes for pesticide residues that can be predicted from historical data of residues and pesticide use

   b. Base sample sizes to reduce existing weighted estimation errors. Weight estimation error range by risk (amount/toxicity/endpoint of concern).

8. Future work

   a. Developing "synthetic estimates" for pesticides/crop combinations with limited or no data

      i. Model residue measurements as influenced by portion of the food sampled, time of sampling, decay rate of pesticide and metabolites, when applied, systemic pesticides which are taken up by the plant, and extent and changes in pesticide use

      ii. Identify cases for which estimates cannot be made or are statistically weak

iii. Evaluate the robustness of aggregate measures to identify significant changes or trends in the level of pesticide residues for a given set of chronic effects, i.e., cancer, neurotoxic, etc.

b. Additional sources toi nclude

i. Total diet study
ii. USDA's monitoring of meat milk and eggs
iii. state monitoring

c. Estimating sampling variance - individually and in aggregate for common mechanisms

d. Clustering and other multivariate techniques to identify plausible interrelationships of huge data sets

e. Developing relationships between pesticide use parameters, crop and pesticide chemical/physical properties to improve regualtion of pesticides

9. Appendice

a. Crops listed in order of estimated dietary pesticide consumption of children and women of child bearing age- FDA and PDP

b. Pesticides listed in order of estimated dietary pesticide consumption of children and women of child bearing age- FDA and PDP

c. Agricultural pesticides not included in PDP or FDA from 1992 to 1995:

Mathematical Geology

Volume 26, Number 3, April 1994

Contents

Mathematical Geology, Vol 26, No 3, 1994

Estimation of Background Levels of Contaminants

Anita Singh,[2] Ashok K. Singh,[3] and George Flatman[4]

Samples from hazardous waste site investigations frequently come from two or more statistical populations. Assessment of "background" levels of contaminants can be a significant problem. This problem is being investigated at the U.S. Environmental Protection Agency's Environmental Monitoring Systems Laboratory in Las Vegas. This paper describes a statistical approach for assessing background levels from a dataset. The elevated values that may be associated with a plume or contaminated area of the site are separated from lower values that are assumed to represent background levels. It would be desirable to separate the two populations either spatially by Kriging the data or chronologically by a time series analysis, provided an adequate number of samples were properly collected in space and/or time. Unfortunately, quite often the data are too few in number or too improperly designed to support either spatial or time series analysis. Regulations typically call for nothing more than the mean and standard deviation of the background distribution. This paper provides a robust probabilistic approach for gaining this information from poorly collected data that are not suitable for above-mentioned alternative approaches. We assume that the site has some areas unaffected by the industrial activity, and that a subset of the given sample is from this clean part of the site. We can think of this multivariate data set as coming from two or more populations: the background population and the contaminated populations (with varying degrees of contamination). Using robust M-estimators, we develop a procedure to classify the sample into component populations. We derive robust simultaneous confidence ellipsoids to establish background contamination levels. Some simulated as well as real examples from Superfund site investigations are included to illustrate these procedures. The method presented here is quite general and is suitable for many geological and biological applications.

KEY WORDS: robust M-estimators, influence function, background estimation, robust confidence limits, separation of mixed sample

INTRODUCTION

The United States Environmental Protection Agency (U.S. EPA) encounters the statistical problem of mixed samples from two or more populations in Resource Conservation and Reclamation Act (RCRA) and Superfund Amendment and

[1] Received 23 June 1993, accepted 5 November 1993.

[2] Lockheed Environmental Systems and Technologies Company, 980 Kelly Johnson Drive, Las Vegas, Nevada 89119.

[3] Department of Mathematics, University of Nevada, Las Vegas, Nevada 89154.

[4] United States Environmental Protection Agency, Las Vegas, Nevada 89154.

Reauthorization Act (SARA) Evaluation and Remediation. This problem is being considered at U.S. EPA's Environmental Monitoring and Systems Laboratory at Las Vegas (EMSL-LV). This paper presents a solution from a probability distribution-based method. A sample of concentration values of contaminants from a Superfund site can be thought of as a mixed sample of background concentration values plus the concentration values from a plume or plumes At first glance, a statistical analyst could think that the mixed sample from a Superfund site could be separated spatially by a Kriging analysis However, these statistical techniques need data obtained using appropriate statistical designs Unfortunately, regulatory life is not simple. Often only too few samples or improperly spaced data for spatial or time series analysis are available and the required regulatory information is only the mean and standard deviation of the distribution(s). This paper provides a robust probabilistic approach for gaining this information from data that are inadequate for above-mentioned alternative approaches.

The occurrence of mixture samples from two or more normal (lognormal) populations has been well recognized in several applied areas of interest such as biology, geology, medicine, reliability, and environmental science Several classical partitioning methods are available in statistical literature. Sinclair (1976) used normal probability plots for graphical partitioning of mixture samples in mineral exploration studies. Holgersson and Jorner (1978) gave a good review of various methods including graphical, maximum likelihood (MLE), nonlinear least squares, and method of moments. Fowlkes (1979) performed extensive simulations to compare several graphical methods including the usual histogram method, the normal probability Q-Q plot, and the empirical cumulative distribution function. The ability of these classical and graphical methods to identify mixtures in samples is doubtful, especially if discordant observations are also present in these samples. Moreover, the detection of these mixtures becomes extremely difficult in the presence of overlap among the component populations. Campbell (1984) used robust methods to study the effect of anomalies on mixture models. Recently Fleischauer and Korte (1990) used the point of inflection of the normal probability plot to obtain an estimate of threshold background level contamination.

The graphical display, unarguably, is one of the most powerful diagnostic tools in the hands of a researcher. However, a subjective estimate of the point of inflection obtained by looking at these graphs is questionable, especially when more than two component populations are present. The overlap among the component populations generally masks the point of inflection Moreover, the anomalous observations (if any) and the presence of several (unknown) component populations can distort the Q-Q plot to such an extent that the resulting inflection point estimates may not be reliable. If one wants to use the Q-Q plots as a partitioning method, a stepwise procedure is desirable The proposed stepwise

procedure requires construction of a Q–Q plot at each step. Populations with higher concentration levels will be identified first. Each step identifies a sample from a different population. In this article, we propose robust procedures to partition a given mixture sample into its component populations. Data-appraised robust confidence limits for the individual observations placed on the same Q–Q plot produce a more precise estimate of the cutoff point between two adjacent populations This reduces the subjectivity involved in choosing the inflection point from the graph. Several simulated as well as real-life examples have been discussed to illustrate these procedures. The mathematical formulation is given in the second section, the third section has all the examples, and finally, there is a summary of our conclusions and recommendations.

MATHEMATICAL FORMULATION

The density function $f_M(x)$ of a mixture population with $(g + 1)$ unknown component populations is given by

$$f_M(x) = \sum_{i=0}^{g} p_i f_i(x; \mu_i; \sigma_i) \tag{1}$$

where $g \geq 1$, and $f_i(x, \mu_i, \sigma_i)$ is the density function of the ith population Π_i, assumed to be normally (or lognormally) distributed with unknown mean and standard deviation (SD) μ_i and σ_i respectively, and p_i is the unknown mixture proportion for Π_i; $i = 0, 1, 2, \ldots, g$, with $\Sigma p_i = 1$. Throughout the rest of the article, it has been assumed that the researcher has performed a suitable data transformation to achieve normality or near-normality (e g., log-transformation for positively skewed data) before proceeding with the following algorithm. Given a sample $x_1, x_2, \ldots, x_n$ of size n from this mixture model, the objective is to resolve it into its component populations, i.e., find $n_i \geq 0$ such that n_i observations belong to Π_i with $\Sigma_{i=0}^{g} n_i + n_E = n$. Here $n_E \geq 0$ is the number of extreme unusual observations which stand alone and do not belong to any of the given $(g + 1)$ populations. The subsample of size n_i then can be used to estimate the parameters of population Π_i and its proportion p_i, $i = 0, 1, \ldots, g$. The normal probability Q–Q plot is generally used to get an idea about g, the number of populations present. However, inevitable overlap among the component populations and/or the presence of anomalous observations generally distort the Q–Q plot significantly, resulting in masking of some of the component populations, especially those populations which have lower concentration levels Traditionally, theoretical quantiles from a standard normal distribution are plotted along the x-axis in a typical Q–Q plot. However, in this article, we use the theoretical quantiles from $N(\bar{x}, s)$ for the classical Q–Q plot and the theoretical quantiles from $N(\bar{x}^*, s^*)$ for the robust Q–Q plot, here $\bar{x}$ is the sample mean

s the sample standard deviation, and $\overline{x^*}$, s^*, (defined later in this paper), represent their robust versions, respectively.

The initial step in the process is to identify $n_E \geq 0$ highly contaminated observations, which stand alone by themselves on a normal probability plot. These observations may require individual treatment and/or further investigation and should not be included in the subsequent partitioning of the underlying mixture sample. Due to masking effects, the exclusion of these observations from subsequent analysis may be required to identify intermediate populations. This does not mean at all that these observations have been thrown away. The new Q–Q plot will be drawn using the remaining $n - n_E$ observations. This Q–Q plot will reveal if any representative samples from populations with higher concentrations, namely Π_g, Π_{g-1} etc. are present. Robust confidence limits for the individual observation x_i drawn on these Q–Q plots provide an objective (rather than subjective) estimate of the cut-off point between two adjacent populations. The process is repeated until all of the observations have been classified into the various component populations. Each time a population is identified, a new Q–Q plot with the new robust limits is drawn using only the unclassified observations. This process provides a good estimate of the number of remaining populations that need to be identified. At each step, these robust limits correspond to the most dominant population present at that step. If there is such a population present, then this population may be identified first, using these robust limits as the estimates of its cutoff points from the adjacent populations. The separation between two populations is probably most difficult in the presence of overlap. The overlapping populations (if any) should be identified in the very end. All these ideas have been explained by means of several examples presented in the following section.

Here, Π_0 represents the background population and Π_i; $i = 1, 2, \ldots, g$ represents contaminated parts of the site with varying degrees of contamination levels in ascending order of magnitude, with Π_g representing the population with highest contamination levels. A recently proposed redescending PROP (Singh, 1994) influence function used here to identify the discordant observations is given by

$$\begin{aligned} \psi(d) &= d && \text{if } d \leq d_0 \\ &= d_0 \exp(-(d - d_0)) && \text{if } d > d_0 \end{aligned} \qquad (2)$$

where d_0^2 is the (α) 100% critical value of the distribution of $d_i^2 = (x_i - \bar{x})^2/s^2$ which is distributed as an $(n-1)^2\beta(1/2, (n-2)/2)/n$, where n here represents the number of observations used in the computation of $\bar{x}$ and s.

It should be noticed that the number of observations used will be updated each time the process is repeated. For the initial iteration all of the n observations will be used, next $n - n_1$ will be used, and then the remaining $n - n_1$ of

observations classified into Π_g will be used, and so on. Each observation is assigned some weight according to its extremeness in either of the two tails of the distribution. These weights provide a very effective way of obtaining estimates of the degrees of freedom needed to compute the individual robust confidence limits at each step. The resulting M-estimators for a given sample are:

$$\overline{x^*} = \Sigma\omega_1(d_i)x_i/\Sigma\omega_1(d_i),$$

and

$$s^{*2} = \Sigma\omega_2(d_i)(x_i - \overline{x^*})^2/\nu \tag{3}$$

$$\omega_1(d_i) = \psi(d_i)/d_i,\ \omega_2(d_i) = [\omega_1(d_i)]^2$$

$\nu = \Sigma\omega_2(d_i) - 1$. The robustified distances $d_i^{*2} = (x_i - \overline{x^*})^2/s^{*2}$ follow a $\nu^2\beta(1/2, (\nu - 1)/2)/(\nu + 1)$ distribution. The two-sided robust limits for the individual observation x_i are given by the following probability statement:

$$P(LTL \le x_i \le UTL) = 1 - \alpha, \qquad i = 1, 2, \ldots, n \tag{4}$$

where $LTL = \overline{x^*} - s^*d_\alpha^*$ and $UTL = \overline{x^*} + s^*d_\alpha^*$, $\overline{x^*}$ and s^* are given by (3), and d_α^{*2} is the (α) 100% critical value from the distribution of d_i^{*2}. The one-sided $(1 - \alpha)$ 100% robust limit for individual x_i can be obtained similarly. The index i runs over the number of observations used in a typical step. Once the n_E extreme observations have been identified and removed from the data set, new Q–Q plot using the rest of the $n - n_E$ observations is drawn. It should be emphasized that the limits used here are for the individual observations x_i and not for the population mean μ, as is sometimes done in practice. For example, in the context of background level estimation, individual observations are being compared (and not the population mean μ) to these threshold limits. Therefore, these limits should be computed using the appropriate interval. A brief description of the various intervals and limits is given in Singh and Nocerino (1993).

The robust limits given by (4) when drawn on the same probability plot provide a good initial estimate of the cutoff point between the adjacent populations. An estimate of the cutoff point c_g between populations Π_g and Π_{g-1} will be obtained first from this Q–Q plot. All of the unclassified observations $x \ge c_g$ (not including the n_E extreme observations) will be used to obtain the robust interval $I_g = (LTL_g, UTL_g)$ for the gth population Π_g. All of the unclassified observations belonging to this interval will be declared as coming from Π_g. Next, all the observations $x > LTL_g$ will be deleted from the subsequent partitioning and a new Q–Q plot with the new robust limits will be obtained using the remaining observations. An estimate of C_{g-1}, the cutoff point between populations Π_{g-2} and Π_{g-1}, will be obtained from this plot. All unclassified observations $x \ge c_{g-1}$ will be used to obtain the robust boundaries given by

$I_{g-1} = (LTL_{g-1}, UTL_{g-1})$ for the $(g-1)$th population Π_{g-1}. All observations belonging to I_{g-1} will be declared as coming from Π_{g-1}. In case of any overlap between Π_{g-1} and Π_g, i.e., when $LTL_g \leq UTL_{g-1}$, observations in the range (LTL_g, UTL_{g-1}) can be assigned to either of the two populations Π_g or Π_{g-1}. However, the PROP influence function (2) used in the derivation of the robust limits given by (4) minimizes the overlap between the estimates for the two adjacent populations by down-weighing the extreme observations appropriately in either of the two tails of the distribution of the underlying populations. Moreover, when the two adjacent populations have disjoint boundaries, the observations (if any) belonging to this unclaimed region (LTL_i, UTL_{i+1}) should be assigned to their nearest neighbor.

This process will be repeated as many times as required until all of the observations have been classified into their respective populations. At the final step, the threshold values for the background population Π_0 will be estimated. The remaining unclassified observations will be used to estimate UTL_0, which is given by the one-sided probability statement:

$$P(x_i < UTL_0) = 1 - \alpha$$

where UTL_0 can be obtained using (4) by replacing α with $2 * \alpha$

Observations smaller than UTL_0 will be declared as coming from Π_0. As before, if there is overlap between Π_0 and Π_1, i.e., $LTL_1 \leq UTL_0$, then observations in the overlapping range (LTL_1, UTL_0) can be assigned to either of the two populations Π_0 or Π_1. Once the boundaries for the various component populations have been established, the complete classification procedure can now be described in various steps as follows:

1. First of all, identify all of the extreme observations $n_E \geq 0$ These will not be used in any of the subsequent partitioning of the underlying sample.

2. Next define a_i = no. of observations ϵ the overlapping region (LTL_i, UTL_{i-1}) between populations Π_{i-1} and Π_i, with $a_{i-1,i} \geq 0$ of these $\epsilon\Pi_{i-1}$ and $a_{i,i} \geq 0$ of these $\epsilon\Pi_i$, $i = 1, 2, \ldots, g$. and b_i = no of observations ϵ the unclaimed region (UTL_{i-1}, LTL_i) between populations Π_{i-1} and Π_i, with $b_{i,i} \geq 0$ of these $\epsilon\Pi_i$ and $b_{i-1,i} \geq 0$ of these $\epsilon\Pi_{i-1}$, $i = 1, 2, \ldots, g$

3. Identify all of the non-overlapping observations $\epsilon\Pi_0$ Then n_0 = (no. of non-overlapping observations $\leq UTL_0$) + $a_{0,1} + b_{0,1}$

4. In general, the number of observations $\epsilon\Pi_i$ is given by n_i = (no. of non-overlapping observations ϵI_i) + $a_{i,i} + a_{i,i-1} + b_{i,i} + b_{i,i-1}$, where $i = 1, 2, \ldots, g-1$

5. Identify all of the non-overlapping observations $\epsilon\Pi_g$ Then n_g = (no of non-overlapping observations $\geq LTL_g$) + $a_{g,g} + b_{g,g}$

6 Once, the number $(g+1)$ of populations present, and the respective

subsample sizes n_i, $i = 0, 1, \ldots, g$ have been estimated, the $(g + 1)$ population proportions are estimated using the following formula:

$$p_i = n_i/(n - n_E), \qquad i = 0, 1, \ldots, g$$

7. Finally, using these n_i observations, the robust estimates of the parameters of population Π_i, $i = 0, 1, \ldots, g$ will be obtained using (3).

In order to illustrate the proposed statistical procedure, we now present some simulated as well as real examples.

EXAMPLES AND DISCUSSION

The procedure described here has been applied to two simulated datasets as well as a real dataset from the Sacramento Army Depot Superfund Site from Region 9 EPA. There were six primary contaminants at the Sacramento Army Depot Superfund Site: Cadmium (Cd), Chromium (Cr), Copper (Cu), Lead (Pb), Nickel (Ni); and Zinc (Zn). A total of 45 samples were analyzed for the above contaminants, six from uncontaminated regions of the site; which will be referred to as the site-specific background sample; and 39 from contaminated regions of the site. Moreover, the procedure outlined here has been used on a simulated data set representing a sample from a mixture of two lognormal populations. Three simulated data sets and the Sacramento Army Depot Superfund Site data set are given in the Appendix. In the following, all letters with * as a superscript represent robust estimates, else, they are the classical maximum likelihood estimates (MLEs). All the computations have been done using the statistical software package SCOUT developed by the Lockheed Environmental Systems & Technologies Company (LESAT) for the U.S EPA.

Example 1. A mixture sample of size 100 was generated from two reasonably separated normal populations with 90% ($p_1 = 0.9$) observations coming from a normal population Π_0 with mean 10 and SD 3 $\sim N(10, 3)$ and 10% ($p_2 = 0.1$) observations coming from $\Pi_1 \sim N(27, 8)$. Observations for the first sample ranged from 2.485 to 18.598, whereas observations for the second sample ranged from 9.489 to 43.998, indicating some overlap between the two populations. This is the data set no 1, given in the Appendix The normal probability Q–Q plots for the whole data set with the classical and the robust limits placed on them are given in Figs. 1a and b, respectively. From both graphs, it is obvious that there are two populations present. The upper robust limit 15.99 for the dominating population Π_0 provides an estimate of the cutoff point c_1 between the two populations (Fig. 1b). Next, using all observations $\geq c_1$, the 95% robust one-sided lower boundary for the population Π_1 with higher concentrations is given by $LTL_1 = 17.5$ (Fig. 1c) Therefore, all of the observations greater than LTL_1 are classified as coming from Π_1 Using the remaining

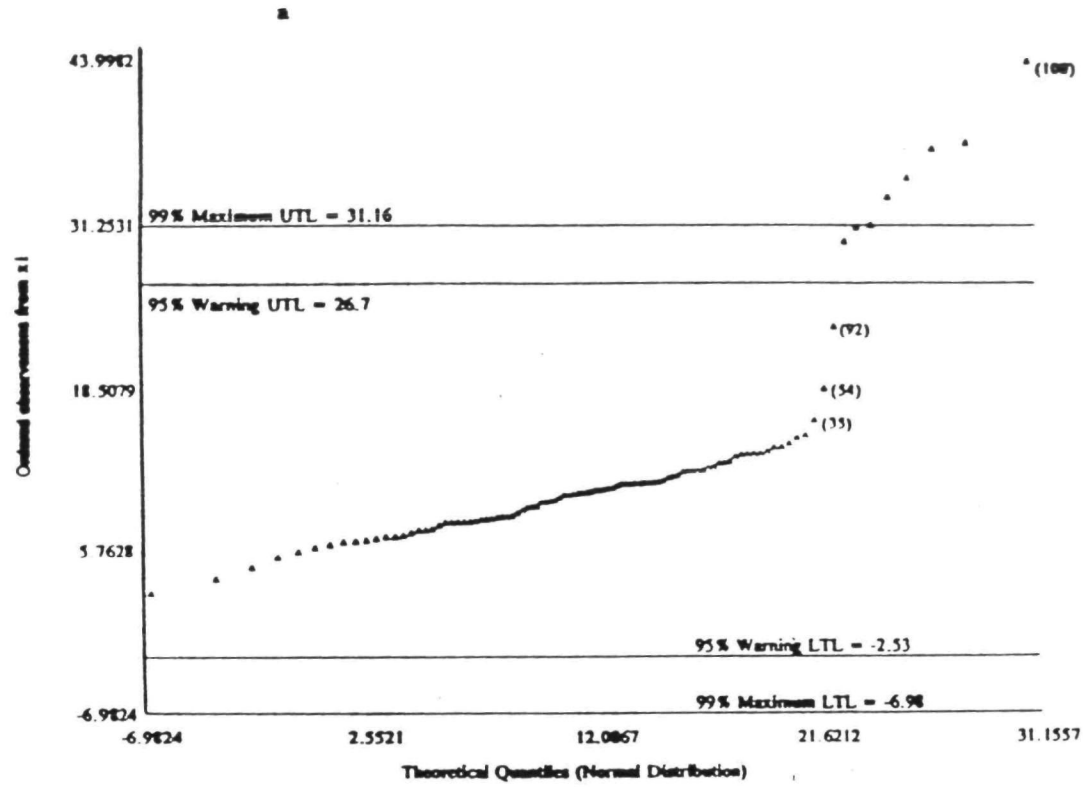


Fig. 1a. Mixture of $N(10, 3)$ and $N(28, 7)$-classical Q–Q plot.

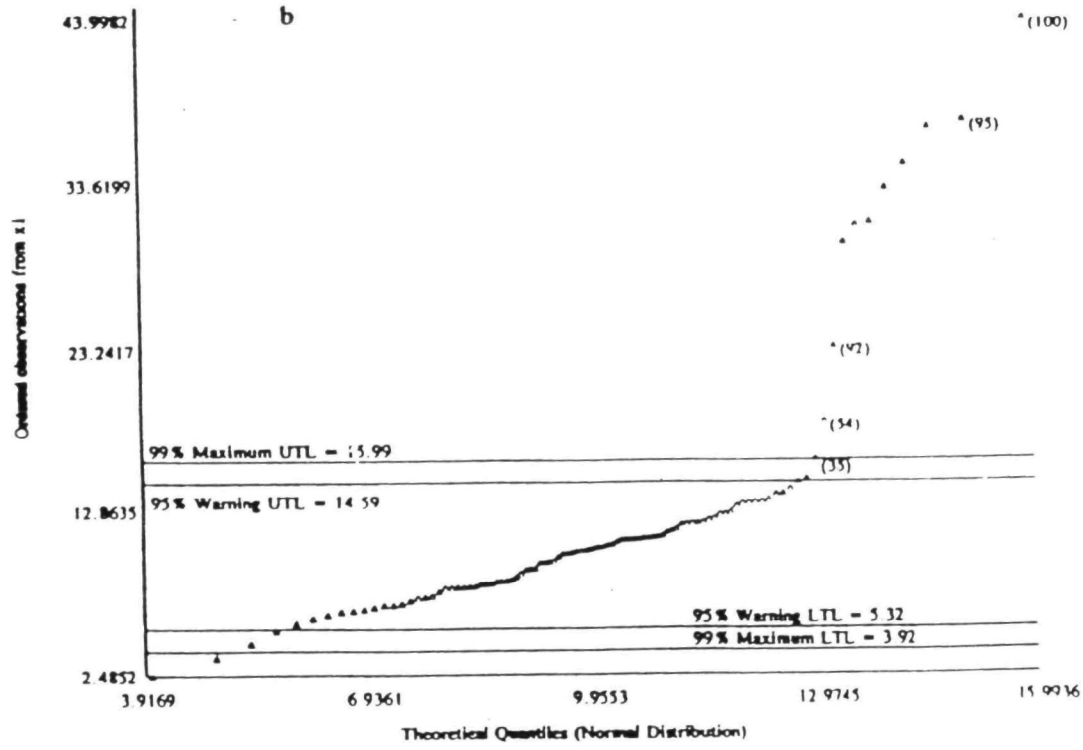


Fig. 1b. Mixture of $N(10, 3)$ and $N(27, 8)$-robust Q–Q plot.

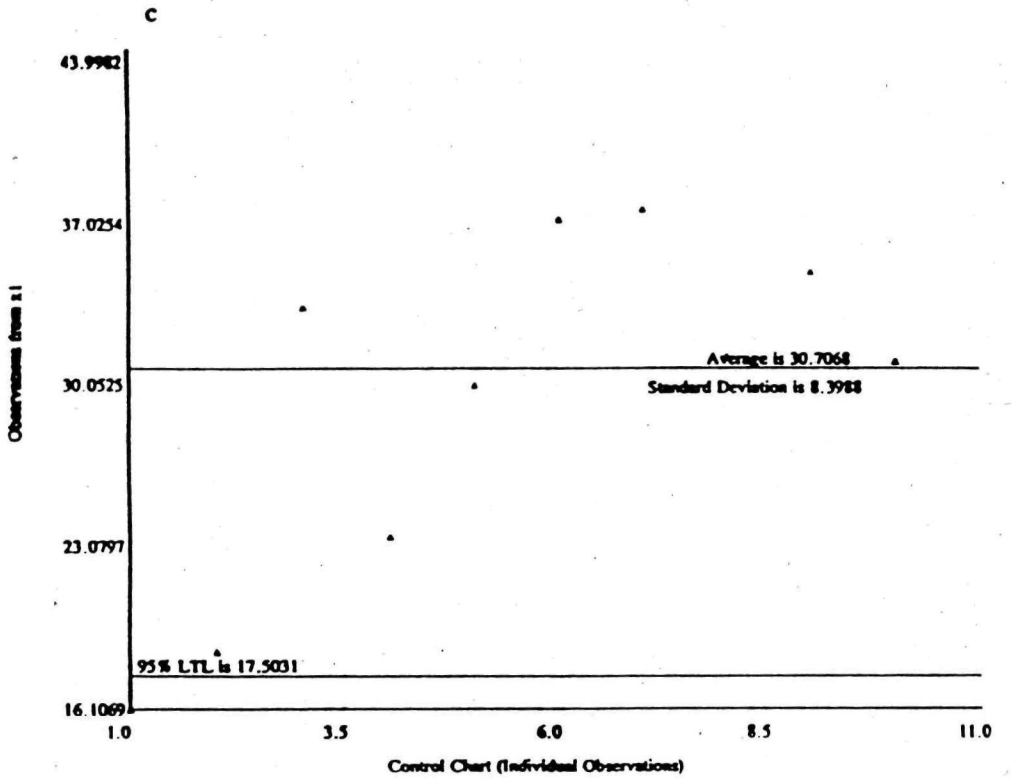


Fig. 1c. Mixture of $N(10, 3)$ and $N(27, 8)$-robust chart-$N(27, 8)$ contam. sample.

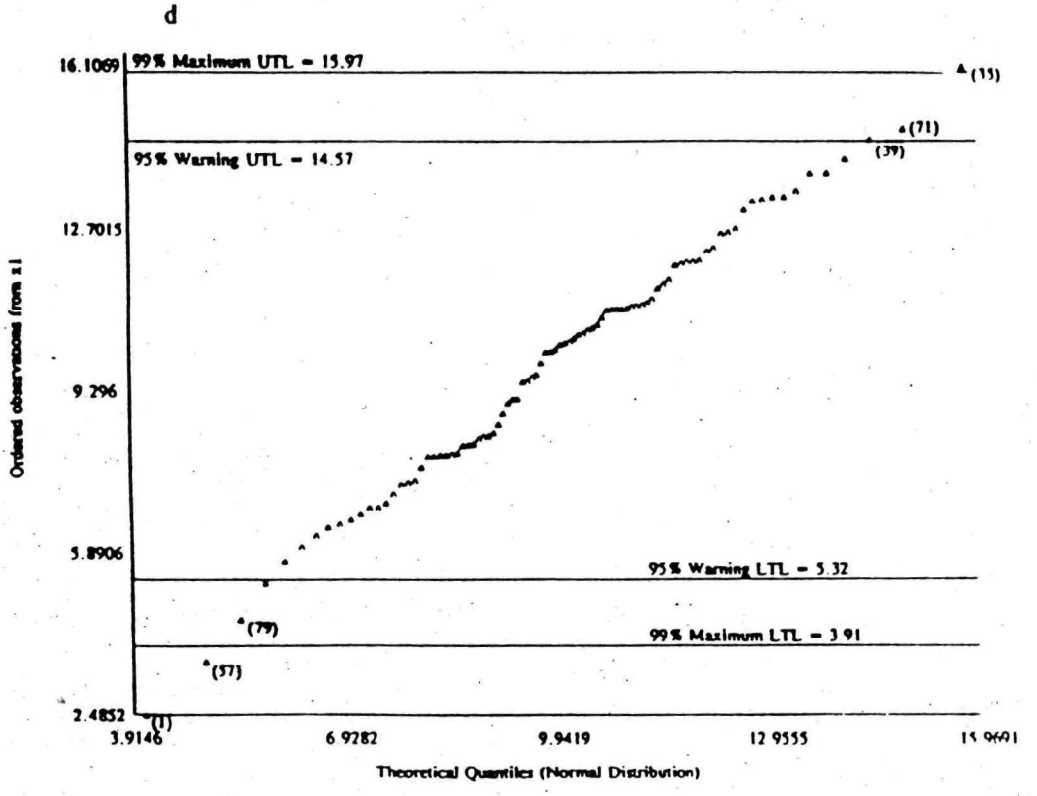


Fig. 1d. Mixture of $N(10, 3)$ and $N(27, 8)$-robust Q-Q plot unclassified sample.

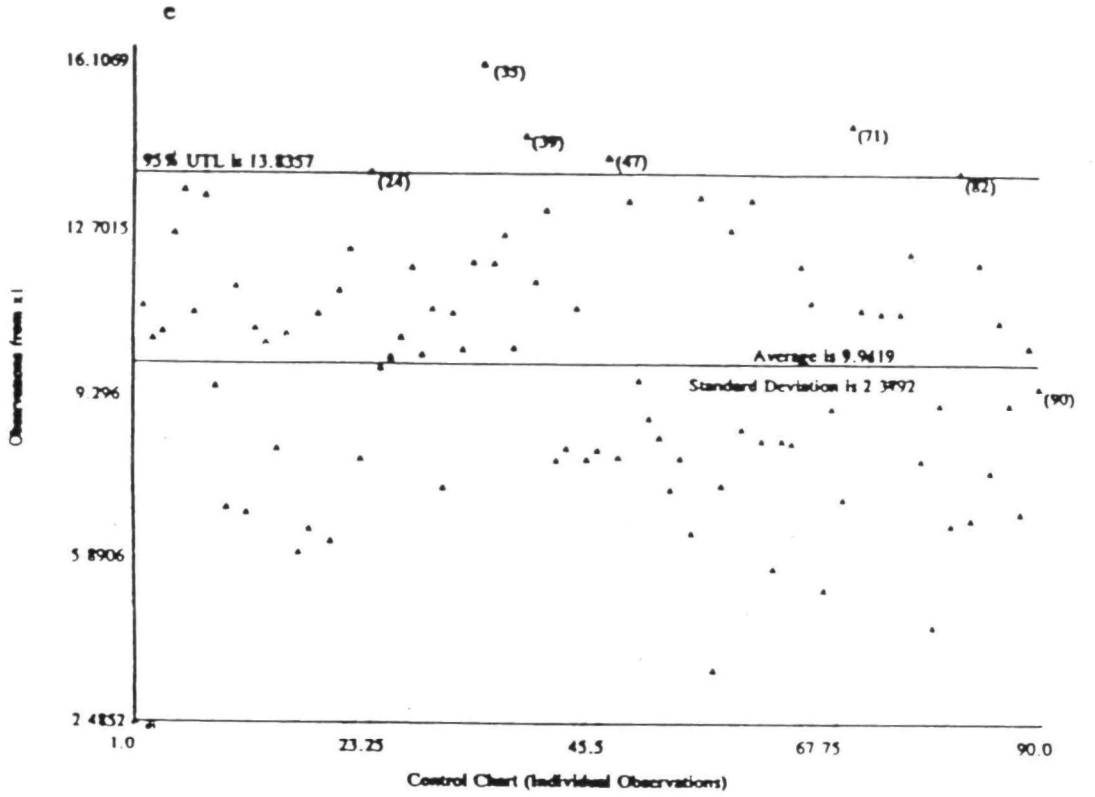


Fig. 1e. Mixture of $N(10, 3)$ and $N(27, 8)$-robust chart background $N(10, 3)$ sample.

unclassified observations (smaller than 17.5), the Q–Q plot with the robust limits placed on it is shown in Fig. 1d. From this figure, it is obvious that there is only one population left. The 95% one-sided robust upper boundary $UTL_0 = 13.84$ for Π_0 is given in Fig. 1e. All observations less than 13.84 are classified as coming from Π_0. Observations in the range (UTL_0, LTL_1) will be assigned to their nearest neighbor. Thus the observation 16.107 (the only observation in this range with $b_1 = 1$) will be assigned to Π_1. Two observations from Π_0, namely, 16.107 and 18.598 are misclassified into Π_1, and one observation 9.489 of Π_1 has been misclassified into Π_0. All of the relevant estimates of the population parameters after the final classification are summarized in Table I.

Example 2. In this simulated example, we consider a three population mixture model with ten observations from an $N(20, 4)$ population, 100 from an $N(0,1)$ population, and 30 from an $N(5, 1)$. Moreover, in order to show the extent of distortion of the Q–Q plot by the presence of extreme observations, two extreme observations from an $N(100,10)$ are also included in this mixed sample. This is data set no. 2 in the Appendix. The classical and the robust Q–Q plots using all of the 142 observations are given in Figs. 2a and b, respectively. Both graphs identify the two extreme observations. Moreover, both graphs give indications of the presence of a sample from a population with higher concentrations (observations no. 1–10). However, due to the large vari-

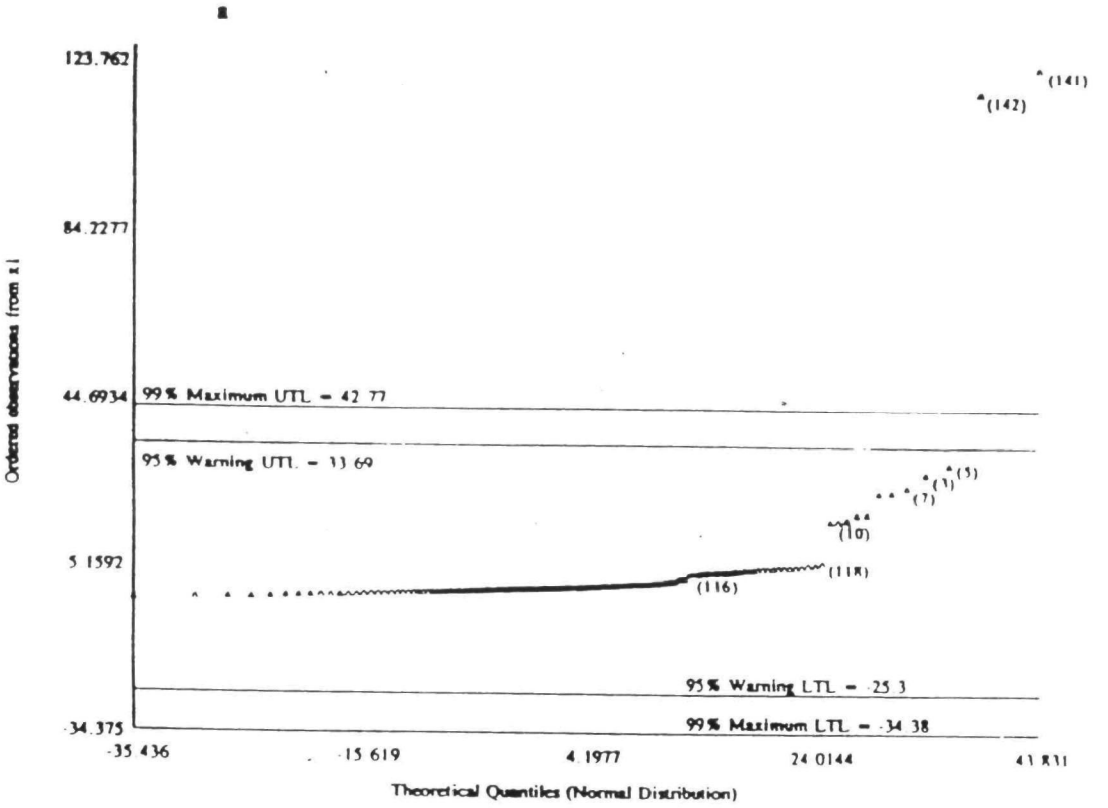


Fig. 2a. Mixture of $N(0, 1)$, $N(5, 1)$, $N(20, 4)$, $N(100, 10)$-classical Q-Q plot.

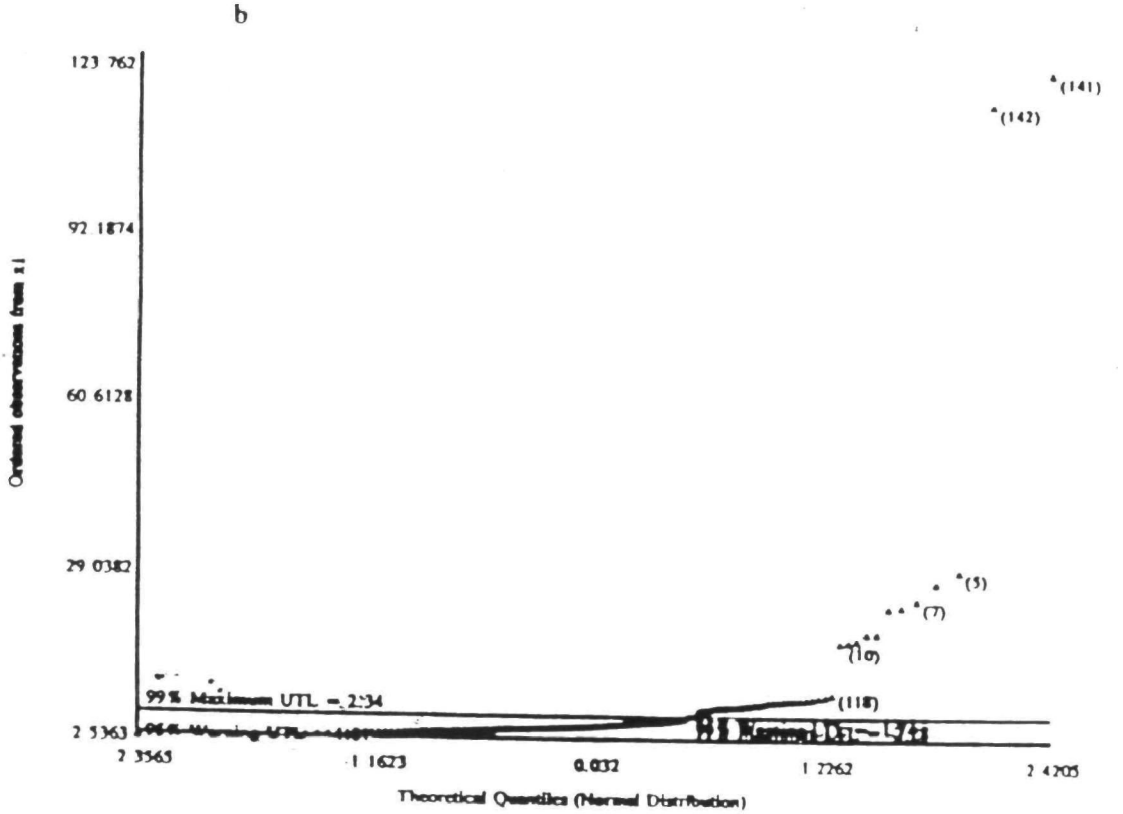


Fig. 2b. Mixture of $N(0, 1)$, $N(5, 1)$, $N(20, 4)$, $N(100, 10)$-robust Q-Q plot.

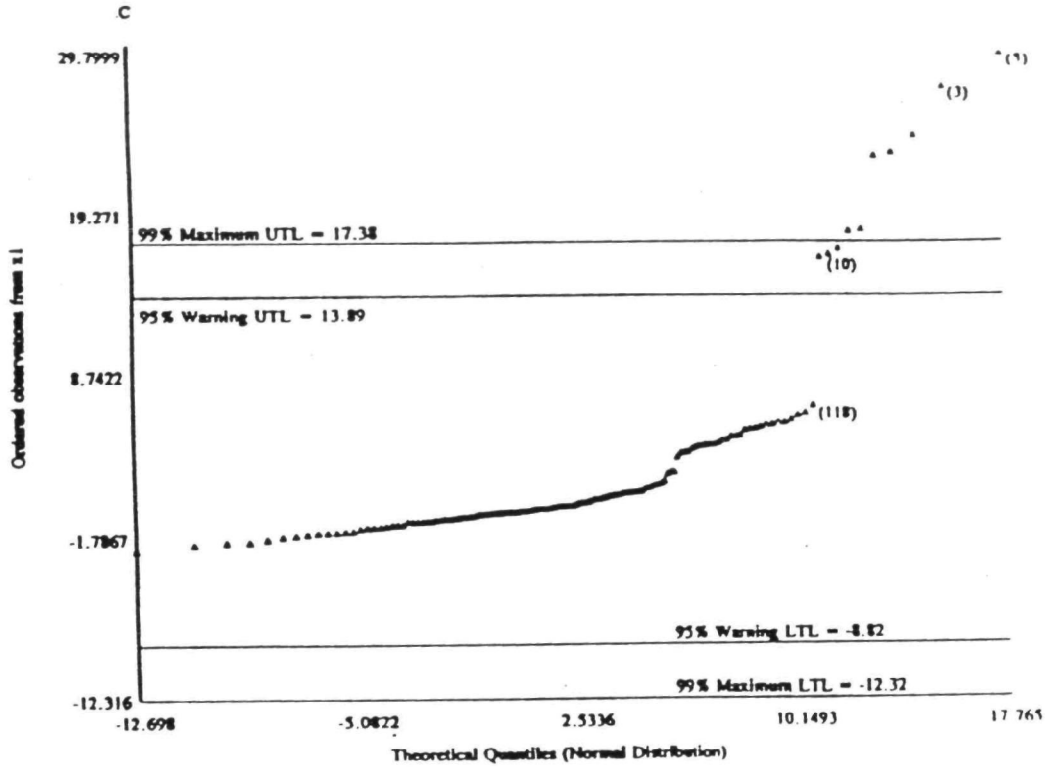


Fig. 2c. Mixture of $N(0, 1)$, $N(5, 1)$, $N(20, 4)$-classical Q–Q plot/two extremes removed.

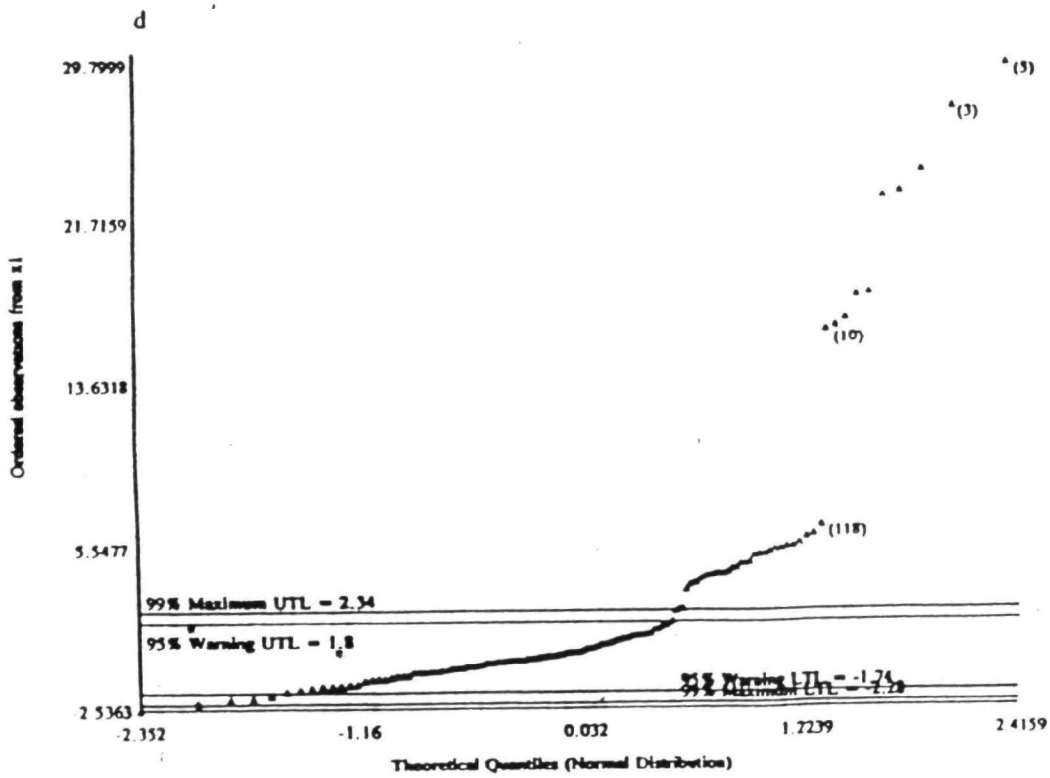


Fig. 2d. Mixture of $N(0, 1)$, $N(5, 1)$, $N(20, 4)$-robust Q–Q plot/two extremes removed.

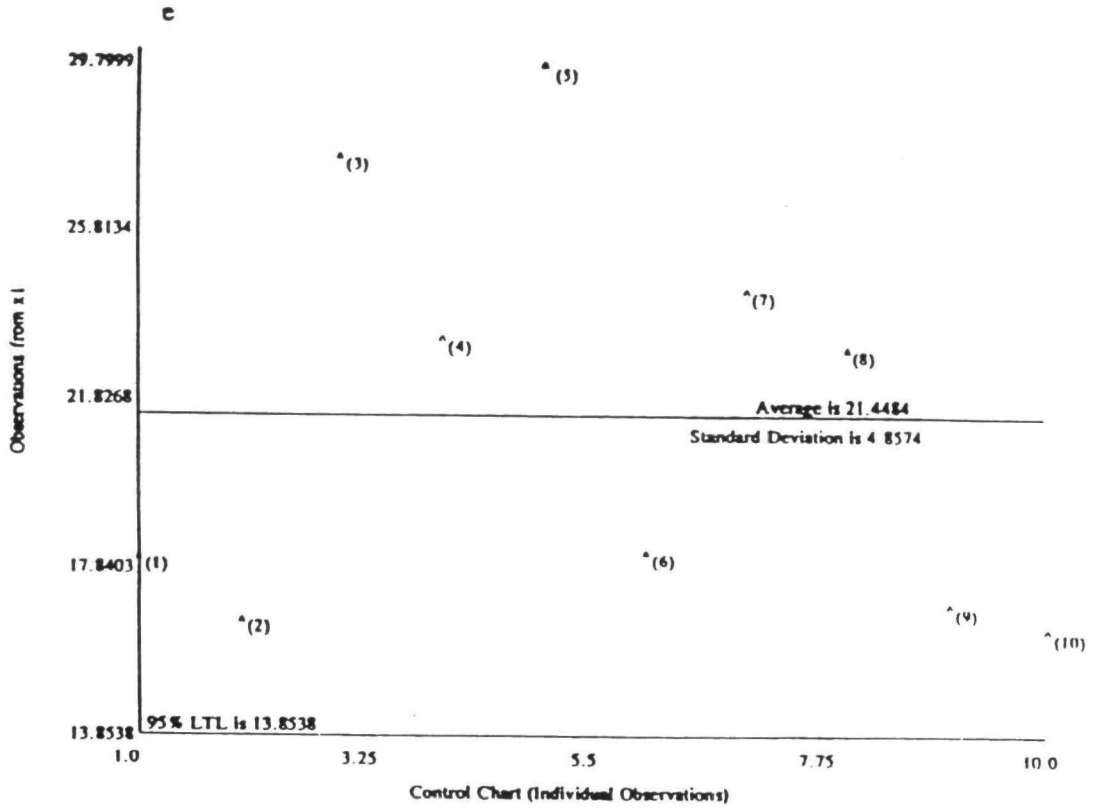


Fig. 2e. Mixture of $N(0, 1)$, $N(5, 1)$, $N(20, 4)$-robust chart-contam. sample $N(20, 4)$.

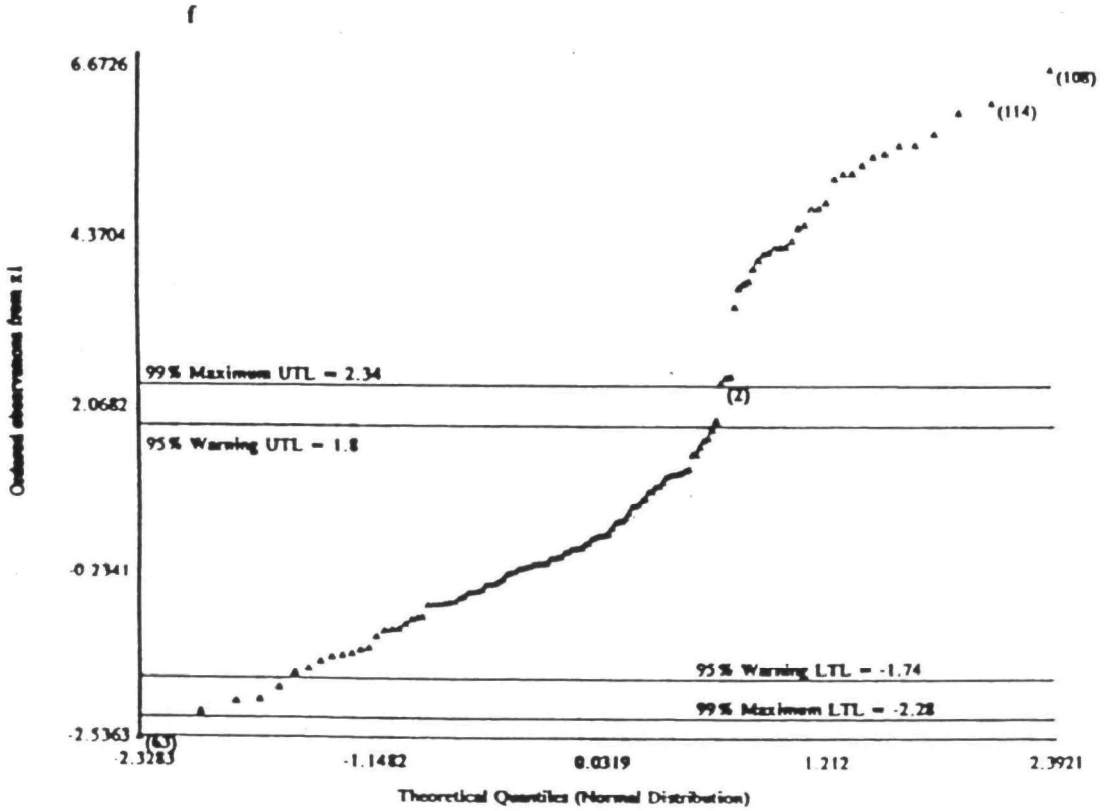


Fig. 2f. Mixture of $N(0, 1)$, $N(5, 1)$, $N(20, 4)$-robust Q–Q plot unclassified data.

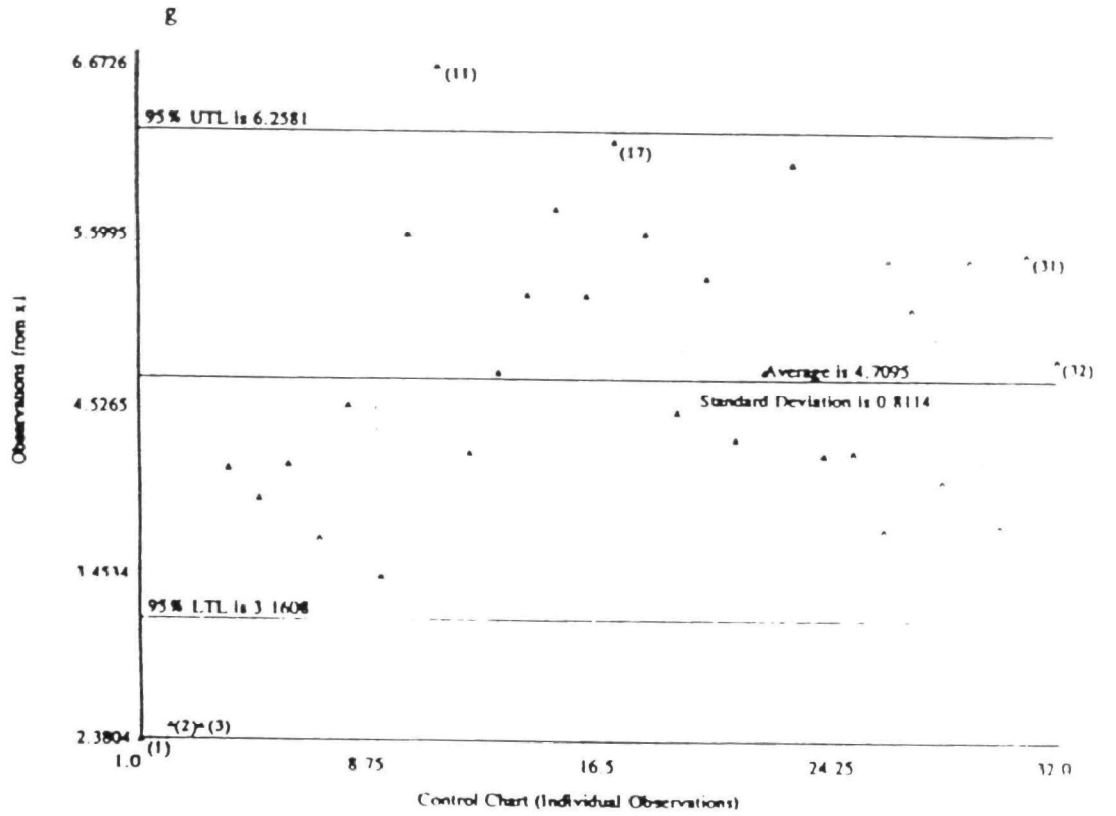


Fig. 2g. Mixture of $N(0, 1)$, $N(5, 1)$, $N(20, 4)$-robust chart-intermed. sample $N(5, 1)$.

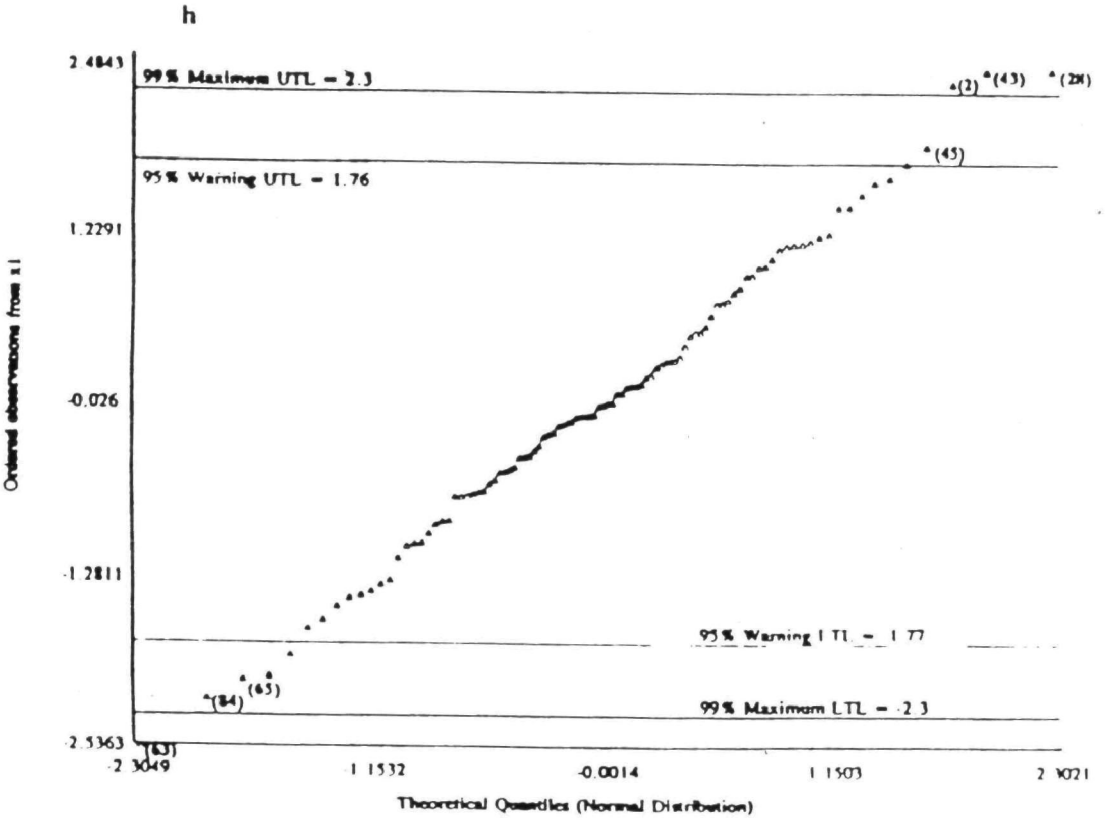


Fig. 2h. Mixture of $N(0, 1)$, $N(5, 1)$, $N(20, 4)$-robust Q–Q plot unclassified data.

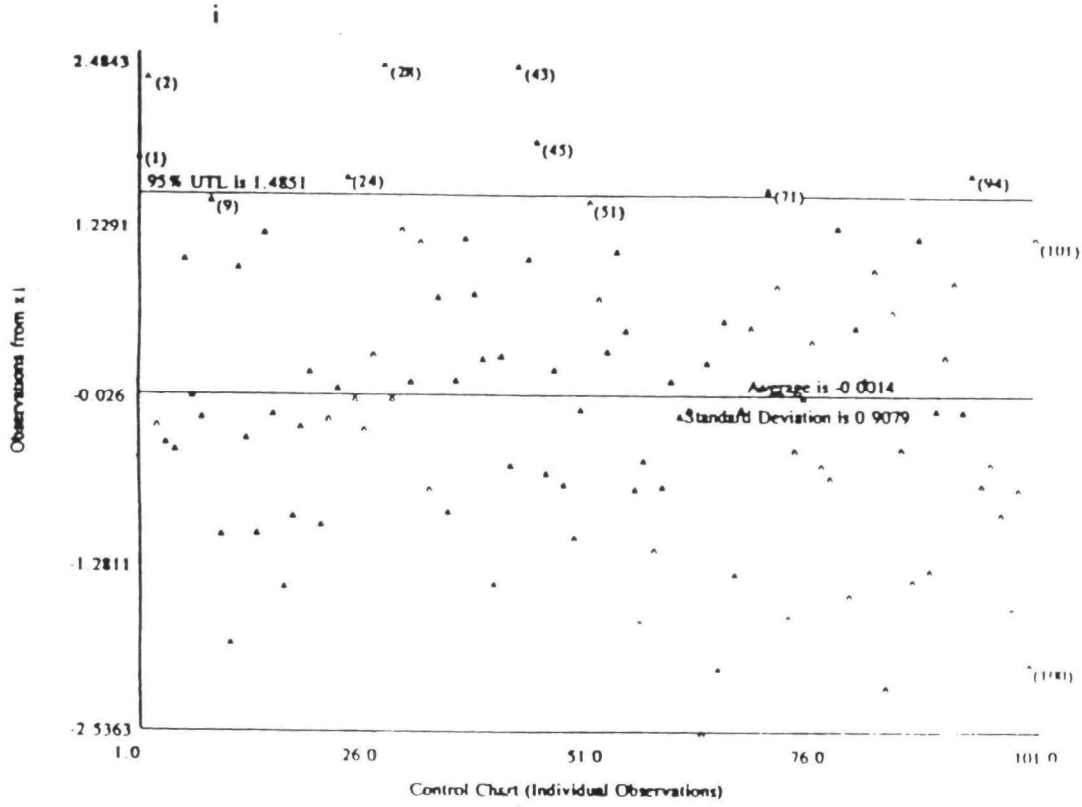


Fig. 2i. Mixture of $N(0, 1)$, $N(5, 1)$, $N(20, 4)$-robust chart-background $N(0, 1)$.

Table I

Popn.	95% Limits	a_i, b_i	n_i	p_i	$\overline{x^*}$	s^*	$\bar{x}$	s
Π_0	$UTL_0 = 13.84$	$b_{0\,1} = 5$	89	.89	9.91	2.36	9.78	2.59
Π_1	$LTL_1 = 17.50$	$b_{1\,1} = 1$	11	.11	30.71	8.40	30.71	8.40

ation in the data set, the intermediate population is masked in Fig. 2a, whereas, Fig. 2b gives a clear indication of the presence of at least three populations. Figures 2c and d represent the same graphs after removal of the two $n_E = 2$ extreme observations. From Fig. 2c, one can wrongly conclude that there are two populations present with observation no. 118 = 6.67 as the inflection point. However, this is not the case here, as is obvious from Fig. 2d. Using observation no. 10 as the cutoff $c_2 = 16.41$ between populations Π_1 and Π_2, the classical as well as the robust (same) lower boundary for population Π_2 is given in Fig. 2e. All observations greater than $LTL_2 = 13.85$ will be classified into Π_2. A new Q–Q plot using the remaining unclassified observations is given in Fig. 2f. which leads to $c_1 = 2.34$ as the cutoff point between populations Π_1 and Π_0. It

should be noticed that the robust procedure used here has produced the same cutoff point of 2.34 between populations Π_0 and Π_1, as can be seen from Fig. 2b, d, and f. Using all unclassified observations ≥ 2.34, the two-sided 95% robust boundary for population Π_1 is (3.16, 6.26), as given in Fig. 2g. Next all observations less than 3.16 have been used to draw the robust Q–Q plot, given in Fig. 2h. From this graph it is obvious that there is only one population Π_0 left at this stage. Using these observations, the 95% robust upper boundary for the background population is given by $UTL_0 = 1.485$. All observations less than this threshold will be classified into the background population Π_0. Once the boundaries have been set, observations in the overlapping and the unclaimed regions have been classified according to rules described above. All of the relevant statistics using the final classification are summarized in Table II.

Example 3. In this example, we consider the data set from a Superfund Site with six samples known to come from the background population (observations 33–38). As mentioned earlier, the site was sampled for six contaminants, but the results for cadmium concentrations alone are included in this article. The data for the 45 collected samples (background samples included) is given in data set no. 3 in the Appendix.

The average site-specific background level of a contaminant plays an important role in remediation decisions. As such, the estimation of the average site-specific background of a contaminant is an important problem. We now show the results obtained by using the proposed procedure on Cadmium concentrations. The classical as well as the robust Q–Q plots for cadmium are given in Figs. 3a and b, respectively. From these figures, it is obvious that observation nos. 15, 9, 22, and 21 represent extremely contaminated samples and should be treated individually. From Fig. 3b, there is a clear indication of the presence of at least three populations. Figure 3c represents the robust Q–Q plot after removal of these $n_E = 4$ extremes, which also indicates the presence of at least three populations. Using $c_3 = 260.27$ (observation no. 19 after the removal of extremes) as the cutoff point between populations Π_2 and Π_3, all observations greater than c_3 will be used to estimate the parameters of Π_3, the population with high concentrations. Figure 3d indicates that these observations are from

Table II

Π_i	95% Limits	a_i, b_i	n_i	p_i	$\bar{x}^*$	s^*	$\bar{x}$	s
Π_0	$UTL_0 = 1.48$	$b_{0,1} = 5$	98	7	−0.03	0.89	−0.07	0.97
Π_1	$LTL_1 = 3.16$	$b_{1,1} = 3$	32	23	4.71	0.81	4.59	1.06
	$UTL_1 = 6.26$	$b_{1,2} = 1$						
Π_2	$LTL_2 = 13.85$	$b_{2,2} = 0$	10	07	21.45	4.86	21.45	4.86
Π_3	—	—	2	—	—	—	—	—

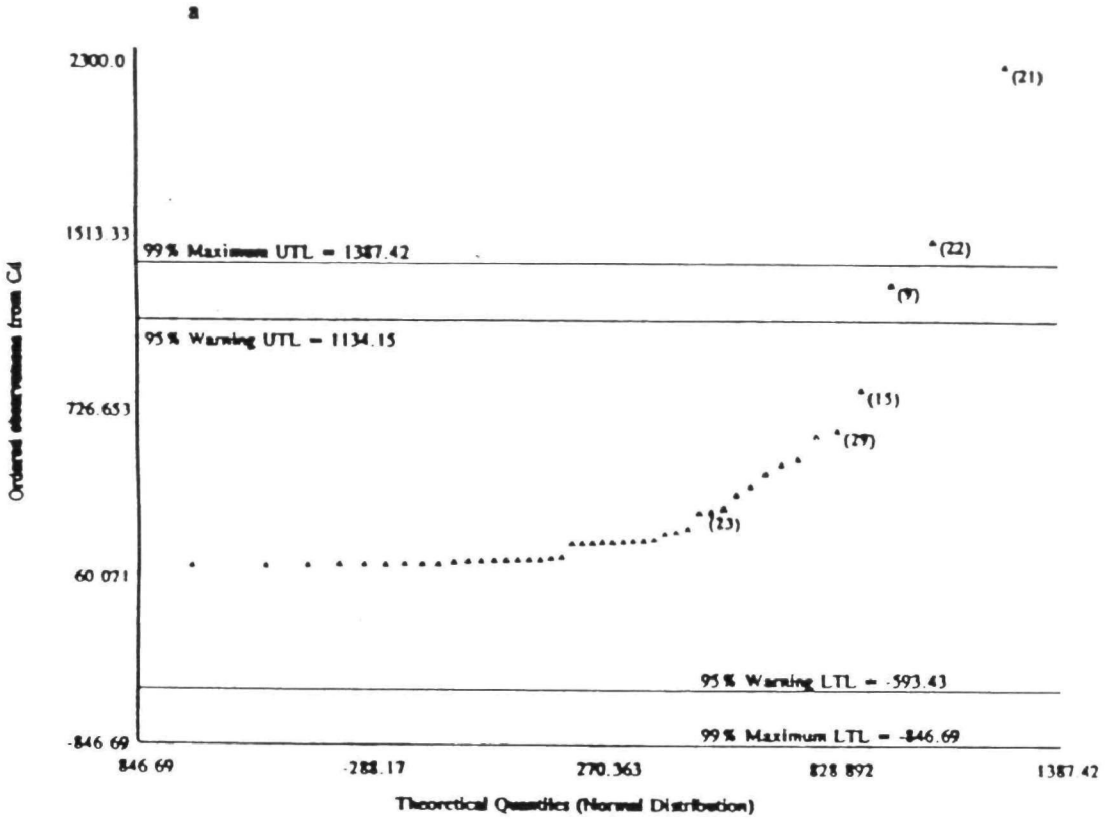


Fig. 3a. Cadmium concentration from a superfund site-classical Q-Q plot.

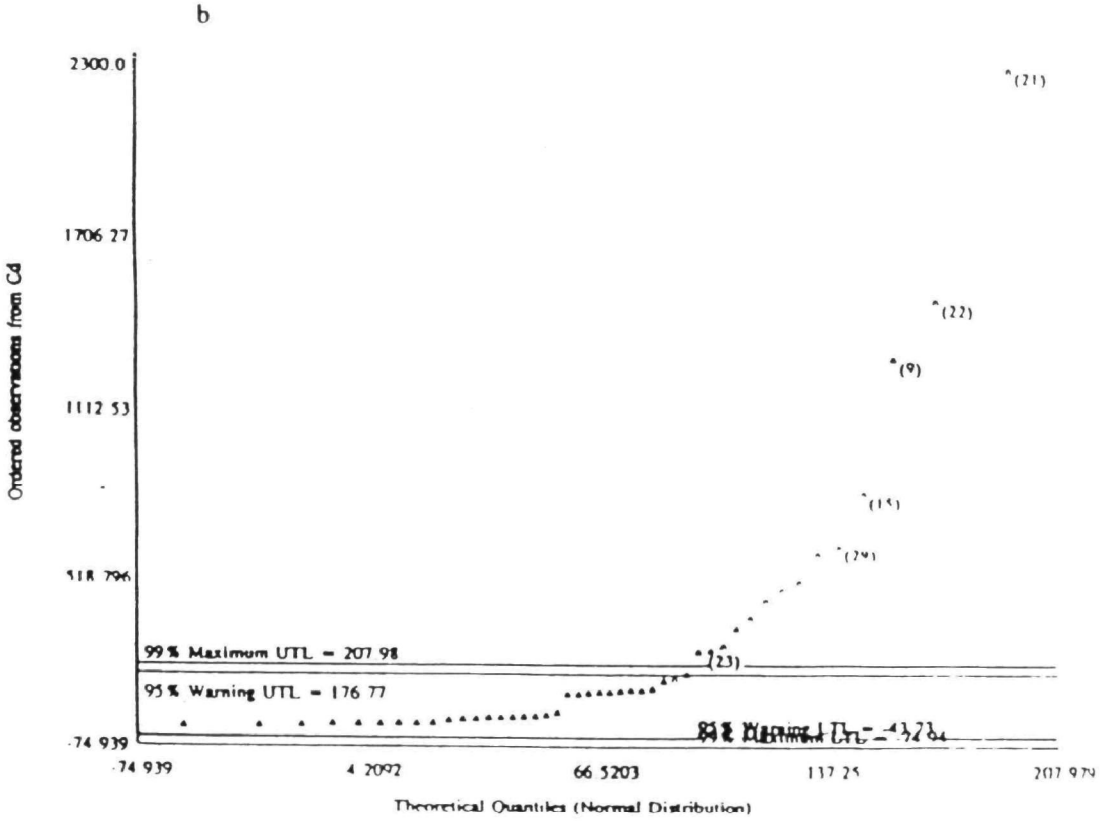


Fig. 3b. Cadmium concentration form a superfund site-robust Q-Q plot.

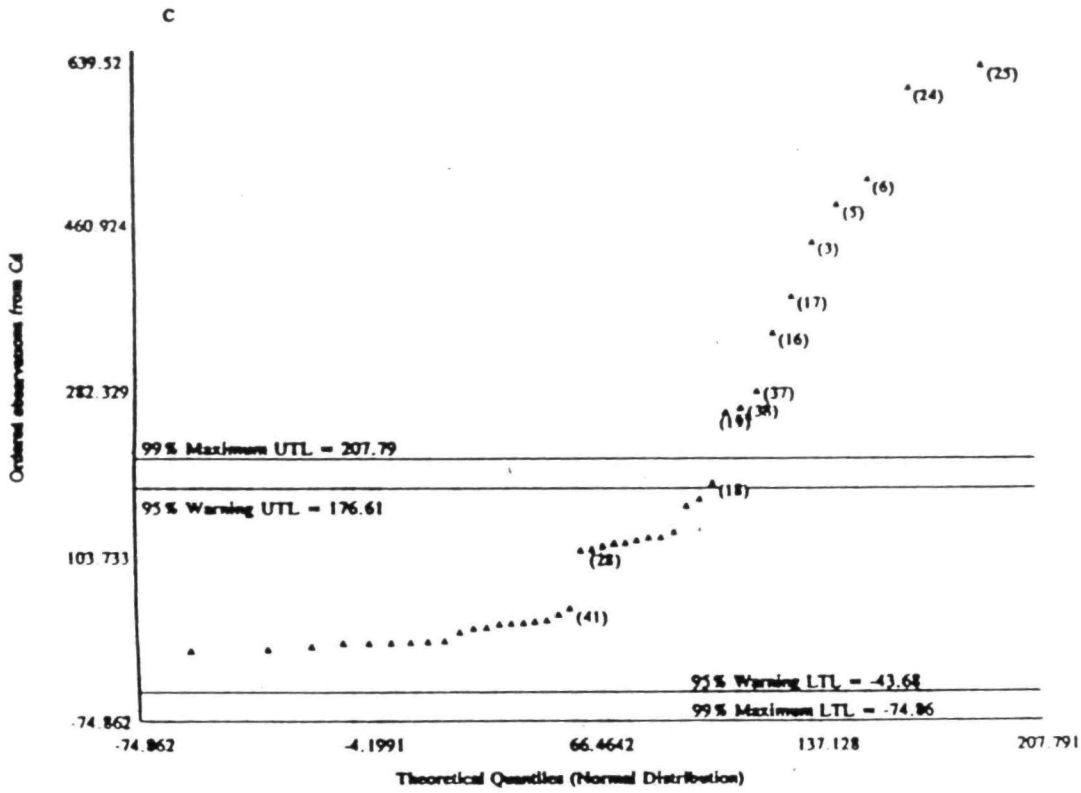


Fig. 3c. Cd conc. from a superfund site-robust Q–Q plot/extremes removed.

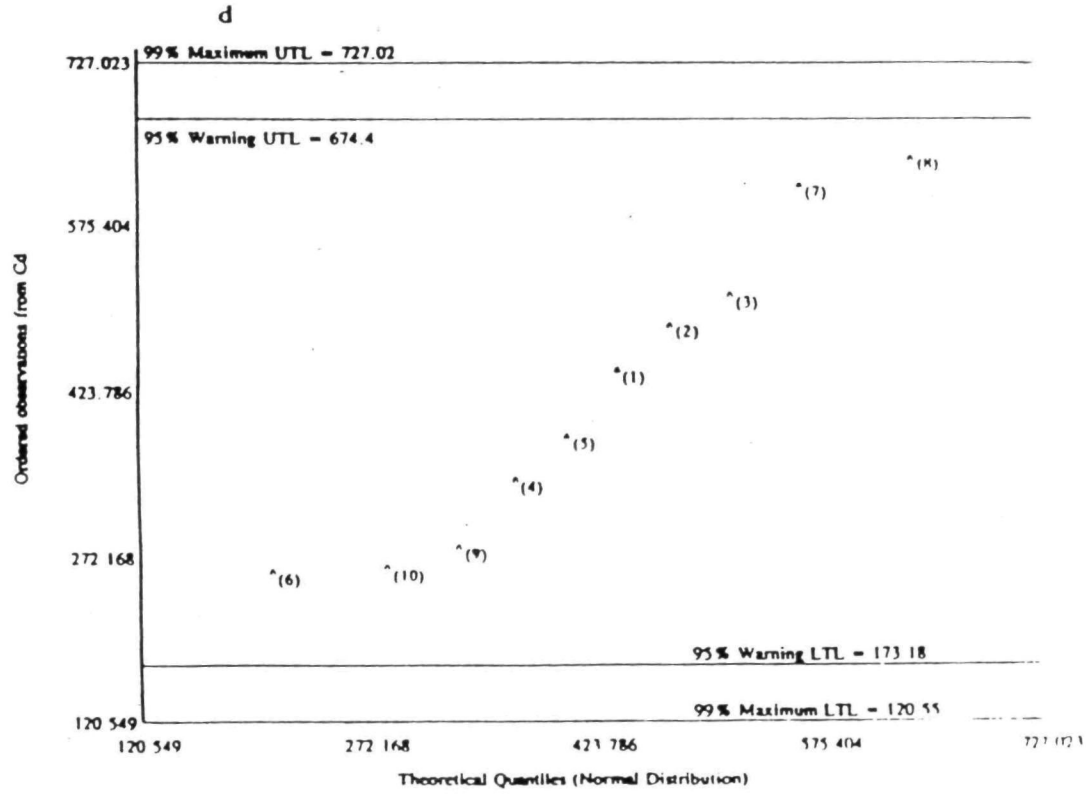


Fig. 3d. Cd. conc. for a superfund site-robust Q–Q plot-high conc.

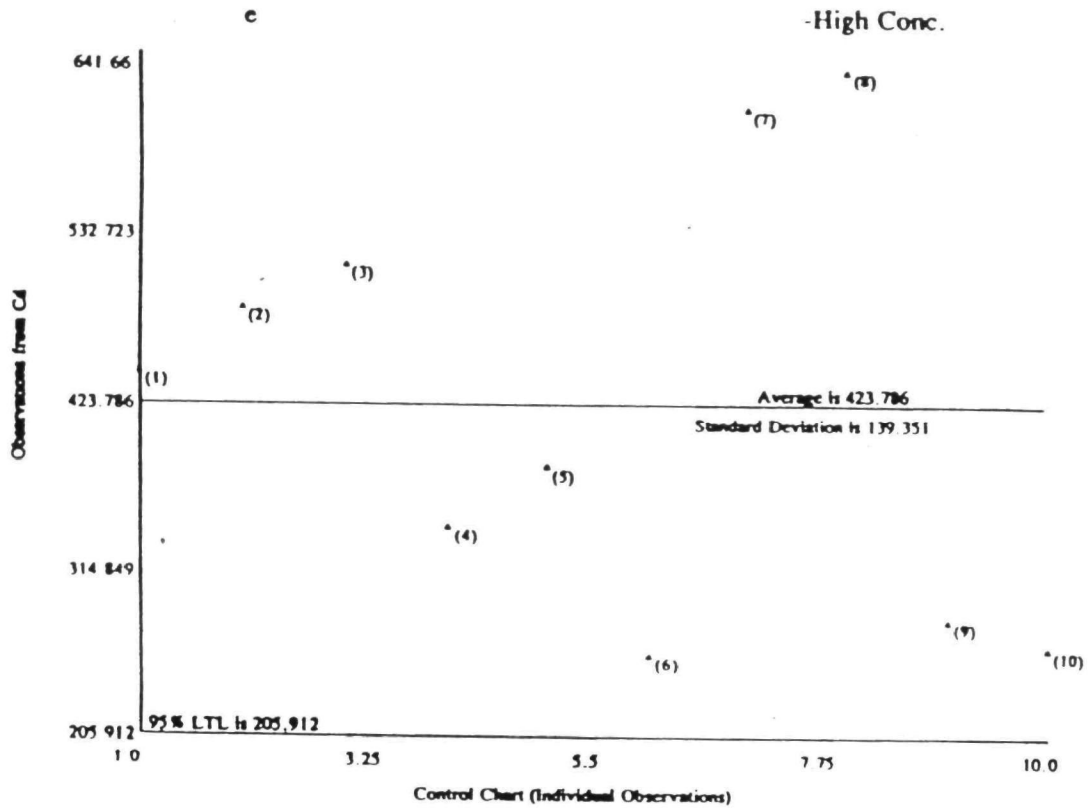


Fig. 3e. Cd. conc. for a superfund site-robust chart-high conc.

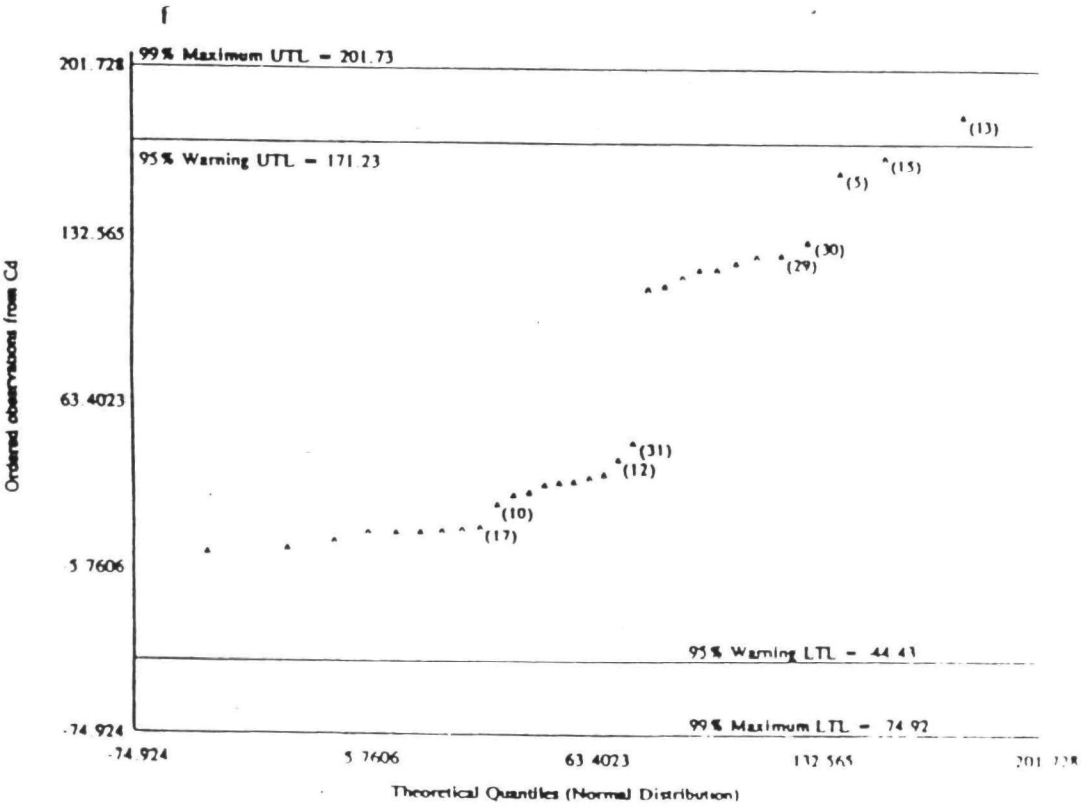


Fig. 3f. Cd. conc. for a superfund site-robust Q-Q plot-high conc. removed.

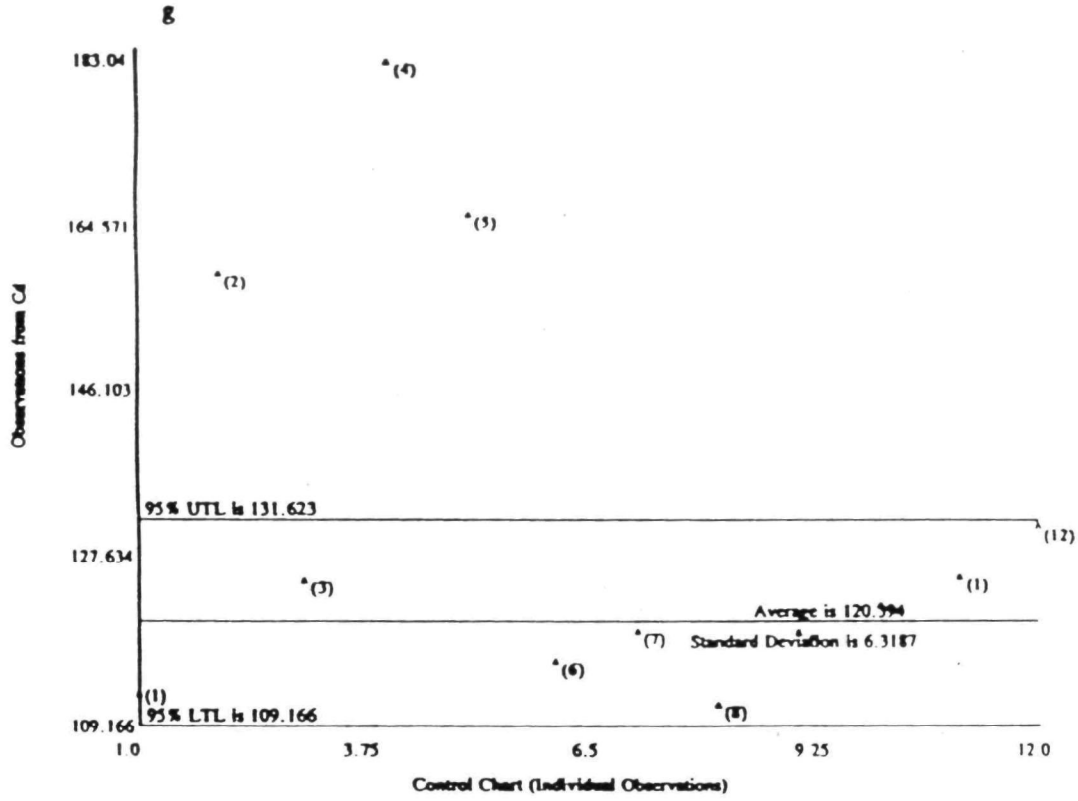


Fig. 3g. Cd. conc. for a superfund site-robust Q-Q chart-intermediate conc.

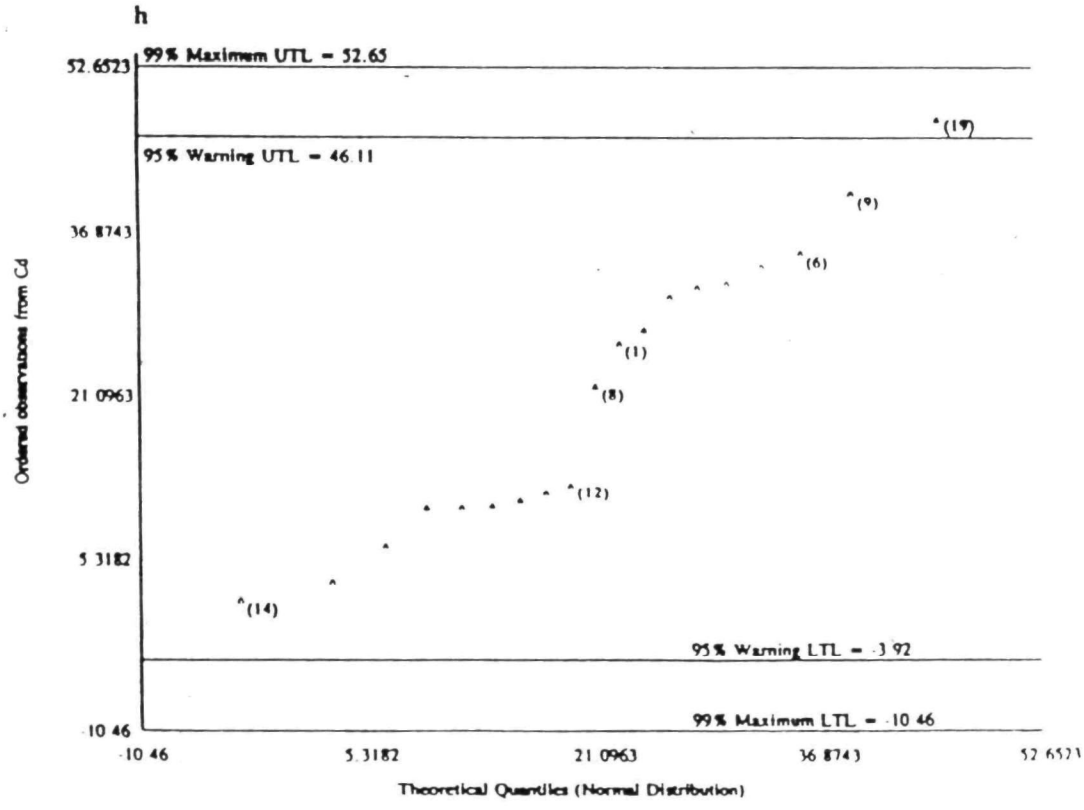


Fig. 3h. Cd. conc. for a superfund site-robust Q-Q plot-highest 2 conc removed

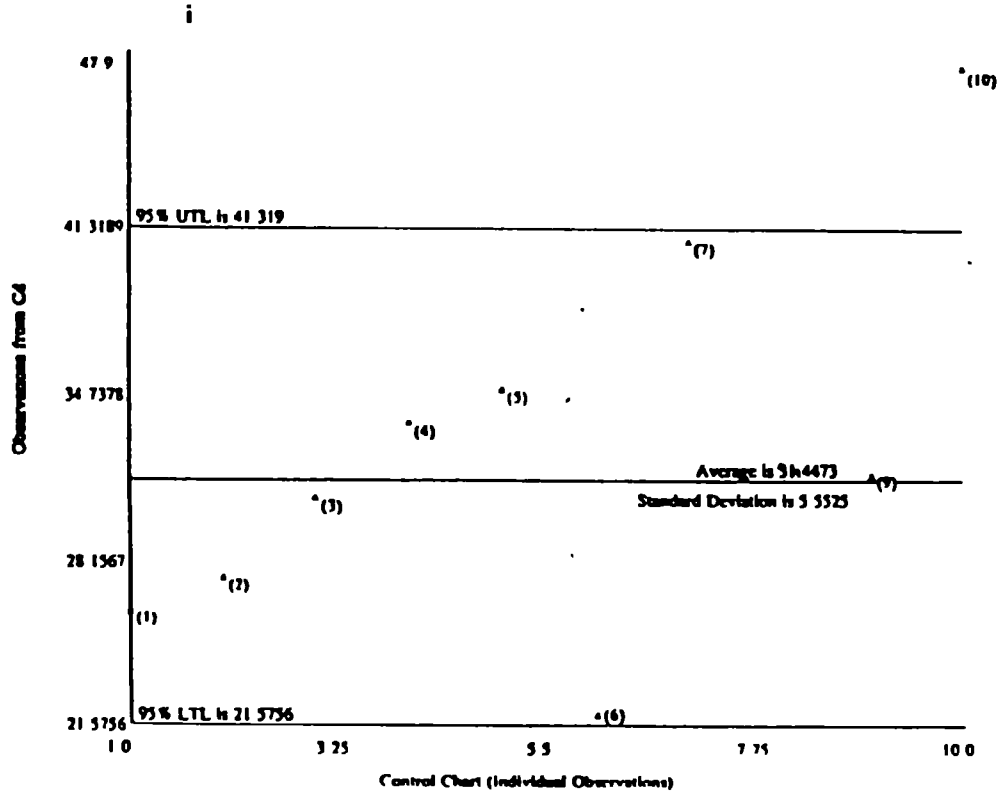


Fig. 3i. Cd conc for a superfund site-robust chart-intermediate conc

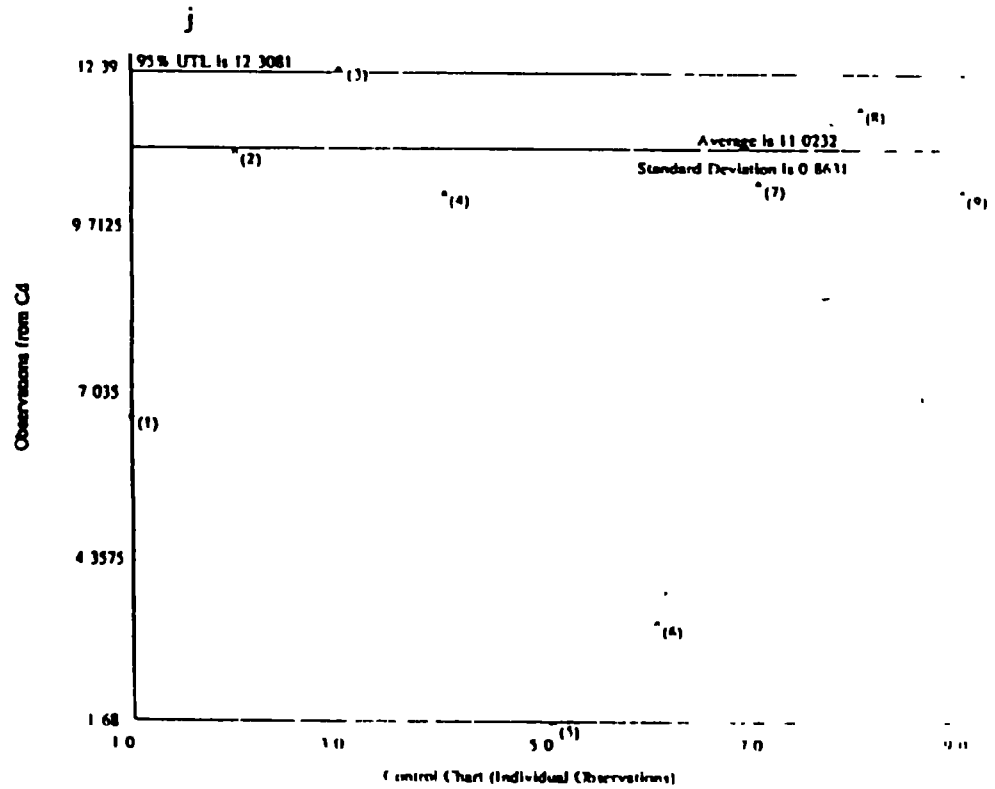


Fig. 3j. Cd conc for a superfund site-robust chart-background conc

a single population. The one-sided lower 95% boundary for this population is $LTL_3 = 205.91$ as given in Fig. 3e.

A new robust Q–Q plot using only the unclassified observations is given in Fig. 3f. There is a clear indication of the presence of three more populations. The robust boundary (109.166, 131.623) given in Fig. 3g for the intermediate population Π_2 is obtained using the top 12 observations of Fig. 3f, with $c_2 = 111.60$ as the cutoff point. Observation nos. 2, 4, and 5 (with $b_3 = 3$) of Fig. 3g belong to the unclaimed region (UTL_2, LTL_3) and will be assigned to appro-

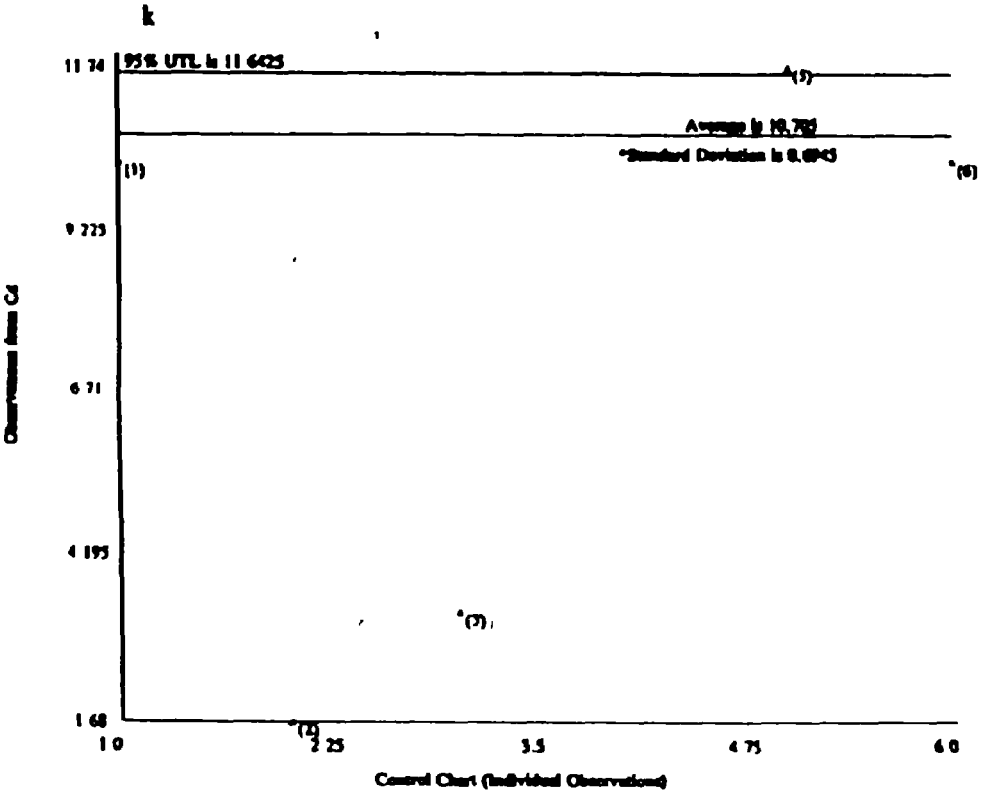


Fig. 3k. Cd conc for a superfund site-robust chart-known background conc

Table III

Π_i	95% Limits	a_i, b_i	n_i	p_i	$\bar{x}^*$	s^*	$\bar{x}$	s
Π_0	$UTL_0 = 12\ 31$	—	9	22	11 02	0 86	8 66	3 85
Π_1	$LTL_1 = 21\ 56$ $UTL_1 = 41.32$	$b_{1,2} = 1$	10	24	31 45	5 55	32 78	7 39
Π_2	$LTL_2 = 109\ 2$ $UTL_2 = 131\ 6$	$b_{2,3} = 2$	11	27	120 39	6 32	128 07	18 11
Π_3	$LTL_3 = 205\ 9$	$b_{3,4} = 1$	11	27	401 9	150 82	401 9	150 82
n_i	—	—	4	—	--	—	—	--

priate populations using the nearest neighbor technique (see Table III). Next, a new Robust Q–Q plot using only the remaining unclassified observations is given in Fig. 3h, giving a clear indication of the presence of two populations with the cutoff point $c_1 = 22.05$ (observation no. 12 in Fig. 3h). The 95% robust boundary = (21.576, 41.319) for population Π_1, using the top ten observations of Fig. 3h, is given in Fig. 3i, with one observation belonging to the unclaimed region (UTL_1, LTL_2), with $b_2 = 1$. Finally, using the last nine observations, the 95% upper threshold value for the background level contamination is $UTL_0 = 12.308$ as can be seen in Fig. 3j. However, in this case, six samples from the background were also available. The robust 95% upper boundary using these six background samples is given in Fig. 3k. The values in Figs. 3j and k are in close agreement, establishing the correctness and validity of the procedure described in this article. All relevant statistics after the final classification have been summarized in Table III.

Example 4. In this example, we consider a simulated data set (given in the Appendix) which consists of a mixture sample from two lognormal populations with some overlap. A sample of size 20 is obtained from a log $N(0, 1)$ population and a sample of ten is generated from a log $N(4, 2)$. We use this example to show the effectiveness of the proposed robust procedure in decomposing the mixture into component populations The classical Q–Q plots of the untransformed and the log-transformed data are given in Figs. 4a and b, respectively. From Fig. 4a, it can be concluded that the sample is from a single positively skewed population with observation no. 22 as an extreme observation. This may lead the user to take the log-transformation From Fig. 4b, one can conclude that the mixture sample comes from a lognormal distribution with observation no. 22 to be slightly discordant. The corresponding robust Q–Q plot before and after the log-transformation are given by Figs. 4c and d, respectively Figure 4c suggests that there are more than one population present. Figure 4d clearly separates the two underlying log-normal populations with cutoff point $c_1 = 1.61$. All of the relevant statistics are summarized as follows in Table IV.

CONCLUSIONS AND RECOMMENDATIONS

The proposed robust procedure works quite effectively in classifying a mixture sample into its component populations. In all of the examples discussed here, the procedure described here classified the observations correctly into their respective populations. When the data represent a mixture from lognormal populations, the procedure based upon the classical MLE estimates may identify some of these observations as anomalous However, the robust procedure described here gives an indication that there is more than one population present (e.g., see Fig 4c) This in turn, forces the user to verify the distributional assumptions It is assumed that the user has some familiarity with symmetric

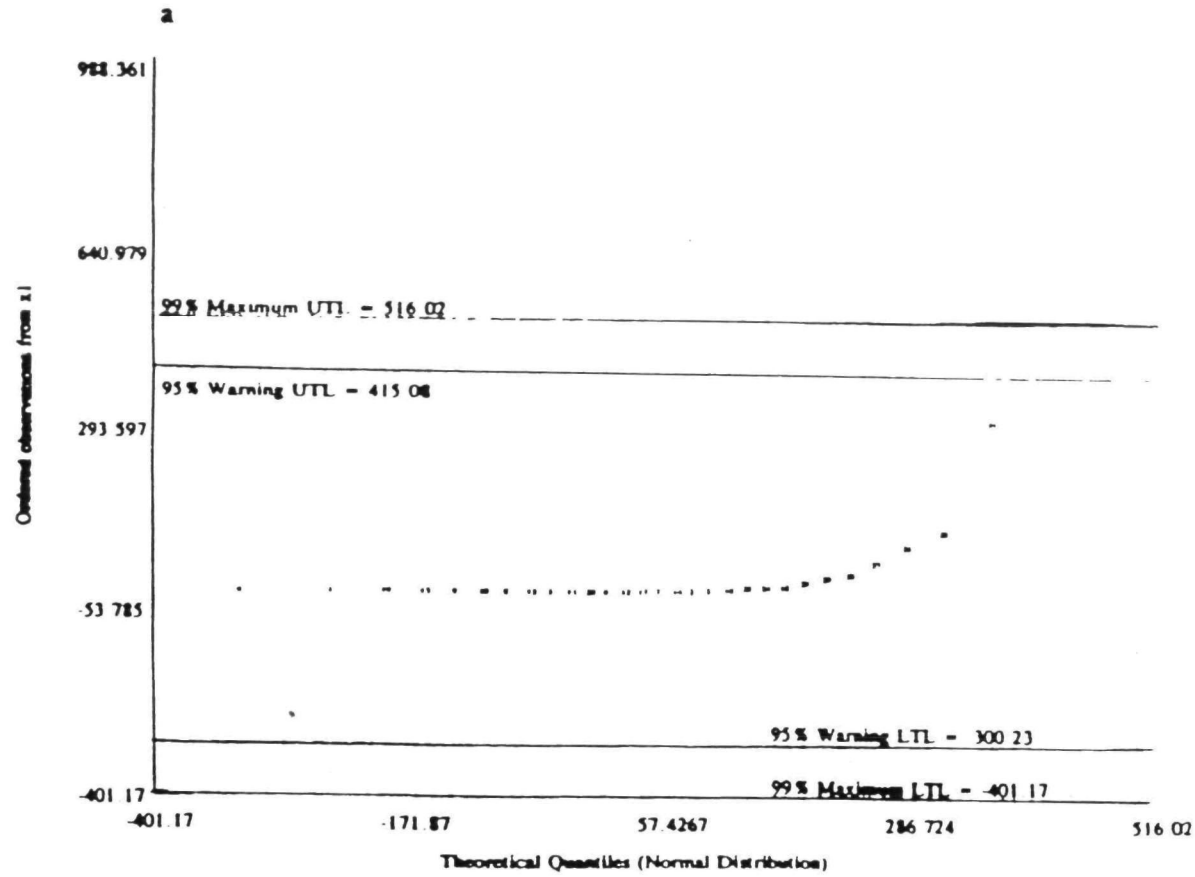


Fig. 4a. Mixture of log $N(0, 1)$ and log $N(4, 2)$ classical Q-Q plot (untransformed).

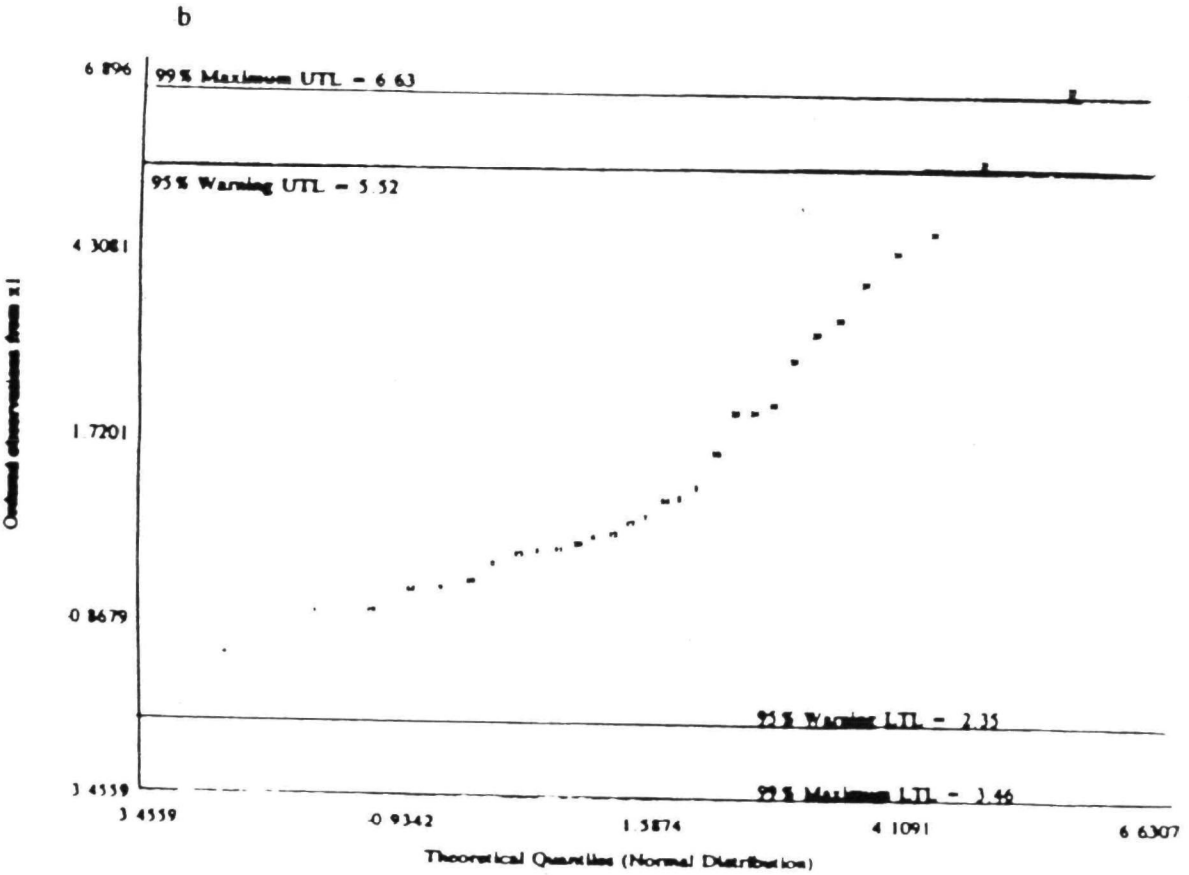


Fig. 4b. Mixture of log $N(0, 1)$ and log $N(4, 2)$ classical Q-Q plot (transformed).

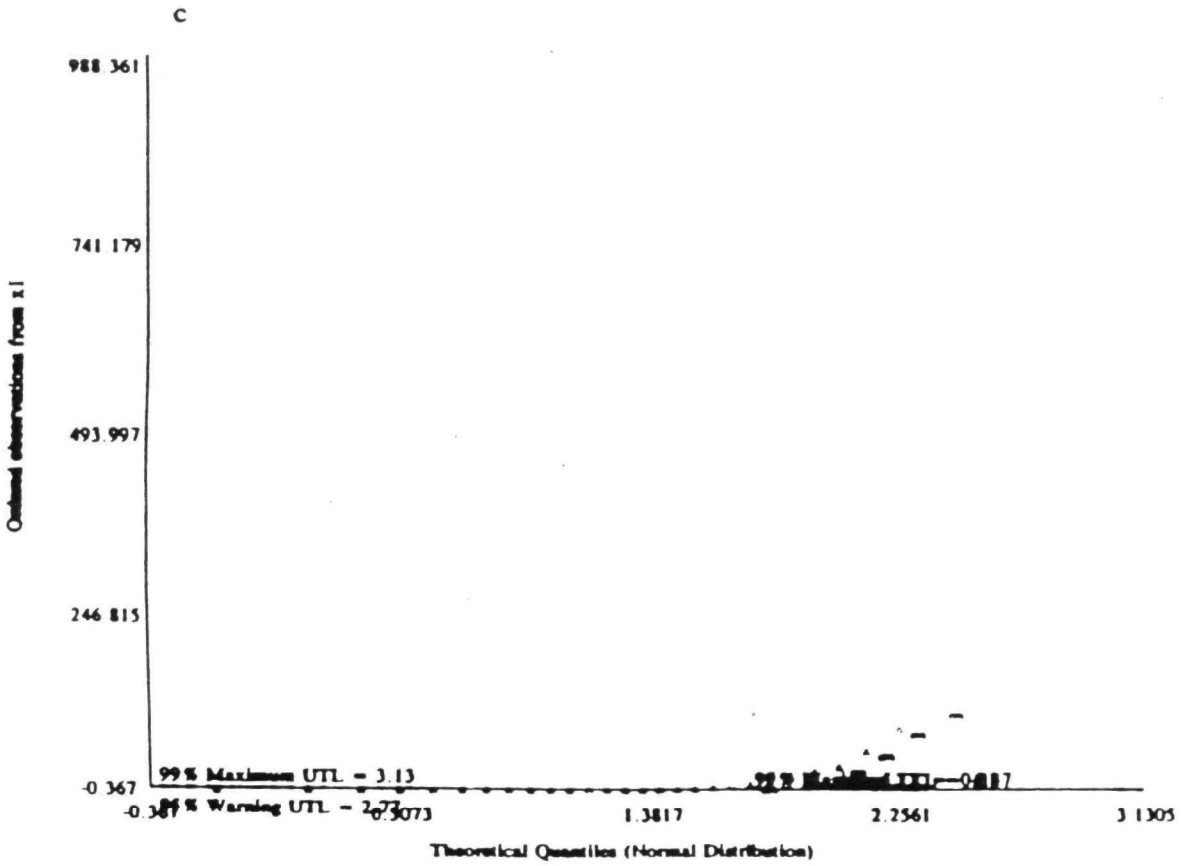


Fig. 4c. Mixture of log $N(0, 1)$ and log $N(4, 2)$ robust Q–Q plot (untransformed).

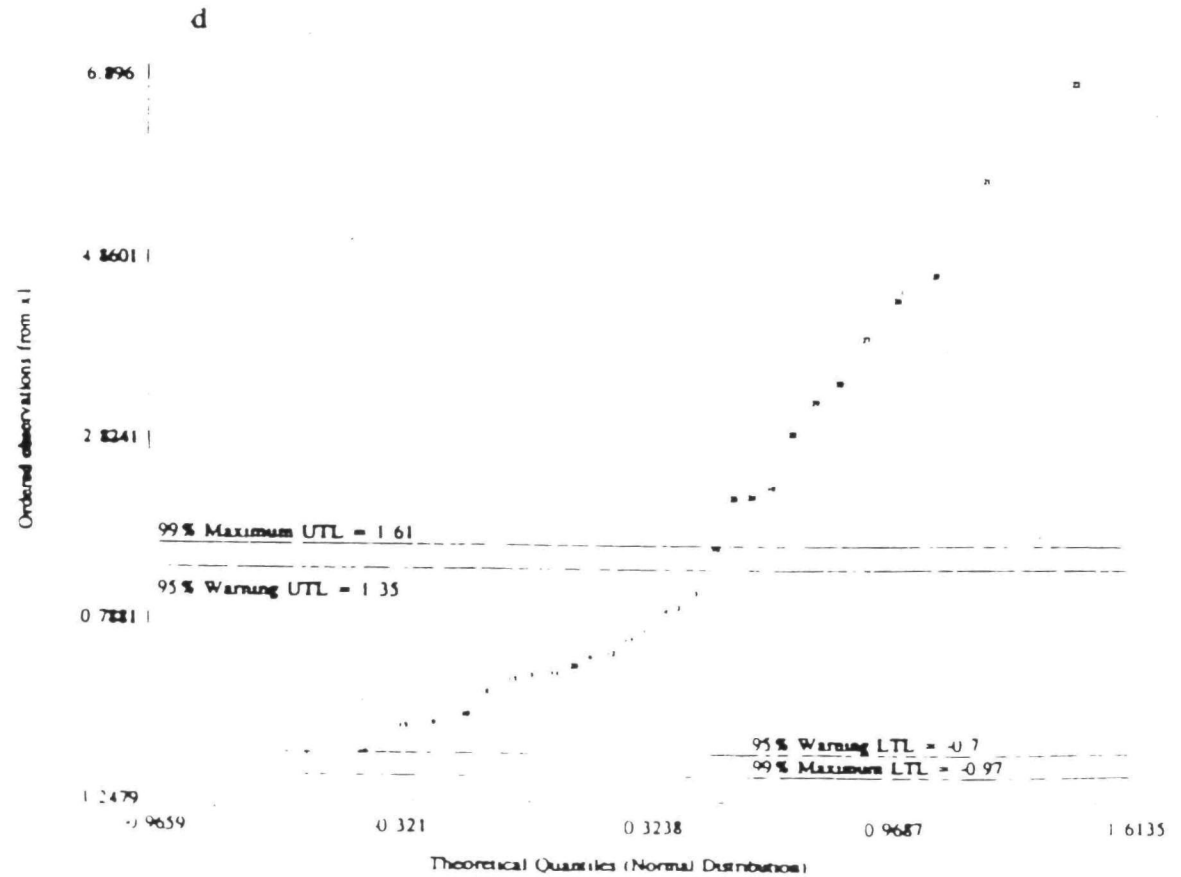


Fig. 4d. Mixture of log $N(0, 1)$ and log $N(4, 2)$ robust Q–Q plot (transformed).

Table IV

Popn	95% Limits	a_i, b_i	n_i	p_i	$\bar{x}^*$	s^*	$\bar{x}$	s
H_0	$UTL_0 = 1\ 143$	—	18	0 6	0 242	0 563	0 187	0 643
H_1	$LTL_1 = 1\ 419$	—	12	0 4	3 503	1 324	3 688	1 58

and skewed distributions. It is the user's responsibility to achieve near-normality (or at least symmetry) for each of the component populations before using the procedure described here. The robust procedure described here works quite effectively in decomposing a mixture sample into its component lognormal populations as well (see Fig. 4d). The stepwise procedure described here combines the natural separation between the component populations. The sample from the Sacramento Army Depot Superfund Site included a known site-specific background sample This, however, is not the case for many Superfund sites The proposed statistical procedure will be a very useful tool for estimation of site-specific background for such Superfund sites.

ACKNOWLEDGMENTS

The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), partially funded and collaborated in the research described here It has been subjected to the Agency's peer review and has been approved as an EPA publication The U S Government has a non-exclusive, royalty-free license in and to any copyright covering this article The authors wish to thank Ken Brown of U.S. EPA\EMSL-Las Vegas for providing us the Superfund site data and for helpful suggestions during the preparation of this paper

APPENDIX

Dataset 1

Normal mixture generated from populations $N(10, 3)$ and $N(27, 8)$ .90 observations are from $N(10, 3)$ and 10 are from $N(27, 8)$ 2 49, 11 15 10 47, 10 62, 12 65, 13 52, 11 02, 13 40, 9.50, 6 93 11 54, 6 83, 10 68, 10 38, 8 16, 10 57, 6 02, 6 49, 10 98, 6 25, 11 45, 12 31, 7 95, 13 89, 9 87, 10 10, 10 50, 11 95, 10 16, 11 09, 7 35, 11 01, 10 26, 12 06, 16 11, 12 03, 12 62 10 29, 14 63, 11 65, 13 13, 7 93, 8 18, 11 11, 7 95 8 15, 14 20, 7 99, 13 31, 9 63, 8 82, 8 42, 7 32, 18 59, 7 97, 6 43, 13 39 3 59 7 40, 12 73 8 59 13 34, 8 34, 5 71, 8 34, 8 29, 11 99, 11 23, 5 26 9 04 7 12, 14 85 11 08

10 11, 11.01, 9.57, 11 01, 12.25, 7.93, 4.48, 9.13, 6.58, 13.89, 6.70, 12 04, 7.69, 10.84, 9.13, 6.84, 10.33, 33.38, 23.49, 30.01, 37.23, 37.66, 31 27, 34.94, 9.48, 31.08, 43.99.

Dataset 2:

Normal mixture with ten observations from $N(20, 4)$, 100 from $N(0, 1)$, 30 from a $N(5, 1)$, and two extreme observations from $N(100, 10)$ 18 12, 16.60, 27.60, 23.27, 29.80, 18.24, 24.40, 23.04, 16.98, 16.41, 1 77, 2 38, −0 22, −0.35, −0.40, 1.00, −0.01, −0.16, 1.44, −1.03, −1.84, 0.94, −0 31, −1 03, 1.19, −0.14, −1.42, −0.89, −0.23, 0.18, −0 96, −0.17, 0 06, 1.62, −0.03, −0.25, 0.30, 2.48, −0.02, 1.23, 0.10, 1.13, −0.69, 0 72, −0.86, 0 11, 1 16, 0 75, 0 27, −1.40, 0.29, −0.52, 2.47, 1.01, 1 89, −0 58, 0 20, −0.66, −1 05, −0.10, 1.44, 0.72, 0.33, 1.06, 0 48, −0 69, −0 48, −1 13, −0 67, 0.12, −0 15, −0.10, −2.54, 0.25, −2.04, 0 55, −1.32, −0 09, 0 51, 0.06, 1 54, 0 81, −1.65, −0.39, −0.01, 0.41, −0 51, −0 60, 1 24, −1.48, 0.51, 0 13, 0.93, −2.17, 0.63, −0.39, −1.37, 1 17, −1 29, −0 10, 0.30, 0 84, −0 11, 1.66, −0.66, −0.50, −0 87, −1 59, −0.69, −2 01, 4 16, 3 97, 4 18, 3 71, 4.55, 3.45, 5.62, 6.67, 4 25, 4.76, 5 24, 5 78, 5 23, 6.20, 1 18, 5 62, 4.51, 5.35, 4.34, 4.77, 6.07, 4 24, 4 26, 3 77, 5 16, 4 07, 5 46, 3.80, 5 50, 4.84, 123.76, 117.61.

Dataset 3.

Cadmium concentrations from the Sacramento Army Depot Superfund Site 26 20, 27 55, 445 01, 30 77, 486.31, 513.79, 112 81, 159.30, 1300, 6 68, 33 72, 35 01, 10.99, 22 05, 830 94, 125.07, 40.84, 345 52, 384 80, 183 04, 2300, 1500, 260.27, 32.09, 166.16, 31.68, 12.39, 614 53, 639.52, 116 24, 119 43, 111.60, 10 29, 1.68, 3.34, 10.47, 11.74, 10 32, 122 30, 283 03, 265.08, 125 49, 131.06, 47.90, 119.34.

Dataset 4:

Mixture of 20 observations from lognormal $N(0, 1)$ and 10 from lognormal $N(4, 2)$ 0 5300, 2.7538, 3 2237, 0.2871, 1.2915, 1 5795, 2 0817, 1 0633, 0 7486, 0 8284, 1 3252, 1 6477, 1 2311, 2 6518, 0 7258, 5 2913, 1 9187, 10 3898, 0 5373, 1 4311, 332.1949, 988.3606, 19 3491, 9 3424, 9 2353 88 1362, 56 3981, 115 9378, 27.8464, 34.4647

REFERENCES

Campbell, N. A., 1984, Mixture Models and Atypical Values: Math. Geol., v. 16, n. 5, p. 465–477.

Fleischhauer, H., and Korte, N., 1990, *Formation of Cleanup Standards for Trace Elements with*

Probability Plots, Environmental Management (Vol 14, No 1) Springer-Verlag, New York, p. 95-105

Fowlkes, E B , 1979, Some Methods for Studying the Mixture of two Normal (Lognormal) Distributions *J Am Stat Assoc* , v 74, n 367, p 561-575

Holgersson, M , and Jorner, U , 1978, Decomposition of a Mixture into Normal Components, A Review J Bio-Med Comp , v 9, p 367-392

Sinclair, A J , 1976, *Applications of Probability Graphs in Mineral Exploration, Assoc of Exploration Geochemists*. Rexdale, Ontario, p 95

Singh, A , 1994, Omnibus Robust Procedures for Assessment of Multivariate Normality and Detection of Multivariate Outliers, *in* G. P Patil and C R Rao, eds , *Multivariate Environmental Statistics* North Holland, Elsevier Science Publishers, p 445-488.

Singh, A , and Nocerino, J 1994, Robust QA/QC for Environmental Applications, *in The Proceedings of the Ninth International Conference on Systems Engineering* University of Nevada, Las Vegas, p. 370-374

Representativeness in Statistics and Quality Assurance

John Warren

Quality Assurance Division
Office of Research & Development

EPA Conf – April 1, 1997

Representativeness Influences:

Data aggregation:

- Merging data sets having similar Quality Assurance protocols collected using probabilistic sampling frames

- Merging data sets having a probability basis with similar data with a non-probabilistic basis

Representativeness Influences:

Hypothesis testing:

- Comparing data sets with different extraction methods and different sample matrices

- Comparing data sets having both within and between differences in the setting of the minimum detection levels and data editing

Factors Influencing Representativeness

Sample Selection Technique:

o Probabilistic:

- Systematic with SRS
- Composite with SRS
- Adaptive with any other

o Non-probabilistic:

- Judgmental
- "Found data"

Factors Influencing Representativeness

Sample Analysis Methodology:

- o Intra/Inter laboratory differences
- o Method equivalence problems
- o Heterogeneous sample matrices
- o Variation in Quality Control
  - Calibration frequencies
  - Detection levels
  - Laboratory protocols
  - Extraction efficiencies

Statisticians Are Little Help

A Dictionary of Statistical Terms
F.H.C. Marriott, 1990 International Statistical Institute

Representative Sample:

In the widest sense, a sample which is representative of the population. Some confusion arises according to whether 'representativeness' is regarded as meaning 'selected by some process which gives all samples an equal chance of appearing to represent the population'; or, alternatively, whether it means 'typical in respect of certain characteristics, however chosen'. On the whole, it seems best to confine the word 'representative' to samples that turn out to be so, however chosen, rather than apply it to those with the objective of being representative.

Kruskal and Mosteller : 1979

Three papers in *International Statistical Review*
"Representative Sampling" commonly applied to:

1. as a "seal of approval"
2. to denote "absence of selective forces"
3. as a "miniature of the population"
4. as being a "typical or ideal case"
5. to denote "coverage of a population"
6. as a "vague term to be made more precise"
7. as a "specific sampling method"
8. as "permitting good estimation"
9. as "good enough for a particular purpose"

"Seal of approval"

No explanation provided of what process was used to go from target population to sampled population

Use of "representative" is to convince the reader to have faith in the reported results and therefore the truthfulness of the conclusions

"Absence of selective forces"

Used to imply that the sampling method used deliberately excluded selective forces that might over-represent some sub-population

Highly vulnerable to personal bias in elimination methodology:

"Miniature of the population"

Implies that every nuance of the population is reflected in the sample i.e. identical frequency distributions for sample and population.

In practice, it is obvious this cannot be acheived

"Typical or ideal case"

Inevitably only a single specimen from the population has been selected

Tremendous possibility of bias but the implication is that an "ideal specimen" has been selected without true definition of whether this implies "average", "worst case", or "best case"

"Coverage of the population"

The implication is that the sample selected has a wide range across the population. At least one example from each class or potential partition (stratum) has been collected but the appropriate weighting factors not made available.

"Vague term to be made more precise"

The word "representativeness" is used as a promise of things to come from a more detailed (not specified) technical consideration of the problem. The use of the term is intended to give permission to discuss a problem without getting sidetracked by technical details

"Specific sampling method"

This is the use of "representative sampling" when really the true kind of sampling has been deemed by the author to be too complex for the audience's comprehension. The intent of the author: understanding by the majority, over-riding the true comprehension of the minority (often statisticians)

"Permitting good estimation"

The connotation that because some sample can be labeled "representative" it will therefore allow for satisfactory estimation without the necessity of defining what this actually implies.

"Good enough for a particular purpose"

This is the use of a sample to illustrate a particular theory or hypothesis. It is a variation on the concept of using a sample size 1 in that a counter-example (non-random sample) can be enough to prove a case.

Representativeness as an Indicator

Data Quality Indicators: **PARCC**

Precision
Accuracy (really Bias)
Representativeness
Comparability
Completeness

PARCC: Representativeness

- Qualitative measure
- Open to individual interpretation
- Depends on media homogeneity
- Difficult to ensure
- Often demands many samples
- Needs expert opinion

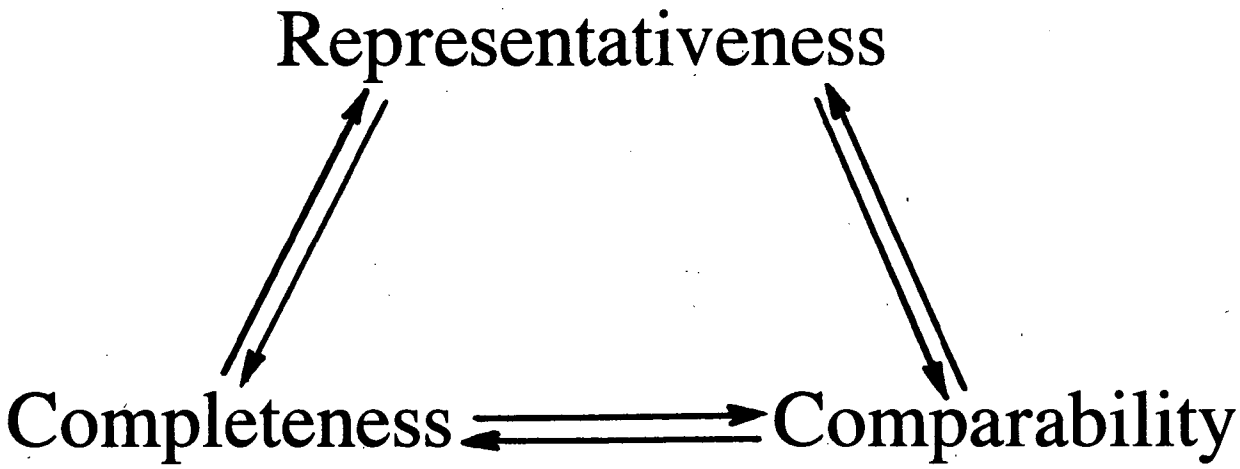
PARCC: Comparability

- o Qualitative measure
- o Expresses a degree of confidence
- o Requires same variables of interest
- o Needs units convertible to a standard
- o Requires similar analytical procedures
- o Needs compatible rules for data editing
- o Requires similar sampling frames
- o Needs meaningful temporal limits
- o Requires expert opinion

PARCC: Completeness

- Quantitative
- Influence depends on sample design
- If unbiased - loss of power
- If biased - loss of validity
- Needs expert opinion

PARCC are Interrelated



Regulatory use of Representativeness

Essentially never defined

Water (40 CFR 403)	"...samples should be representative of daily conditions"
Air (40 CFR 51)	"...selected on the basis of spatial and climatological (temporal) representativeness"
TSCA (40 CFR 763)	"...at locations representative of the air entering the abatement site"
RCRA (40 CFR 260)	"...a sample of a universe or whole which can be expected to exhibit the average properties of the universe or whole"

Potentially Promising Areas

- o Composite statistics & area of support
- o Combining environmental information
- o Applying Gy's theory of sampling

Composite Statistics & Area of Support

---

o Interpretation of "support"

e.g. Linkage of long-term exposure risk
(10^4sq meters) with remediation technology
(10^3 sq meters) with sampled area
(10^2 sq meters) with physical sample
(10^1 sq meters) with sample analysis
(10^{-1} sq meters) with ...

e.g. geophysical/geostatistical (kriging)

Englund & Flatman: Spatial Statistics Sampling

Composite Statistics & Area of Support

- o Literature and information on composite sampling
  - + *Statistical Methods for Environmental Pollution Monitoring* (R.O. Gilbert)
  - + *Handbook of Statistics vol 12, Chapter 4* (G. Lovison, S.D. Gore, & G.P. Patil)
  - + *Environmental and Ecological Statistics* (Special Edition, G.P. Patil, editor)
  - + *Guidance on Sampling (QA/G-5S)* (Under development by QAD)

Combining Environmental Information

---

o Literature and information on data combining:

+ *Encountered Data, ...and Weighted Distributions* (G.P. Patil) *1991, Environmetrics 2, 377-423*

+ *Using Found Data to Augment a Probability Sample* (J.M. Overton, T.C. Young & W.S. Overton, *1993 Envir. Mon. & Assess 26, 65-83*

+ *Combining Environmental Information I & II* (L.H. Cox & W.W. Piegorsch) *1996, Environmetrics 7, 299- 324*

+ *Guidance on Sampling (QA/G-5S)* (Under development by QAD)

Encountered Data, Statistical Ecology, Environmental Statistics, and Weighted Distribution Methods

---

o Weighted distributions used to account for observer bias due to being unable to actually observe an event or sample value

o If an observation (X) has a probability θx of being observed then the observed pdf is the true pdf weighted by $1 - \theta x$

o Regard the problem as one of modelling when samples are drawn without a proper frame

o The paper contains some theoretical properties of weight functions together with some applications

Using Found Data to Augment a Probability Sample

- o If the variable of interest is in both found and probability based samples, then use a pseudo-random sample approach and combine the data in the manner of a stratified sample

- o If not, use a stratified calibration approach - form a predictor equation for found data by regressing variable of interest on the known frame attributes. Then for the probability based sample, use the prediction equation and the frame attributes to predict new variables of interest

- o Extensive example on streams from the National Surface Water Survey

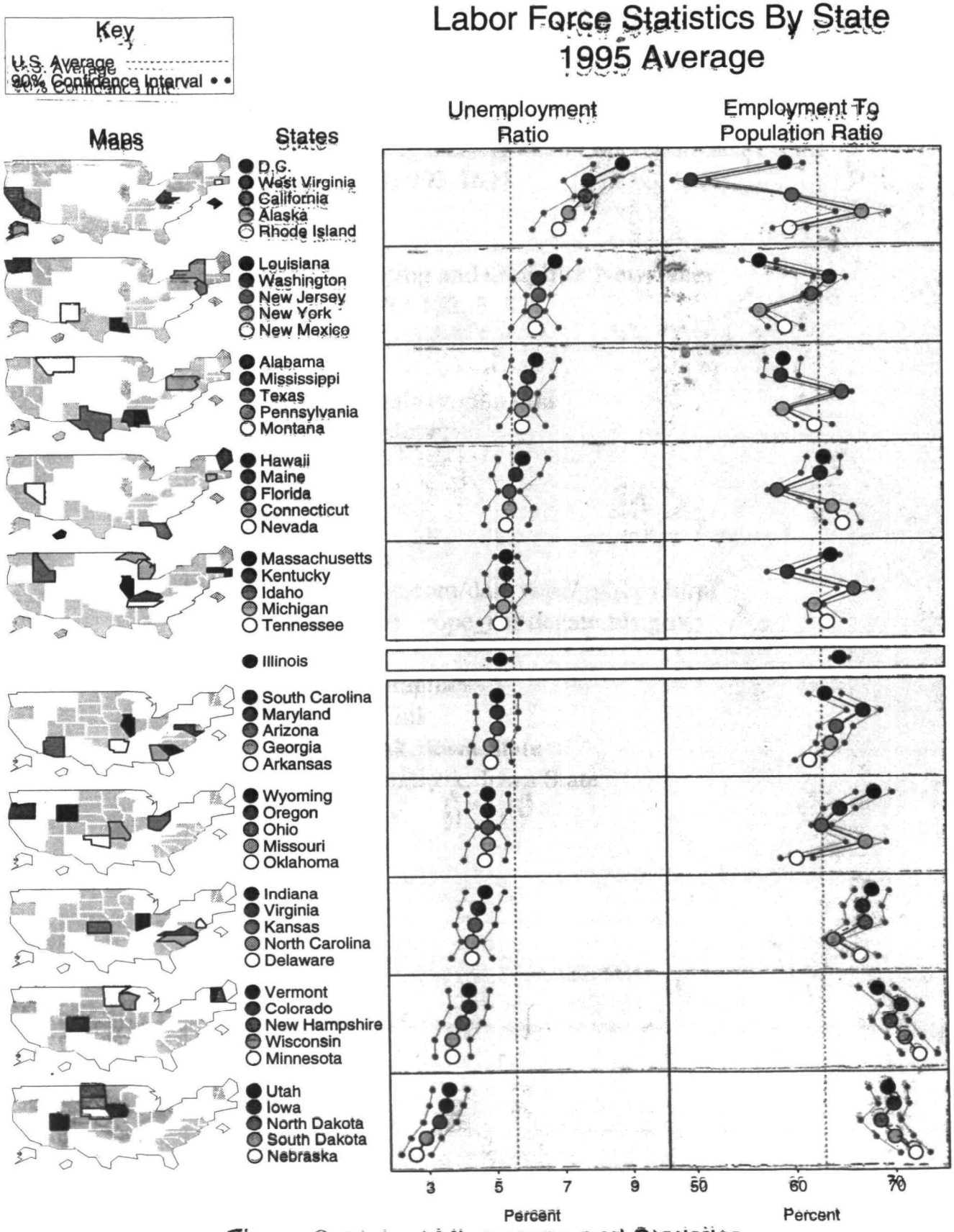


Figure 2: Linked Micromaps and Statistics

Combining Environmental Information I & II

- Two consecutive papers, the first being an overview with potential areas for research, the second considering various applications to epidemiology and toxicology

- The overview includes kriging, non-detect problems, and application to truncated spatial data

- Overview also includes the mathematical aspects of combining p-values (the works of R.A.Fisher, and T. Mathew, B. Sinha, & L. Zhou)

- Examples include passive smoking and dose-response