CLASSIFICATION		DOCUMENT IDENTIFICATION NO. PNL-SA-6697
Battelle Pacific Northwest Laboratories Richland, Washington 99352		COPY AND SERIES NO.	
		DATE JANUARY 1978	
TITLE AND AUTHOR TRAN-STAT, ISSUE #3, JANUARY 1978. TOPICS DISCUSSED: SOME STATISTICAL ASPECTS OF COMPOSITING FIELD SAMPLES. R. O. GILBERT		CONTRACT ☐ - 1830 ☐ - 1831	PROJECT NO.
		RESERVED FOR TECH. INFO. USE	

D I S T R I B U T I O N

NAME	COMPANY	LOCATION	NAME	COMPANY	LOCATION
<div data-bbox="619 1165 1007 1417" data-label="Text" style="border: 1px solid black; padding: 5px; margin: 10px auto; width: 80%;"> <p style="text-align: center;">NOTICE</p> <p>This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.</p> </div> <div data-bbox="1106 1344 1329 1501" data-label="Text" style="font-size: 2em; transform: rotate(-10deg); margin: 20px auto;"> MASTER </div> <div data-bbox="933 1543 1569 1585" data-label="Text" style="color: red; text-align: right; margin-top: 20px;"> DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED </div>					

ROUTE TO	PAYROLL NO.	COMPANY	LOCATION	FILES ROUTE DATE	SIGNATURE AND DATE

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

TRAN - STAT

Statistics for Environmental Transuranic Studies

February 1978
Number 3

Battelle Memorial Institute
Pacific Northwest Laboratory
Richland, Washington 99352

Coordinator: Richard O. Gilbert, Energy Systems
Department, Statistics Section,
(FTS) 444-7511, Ext. 946-2378

This issue is devoted primarily to discussing some statistical aspects of compositing field samples. We begin by listing our summary thoughts that identify some of the potential problems involved in pooling samples. This is followed by more detailed discussions and examples of compositing designs, adequacy of mixing, statistical models and their role in compositing, and other topics. Those readers interested in gaining a general awareness of potential problems may not feel a need to read these later sections. I would hope, however, that most of you will read to the end and be motivated to think more about the problems discussed.

This issue has benefited greatly from comments made on an earlier draft by the following individuals: Gary White (LASL), John Pinder (SREL) Lee Eberhardt, Ted Hinds, Gene Schreckhise and Wallie Weimer (BNWL). I also appreciate the phone discussions I had with these and other scientists who have either composited samples or have considered the possibility.

COMMENTS ON 2ND ISSUE OF TRAN-STAT

I received comments from several of you on the 2nd issue of TRAN-STAT (November 1977). One statistician noted that perhaps we should be looking more closely at the statistical aspects of extreme data points rather than at tests of hypotheses or the estimation of average concentrations; the argument being that even one hot spot or one hot particle could cause a health risk. There should be a considerable statistical literature available concerning the analysis of extreme data points and it should be closely examined for its relevance to Pu field studies.

NEW REFERENCE ON ENVIRONMENTAL MONITORING

I recently purchased the paper-bound "Environmental Monitoring", a report to the U. S. Environmental Protection Agency from the Study Group on Environmental Monitoring, published by the National Academy of Sciences, Washington D.C. 1977. This is volume IV in a series of studies performed by the National Research Council for EPA on the acquisition and use of scientific information in environmental decision making. While the book does not discuss environmental sampling for transuranics it does offer much food for thought concerning the collection, manipulation and use of environmental data. Chapter 4 on the management of scientific data, and the paper

by J. Stuart Hunter in Appendix C entitled "Incorporating Uncertainty into Environmental Guidelines" seem particularly relevant to the 2nd issue of TRAN-STAT.

SOME STATISTICAL ASPECTS OF COMPOSITING FIELD SAMPLES

Let us define a "composite sample" to be the mass of material resulting from pooling several individual samples together before any chemical analysis or counting is done. Compositing (pooling) may be done for several reasons including (a) to obtain a more "representative" concentration for fixed costs, and (b) to increase the amount of radioactivity present so that the number of "non-detectable" counts is reduced. In general, compositing becomes attractive when the cost of a single analysis is large relative to the costs of collecting, pooling, and adequately mixing samples.

If pooling is done, either the entire pooled sample may be analyzed yielding a single observed concentration, or one or more (R) aliquots may be withdrawn from the pooled sample and analyzed separately to yield R observations. An aliquot's concentration is an estimate of the value that would be obtained if the entire pooled sample were analyzed. It is always preferable from a statistical point of view to analyze the entire pooled sample, but this may be unduely expensive or simply not feasible in the laboratory. How well an aliquot represents the entire pooled sample depends, clearly, on how well the pooled sample has been homogenized and mixed before aliquots are removed. Most of this issue of TRAN-STAT is concerned with the complications that arise in estimating the variance of a mean when the entire pooled sample is not analyzed.

Most of the information discussed here comes from 3 papers that deal with statistical aspects of compositing; Kussmaul and Anderson (1967), Brown and Fisher (1972) and Rohde (1976). Although none of these papers are directed toward applications to transuranic studies their results appear to be applicable in many respects. Please note, however, that all three papers are primarily concerned with applications where measurement error is negligible, an assumption clearly not applicable to transuranic studies where analytical and counting errors are often large. This is one area where the statistical models discussed here should be extended to meet more fully our needs.

Summary Thoughts and Suggestions Relative to Compositing

- (a) Careful thought should be given to the meaningfulness of pooling samples in each field situation and to the definition of the population from which field samples are drawn. How are the results obtained from pooled samples to be used? Can the scientific questions being asked be answered from pooled samples? Will important variables (changes over time, differences between locations, soil types, vegetation species, microhabitats, etc.) be masked by the pooling and mixing process?
- (b) There are several sampling design models that may be used in pooling samples, some of which allow one to estimate variance components (see the nested designs by Kussmaul and Anderson) and some that do not (see Brown and Fisher, and Rohde). These are discussed in Sections 1 and 3

below.

- (c) Estimates of the variability between the original field samples being pooled cannot be obtained by taking several aliquots from the pooled sample. When variability between field samples is required (perhaps for tests of significance, for estimating the variance of the mean (see Section 3 below), or to compare estimates from other studies) compositing should be avoided or special samples analyzed to obtain such information.
- (d) If several aliquots are withdrawn from a composite sample, the usual estimate of variance computed using the several aliquots will underestimate (on the average) the actual (true) variance. Hence, the estimated variance of the mean will also be too small on the average. This happens because of the correlation that exists between aliquots from the same pooled sample. This correlation exists because each aliquot is made up of portions of the same field samples. See Section 3 for discussion.
- (e) The true variance of an aliquot from a pooled sample depends on the components of variance present in the field samples being pooled together, as well as on a variance component due to the mixing process which we call the "mixing variance" and denote by σ_c^2 . If $Y_1, Y_2, \dots, Y_n$ are the Pu concentrations of n samples pooled into a composite sample, then the concentration of an aliquot withdrawn from the composite is given by (in the notation of Rohde)

$$Z = \alpha_1 Y_1 + \alpha_2 Y_2 + \dots + \alpha_n Y_n, \quad (1)$$

where α_j is the proportion that the j^{th} sample makes up of Z. The α_j are called mixing proportions. These proportions will vary somewhat for each aliquot withdrawn. The magnitude of this variance (σ_c^2) depends on the mixing process, i.e., on how well the n samples have been homogenized and mixed.

- (f) There are two basic approaches that can be taken to estimate σ_c^2 . First, one can assume that "perfect" mixing has occurred (defined in Section 2), and attempt to find an appropriate estimate of σ_c^2 based on statistical distributional assumptions. We suggest that the statistical distributions used thus far for obtaining estimates of σ_c^2 (see Section 4) need to be evaluated and perhaps modified to better meet the needs for transuranic studies. Second, one can make no assumptions about the degree of mixing achieved and conduct special field and laboratory studies to empirically estimate σ_c^2 for a proposed mixing procedure (see Section 5).

Since there appears to be little if any data currently available that were obtained for the specific purpose of estimating σ_c^2 , we suggest that experiments along the lines discussed in Section 5 be undertaken. Such studies should certainly precede any major field sampling effort that contemplates the pooling of samples.

- (g) A large mixing variance could result in a substantial increase to the total variance between aliquots withdrawn from a pooled sample if the medium (soil, vegetation, animal tissue, etc.) has not been reduced to sufficiently small particles that have been thoroughly mixed so that the Pu particles are well mixed throughout the medium. A strict uniform mixing of Pu particles may be impossible to achieve except, e.g. in vegetation samples where the Pu has been completely dissolved and aliquots of the liquid are analyzed. Hence, at least a small contribution to the total variance is likely to result from the mixing process, even if the Pu particles have been well mixed.
- (h) If the entire pooled sample is analyzed as a single sample then the mixing variance clearly becomes irrelevant. Hence, the statistical problems involved in estimating σ_c^2 simply go away.
- (i) The variance formulae derived in the three above mentioned papers are appropriate only under the simplifying assumptions used in their derivation. The validity of these assumptions for each field study should be checked before their formulas are used. New expressions for estimating variances under less restrictive assumptions may need to be derived. See Sections 3 and 6 for discussion.
- (j) Eberhardt (1976) notes that the z statistic given by

$$z = \log_e \bar{x} - \overline{\log_e x}$$

gives a "convenient notion of the effects of pooling or compositing samples". In Section 7 we note that under certain conditions, e^z should be approximately equal to the ratio of the arithmetic mean to the geometric mean. If imperfect mixing has occurred then e^z should deviate from this ratio. An interesting avenue of statistical research would be to obtain the statistical distribution of z or e^z under conditions of perfect as well as imperfect mixing (perhaps using computer simulations). Then it might be possible to base a statistical test on z or e^z to test for adequacy of mixing.

- (k) Kussmaul and Anderson consider the problem of estimating the optimum number of locations, samples within locations, and measurements to take for a given cost function, where a "measurement" is on the entire pooled sample. This approach may be useful in estimating the number of samples that should be pooled for given costs associated with sampling and pooling. However, their model needs to be expanded to take into account measurement error and the eventuality that the entire pooled sample cannot be analyzed.
- (l) A topic not discussed in this issue of TRAN-STAT is how compositing affects common tests of hypotheses such as the "t" test. This is an important aspect that needs attention.

- (m) We note that Rohde lists several references on compositing where the emphasis is on detection rather than estimation. These might be useful with regard to field sampling for cleanup purposes.
- (n) We recently communicated with Dr. Charles Rohde of the Department of Biostatistics, School of Public Health, of Johns Hopkins University, Baltimore, Maryland. He generously sent us rough drafts of two papers he has written, entitled "Sampling Aquatic Environments" and "Batch, Bulk and Composite Sampling". We have not had time to adequately review these papers for discussion here. He lists several new (to me) references on compositing. Two of these are Brown and Robson (1975) and Sobel and Elashoff (1975).

This issue is concluded with a page of notation used by Rohde, and Brown and Fisher for the convenience of those who may want to study and compare their papers more fully (Section 8).

SECTION 1. Compositing Designs

Since Brown and Fisher and Rohde use a different compositing scheme than Kussmaul and Anderson, we illustrate these two approaches in this section. Of course, other compositing designs could also be used. These would require different statistical models than discussed here.

Consider the collection of several samples at each of several locations. The variance formulas developed by Brown and Fisher and Rohde that are given in later sections below are for the case where all samples are pooled. For example, consider the following diagram.

The pooled sample may be analyzed in total, or R aliquots removed and analyzed individually. Either way, the variance components between locations and between samples within locations cannot be estimated from this type of pooled sample.

Kussmaul and Anderson consider a two-stage nested model where compositing does not occur across locations. A simple example is:

This scheme allows the estimation of the variance component between locations and between samples within locations. They assume that the measurement from each pooled sample is the average of the measurements the two individual samples would have provided. This assumption seems reasonable if each of the pooled samples is analyzed in its entirety (ignoring measurement errors). However, if only a portion of each pooled sample is analyzed this assumption would hold only when "perfect-uniform" mixing had occurred (see Section 2 for definition).

SECTION 2. Degree of Mixing: Some Terminology

In this section we propose some terms for the degree of mixing achieved by a given mixing procedure. These definitions are useful for discussion purposes since the magnitude of the mixing variance is related to the degree of mixing.

<u>Degree of Mixing</u>	<u>Terminology</u>	<u>Description</u>
1	"Unmixed"	Samples pooled together with no or little attempt made to mix them. This procedure is acceptable <u>if the entire</u> pooled sample is analyzed or <u>if it is known</u> that all samples have the same concentration (a risky assumption). It is not acceptable if only a portion of the mixed sample is analyzed.

<u>Degree of Mixing</u>	<u>Terminology</u>	<u>Description</u>
2	"Mixed"	Samples pooled and mixed by grinding, kneading, ball-milling, shaking, etc. for some period of time. The adequacy of mixing is usually not evaluated although the assumption may be made that "perfect" mixing (see next definition) has been achieved. σ_c^2 may be large if this assumption is false.
3	"Perfect" mixing	Samples pooled and mixed using methods and mixing times such that continued mixing would achieve no greater homogenization and mixing. This is the situation for which the formulas for σ_c^2 derived by Brown and Fisher, and Rohde apply. σ_c^2 will not be zero but it may be small.
4	"Perfect-Uniform" mixing	Samples pooled and mixed so well that each aliquot withdrawn has <u>exactly</u> the same concentration. This <u>ideal</u> would be achieved, presumably, in the case of vegetation samples that are reduced to a liquid, with aliquots being withdrawn from the liquid. Observed concentrations would still vary between these aliquots, but would be due entirely to counting errors or analytical errors introduced after the aliquots had been removed from the pooled sample (liquid). "Perfect-Uniform" mixing would appear to be unachievable for soil samples pooled together since a "perfect-uniform" mixing would imply that each aliquot contained <u>exactly</u> the same number and size distribution of Pu particles. If the entire pooled soil sample were dissolved to liquid form then we could assume the total liquid could be characterized by "perfect-uniform" mixing.

SECTION 3. Components of Variance and Models

Before the decision is made to composite samples, the sources of variability expected to be present in the resulting data should be identified. This is important because the variance of a mean computed on aliquots withdrawn from composite samples depends on the magnitude of the components of variance associated with these sources of variability.

Example 1: Suppose n soil samples are to be collected at random within a 100 x 100 meter plot. The samples are to be composited and two or more aliquots analyzed for Pu. The several resulting concentrations are averaged to give a mean. We wish to estimate the variance of this mean. A reasonable model for the concentrations of the original samples making up the composite might be

$$Y_j = \mu + e_j, \quad j = 1, 2, \dots, n \quad (2)$$

where

Y_j = concentration of the j^{th} sample going into the composite,

μ = true mean concentration of all possible soil samples that could possibly have been collected from the plot¹, and

e_j = difference between the concentration of the j^{th} soil sample and the true mean μ .

If R aliquots are withdrawn from the composite and we denote the observed concentration of the i^{th} aliquot by Z_i , then the true (unknown) variance of their mean concentration $\bar{Z}$ is shown by Rohde (pp. 276, 277) to be

$$\text{Var}(\bar{Z}) = \frac{\sigma^2}{n} + \frac{n\sigma_c^2}{R}, \quad (3)$$

where

σ^2 = the variance between all possible soil samples (Y_j) in the field plot, and

σ_c^2 = "mixing variance", i.e., the true variance due to less than "perfect-uniform" mixing of the Pu particles.

We note that σ^2 includes the variance due to changes in concentration level over the plot as well as analytical and counting errors.

¹In soil sampling it is reasonable to assume that each soil sample will have a different true mean, i.e. the mean of Y_j is μ_j . Then μ may be considered to be the average of the μ_j 's for all possible samples over the 100 x 100 foot plot.

The form of Equation (3) is a consequence of choosing the particular model in Equation (2). Other models would result in other expressions for the variance of the mean as shown in the next example.

Example 2: Suppose that J "locations" within a 100 x 100 meter plot have been selected at random. Define a "location" to be a unit of land of dimension 1 x 1 meter. Select K soil samples at random within each location for a total of JK samples. A reasonable model might be:

$$Y_{jk} = \mu + t_j + e_{jk} \quad \begin{matrix} j = 1, 2, \dots, J \\ k = 1, 2, \dots, K \end{matrix} \quad (4)$$

where

Y_{jk} = concentration of the k^{th} sample at the j^{th} location within the 100 x 100 meter plot,

μ = true mean concentration of all possible soil samples that could have been collected over the entire plot,

t_j = difference between the "true mean concentration of the j^{th} location"² and the true plot mean μ ,

e_{jk} = difference between the k^{th} observed concentration at the j^{th} location (Y_{jk}) and the true mean concentration of the j^{th} location.

Now, suppose all KJ samples are pooled into a single composite sample and R aliquots withdrawn for analysis. The variance of the mean of the resulting R concentrations Z_i , $i = 1, 2, \dots, R$ is shown by Rohde (pp. 278-279) to be

$$\text{Var}(\bar{Z}) = \frac{\sigma_t^2}{J} + \frac{\sigma^2}{KJ} + \frac{KJ \sigma_c^2 \sigma^2}{R} + \frac{K^2 J(J-1) \sigma_c^2 \sigma_t^2}{R(KJ-1)}, \quad (5)$$

where

σ_t^2 = variance between the true mean concentrations of all possible "locations" that could conceivably have been selected,

σ^2 = variance between all possible samples that could have been collected within a location,² and

σ_c^2 = mixing variance as defined in Example 1 above.

²We assume in this model that σ^2 does not vary from one location to another and that σ^2 includes measurement errors.

Equations (3) and (5) differ because the field sampling procedure and hence the underlying model are different. Once a field sampling procedure and a reasonable model are in mind then formulas for the variance of the mean of aliquots withdrawn from the composite can be obtained (such as Equations (3) and (5)). These formulas will include the components of variance (e.g., σ^2 , σ_t^2 , σ_c^2).

Example 3: This example uses $^{239-240}\text{Pu}$ soil concentration data kindly offered to us by John Pinder (Savannah River Ecology Laboratory). The data are a portion of a larger set contained in the draft manuscript "Sources of Variation in Soil Plutonium Concentrations" by John E. Pinder, III, and Donald Paine (Rockwell International, Richland, WA) that has been submitted for possible inclusion in the DOE document "Transuranic Elements in the Environment" to be edited by W. C. Hansen (LASL). The purpose of the study was to evaluate the relative importance of 3 components of variance: distance from the source, microtopographical heterogeneity in deposition of Pu particles, and sampling error.

Three soil samples were collected at 1 meter intervals along several transects located various distances from the source of Pu (a 62 meter stack that exhausts filtered air from the internal atmosphere of a reprocessing facility on the Savannah River Plant). In this example we restrict our attention to the transect closest to the stack (182 m) and suppose, for the sake of illustration, that the purpose of sampling was to estimate the mean $^{239-240}\text{Pu}$ concentration along the transect.

Suppose the 3 samples collected along the transect are composited, mixed in some fashion, and R aliquots removed for Pu analysis. The mean $\bar{Z}$ is computed and we wish to estimate $\text{Var}(\bar{Z})$. First we need an appropriate model. The model given by Equation (2) is assumed here as being applicable, where $n = 3$. For this model $\text{Var}(\bar{Z})$ is given by Equation (3).

We digress here briefly, to point out that Rohde shows the usual estimate of the variance of the mean, i.e.

$$\frac{1}{R(R-1)} \sum_{i=1}^R (Z_i - \bar{Z})^2, \quad (6)$$

will tend on the average to underestimate the true variance given by Eq. (3). He shows that if the model given by Equation (2) is correct, then expression (6) above estimates $n\sigma_c^2/R$, not $\sigma^2/n + n\sigma_c^2/R$. In general, no matter what the correct model may be, the use of expression (6) to estimate $\text{Var}(\bar{Z})$, where the Z_i are observations on aliquots withdrawn from a composite sample, will tend to underestimate the true variance of the mean. This occurs because of the

correlation between aliquot concentrations withdrawn from the same composite. That such a correlation must exist can be seen from equation (1) since each aliquot concentration Z is a function of the same set of field sample concentrations Y_j , $j = 1, 2, \dots, n$.

Returning to Equation (3), it is clear that we need estimates of σ^2 and σ_c^2 to estimate $\text{Var}(\bar{Z})$. Pinder and Paine estimated σ^2 to be 12.16 on the basis of two 10 gram aliquots of soil drawn from each of the 3 soil samples collected along the transect. Using this estimate of σ^2 and $n = 3$, we see that Equation (3) gives

$$\text{Var}(\bar{Z}) = 4.05 + 36.5 \sigma_c^2 / R. \quad (7)$$

Now, if we suppose the soil is sufficiently well mixed so that the Pu particles are perfectly and uniformly mixed, then $\sigma_c^2 = 0$ and $\text{Var}(\bar{Z}) = 4.05$. It is more realistic to assume that this level of mixing cannot be attained so that $\sigma_c^2 > 0$. Now, $\text{Var}(\bar{Z})$ can be reduced by analyzing a large number (R) of aliquots. This, however, would defeat one purpose of compositing, i.e. to reduce analysis costs. We suppose here that $R = 2$ is a minimum number since $R = 1$ would give no information on the adequacy of mixing. (The optimum choice for R should be determined taking into account the costs of sampling, pooling, and mixing samples, as well as the magnitudes of the variance components). For $R = 2$, $\text{Var}(\bar{Z})$ from Equation (7) becomes

$$\text{Var}(\bar{Z}) = 4.05 + 18.2 \sigma_c^2, \quad (8)$$

and we are faced with estimating σ_c^2 .

There are two basic approaches we might take in estimating σ_c^2 . First, we could assume that perfect mixing has occurred and derive an expression for σ_c^2 based on further assumptions about the statistical distribution of the mixing proportions ($\alpha_1, \alpha_2, \dots, \alpha_n$) and the sample concentrations ($Y_1, Y_2, \dots, Y_n$). Second, we could make no assumptions about the level of mixing achieved and instead, resort to the analysis of special field samples to empirically estimate σ_c^2 . Since little is known at the present time about the adequacy of mixing, we feel this second approach should be the preferred method. Some possible ways to attempt this are discussed below in Section 5.

Concerning the first approach, (which assumes perfect mixing) Brown and Fisher (p. 667) develop an expression for σ_c^2 based on the n -variate hypergeometric distribution where the characteristic being measured is composed of discrete units. Rohde assumes the Dirichlet distribution to be appropriate and obtain a similar expression. These models are discussed in greater detail in Section 4 below. For this example we use Rohde's formula for σ_c^2 , which is given by

$$\sigma_c^2 = (n-1)/n^2(n+1). \quad (9)$$

There is some rationale for considering this expression to be an approximate upper limit to the true σ_c^2 for soil sampling situations (Section 4), although

we are not prepared at this point to defend this assertion. Assuming for the moment that Equation (9) is applicable to the present example (i.e. that perfect mixing has occurred and the Dirichlet distribution is appropriate) we find that $\sigma_c^2 = 0.0556$ since $n = 3$. Then Equation (8) becomes $\text{Var}(\bar{Z}) = 4.05 + 1.01 = 5.06$, a 25% increase over 4.05.

There remains considerable doubt in our minds whether Equation (9) is the best method for estimating, an upper limit on σ_c^2 . We hope that this discussion and that in Section 4 will stimulate some of our readers to investigate the statistical aspects of this problem to resolve this question. Most surely, the reader must not assume that compositing will necessarily increase $\text{Var}(\bar{Z})$ by the 25% obtained in this example. The magnitude of σ_c^2 clearly depends on many variables involved in the compositing and mixing process and could vary greatly from application to application.

Suppose, e.g., that the number of Pu particles in each of the 3 samples collected along the transect is about 1000 rather than 1. Then, using Equation (15) (see Section 4) we find $\sigma_c^2 = 0.000111$ so that $\text{Var}(\bar{Z}) = 4.05 + 0.00202 = 4.052$, a negligible increase over 4.05. (Recall again that we are still assuming perfect mixing when we make this calculation). It is obviously important to know as much as possible about the number and size of Pu particles at their source (the exhaust stack in this example). Since in general we cannot assume perfect mixing has occurred, information should be gathered on the magnitude of σ_c^2 using field samples under various mixing procedures so that the relative merits of compositing can be evaluated (see Section 5).

SECTION 4. Expressions for σ_c^2 Based on Statistical Models When "Perfect" Mixing has Been Achieved.

Brown and Fisher consider an urn model for material composed of discrete units that results in an expression for σ_c^2 . Each sample is assumed to contain the same number (s) of units. Assume the probability of getting x_1 units of sample 1, x_2 units of sample 2, ..., and x_n units of sample n in an aliquot withdrawn from the "perfectly" mixed pooled sample is given by the n -variate hypergeometric distribution. (Note that if the sample has not been "perfectly" mixed, the expressions for σ_c^2 are not correct. In that case the magnitude of σ_c^2 must be estimated from specially designed studies as discussed in Section 5). For this model they show that the mixing variance is given by

$$\sigma_c^2 = \frac{(n-1)(ns-M)}{Mn^2(ns-1)}, \quad (10)$$

where

- n = number of samples pooled together,
- s = number of units in each sample,
- ns = total number of units in the n pooled samples, and
- M = number of units in the aliquot withdrawn from the pooled sample.

If soil samples contaminated with Pu are being pooled, then it seems reasonable that the "units" of interest are Pu particles. Hence, ns is the number of Pu particles in the n pooled samples. If ns is large, a reasonable assumption in many situations, we may replace $ns-1$ by ns . Then Equation (10) becomes

$$\sigma_c^2 = \frac{n-1}{n^2} (1/M-1/ns). \quad (11)$$

When ns is very large this reduces approximately to

$$\sigma_c^2 = (n-1)/n^2 M. \quad (12)$$

There are at least two main difficulties with using Equations (10), (11) or (12) to estimate σ_c^2 when perfect mixing has been achieved. First, the n -variate hypergeometric distribution model used by Brown and Fisher specifies that s is the same for each sample being composited. This will never be strictly true, although for soil samples collected within a small plot it may be approximately true. Secondly, M is assumed to be a known parameter chosen by the experimenter (drawing M balls from an urn) rather than a random variable that changes with each aliquot withdrawn from the pooled sample. One approach around this second problem would be to specify a statistical distribution for the random variable M . Then the mixing proportions would have a compound distribution (see Johnson and Kotz, 1969, Chapter 8) for which σ_c^2 could, presumably, be found.

Rohde assumes that the vector of mixing proportions for the i^{th} aliquot $(\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{in})$ follows a Dirichlet distribution (Wilks, 1962, p. 177-182). Under this assumption the mixing variance is given by

$$\sigma_c^2 = \frac{\ell_j \left(\sum_{j=1}^n \ell_j - 1 \right)}{\left(\sum_{j=1}^n \ell_j \right)^2 \cdot \left(\sum_{j=1}^n \ell_j + 1 \right)}, \quad (13)$$

where the ℓ_j are known positive numbers (parameters).

Rohde does not discuss the physical interpretation of these ℓ_j in a given sampling context. An interpretation is suggested by the formula for the expected (average) value of α_j if the Dirichlet distribution is appropriate. This expected value is

$$E(\alpha_j) = \ell_j / \sum_{j=1}^n \ell_j. \quad (14)$$

If we interpret ℓ_j to be the number of Pu particles in the j^{th} sample, then Equation (14) is just the proportion of total Pu particles in the pooled sample that was contributed by the j^{th} sample. Hence, this interpretation of ℓ_j makes intuitive sense. Unfortunately, we never know the number of

Pu particles in a sample so that Equation (13) cannot be computed.

If we make the same assumption as used by Brown and Fisher, i.e., that all samples have the same number of Pu particles (denoted by s), then Equation (13) reduces to

$$\sigma_c^2 = (ns-1)/[n(ns)(ns + 1)] \quad (15)$$

which was obtained by Rohde (p. 275). (Note that σ_c^2 computed using either Brown and Fisher's or Rohde's formulas (Equations (12) and (15), respectively) becomes very small as the number of Pu particles in the aliquot or soil sample becomes large). If we set $s = 1$, signifying one Pu particle in each of the n samples being pooled, we find that Equation (15) reduces to

$$\sigma_c^2 = (n-1)/[n^2(n+1)] \quad (16)$$

Recall that Equation (16) was used in Example 3 in the previous section. Excluding the situation where one or more samples have no Pu particles, Equation (16) could perhaps be considered an approximate upper limit on the true σ_c^2 if the Dirichlet distribution is appropriate. We have noted in the previous section our need to investigate the validity of this assertion and of the Dirichlet distribution.

Equation (15) suffers from the same difficulties discussed above relative to Equation (10), i.e. the derivation of Equation (15) is based on the assumption that s is a parameter (i.e. fixed), that does not vary between samples, when in fact it is a random variable that does vary from sample to sample. As mentioned above relative to Brown and Fisher's approach, it may be desirable to develop a compound distribution for the vector of mixing proportions by assuming a distribution for s or ns .

SECTION 5. Some Possible Methods for Estimating σ_c^2 , The Mixing Variance When Perfect Mixing is Unlikely to be Achievable.

In this section, methods mentioned by Brown and Fisher (p. 668) and Rohde (p. 278) are discussed.

Method 1: Direct Estimation Using Mixing Experiments

It may be possible in some situations to estimate σ_c^2 for a given mixing procedure by mixing uncontaminated sample material with sample material spiked with a known amount of contaminant. Suppose, e.g., that 100 g of uncontaminated sample material and 100 g of material spiked to have a Pu concentration of 10 nCi/gram are composited and subjected to a proposed mixing procedure. If 10 gram aliquots are then removed from the mixed composite we would expect (assuming adequate mixing has occurred) each aliquot to have a concentration of about 5 nCi/g, i.e.,

$$[\alpha_1 \times (\text{zero nCi/g})] + [\alpha_2 \times 10 \text{ nCi/g}] = 5 \text{ nCi/g.}$$

An estimate of α_2 can be obtained for each 10 g aliquot that is analyzed since

$$\hat{\alpha}_2 = (\text{observed concentration}) / (10 \text{ nCi/g})$$

If the numerator and denominator of this expression are free of measurement error, then the variance of these estimates of α_2 would be an estimate of σ_c^2 for the particular sample type and mixing procedure used. Since measurement error is always present, this estimate of σ_c^2 will be too large on the average and would need to be adjusted to a smaller value if the magnitude of measurement errors is known or can be estimated. Care would need to be taken to make sure that the samples and mixing procedure used to estimate σ_c^2 are really applicable to field samples and to laboratory mixing procedures used in the actual study. It's possible, of course, that σ_c^2 could vary from one location to another in the field, in which case σ_c^2 estimated on the basis of one type of sample, say one type of soil, would not be applicable to another type of soil.

Method 2: Using Prior Knowledge of Components of Variance and Formulas for $\text{Var}(\bar{Z})$.

Estimates of the components of variance of the underlying model (such as in Equations (2) and (4) above) may be available from prior studies for which compositing was not done. If a new study utilizing compositing is undertaken for which these components are applicable, then σ_c^2 can be estimated in the following steps: (1) compute the estimated variance $\text{Var}(\bar{Z})$ of concentrations from aliquots withdrawn from the composite, (2) equate this estimate of $\text{Var}(\bar{Z})$ to the appropriate variance formula (such as Equations (2) and (4) above), and (3) solve for σ_c^2 .

Example: Suppose the model in Equation (2) is appropriate and suppose we have estimated σ^2 (which includes measurement errors) from prior individual samples from the same region. We take new samples, pool them into a composite, and withdraw R aliquots each of which are analyzed yielding observations Z_i , $i = 1, 2, \dots, R$. We compute $\text{Var}(\bar{Z})$. We can then solve

$$\hat{\text{Var}}(\bar{Z}) = \frac{\hat{\sigma}^2}{n} + \frac{n\sigma_c^2}{R}$$

for σ_c^2 , the solution of which is

$$\hat{\sigma}_c^2 = \frac{R}{n} \left[\frac{\hat{\text{Var}}(\bar{Z})}{\hat{\sigma}^2} - \frac{1}{n} \right]$$

The validity of this approach clearly depends on how good an estimate of σ^2 we have available, and whether that estimate, obtained previously in time when field and laboratory conditions were undoubtedly different to some

degree, is really applicable to the new sampling study.

Method 3: Using Current Estimates of Components of Variance and Formulas for $\text{Var}(\bar{Z})$.

This is essentially the same as Method 2 except for the important difference that the components of variance and $\text{Var}(\bar{Z})$ are both estimated from samples collected at the same time and under the same conditions. This approach is much preferred over Method 2 since there is no need to assume the relevance of variance estimates from prior studies.

Example: Suppose the model given by Equation (4) applies. Collect K soil samples at each of J locations. Randomly assign the K samples at each location to one of two groups. Hence each group will have JK/2 samples. Select one group of samples at random and analyze each sample separately in order to estimate the between location variance (σ_L^2) and the within location variance (σ^2) which includes measurement errors. Pool the other group of JK/2 samples into a composite. Mix the composite in the same manner as planned for field studies, withdraw aliquots for analysis, and compute $\text{Var}(\bar{Z})$, the variance of the resulting mean concentration (Equation 6). Set $\text{Var}(\bar{Z})$ equal to Equation (5) and solve for σ_C^2 using the estimates of σ_L^2 and σ^2 obtained from the other group of JK/2 samples.

Method 4: Analyzing Entire Composite Samples as Well as Aliquots from Composite Samples.

This method involves computing the variance between aliquots from a single composite as well as the variance between several entire composites (no subsampling).

Example: Suppose the simple model of Equation (2) is appropriate, i.e. $Y_j = \mu + e_j$. Suppose a sample is collected at each of n random locations within a 100 x 100 meter area, and that this procedure is repeated P + 1 times. To be specific, suppose n = 10 and P = 5. Composite the 10 samples from each repetition to yield P + 1 or 6 pooled samples. Analyze 5 of these 6 composites in their entirety to yield 5 observations Z_i , where

$$Z_i = \sum_{j=1}^{10} \alpha_{ij} Y_j \quad i = 1, 2, 3, 4, 5.$$

Rohde (p. 276) shows that if the model of Equation (2) is appropriate then the variance of Z_i is

$$\frac{\sigma^2}{n} + n\sigma_C^2 \sigma^2, \quad (17)$$

where n = 10 in our example, and σ^2 and σ_C^2 are defined following Equation (3). Hence the computed variance of the five observations

Z_i , i.e.,

$$\frac{1}{5-1} \sum_{i=1}^5 (Z_i - \bar{Z})^2, \quad (18)$$

is an estimate of expression (17).

Now, suppose the 6th composite sample is not analyzed in its entirety. Instead, suppose the composite sample is mixed using the same procedure to be used in future studies, and that $R = 5$ aliquots are withdrawn from the composite sample yielding 5 observations T_i , $i = 1, 2, 3, 4, 5$. Rohde (p. 277-278) shows that

$$\frac{1}{5-1} \sum_{i=1}^5 (T_i - \bar{T})^2 \quad (19)$$

is an estimate of $n\sigma_c^2$ if Equation (2) is appropriate. Hence expression (18) minus expression (19) is an estimate of σ^2/n . This estimate of σ^2 can then be used with expression (19) to estimate σ_c^2 as follows:

$$\hat{\sigma}_c^2 = [\text{expression (19)}]/n\hat{\sigma}^2 \quad (20)$$

The precision of $\hat{\sigma}_c^2$ obtained using Equation (20) will depend on the size of P and R , and on the variability in concentrations between aliquots withdrawn from a composite sample, and that between different composite samples. These variances can be very large in field samples and the distribution of concentrations is usually skewed. Under these conditions it would seem prudent to let both P and R be at least 30 (Snedecor and Cochran 1967, p. 74-76, 89). Considerably more could be required if long-term field studies are contemplated so that the cost of obtaining a more precise estimate of σ_c^2 is justified.

SECTION 6. Assumptions Underlying the Variance Formulas of Brown and Fisher (1972) and Rohde (1976).

Rohde (p. 279, 2nd paragraph) assumes that the set of mixing proportions $(\alpha_1, \alpha_2, \dots, \alpha_n)$ is independent from aliquot to aliquot. This assumption seems reasonable where the number of aliquots (R) withdrawn from the composite sample is a small proportion of the total number that could conceivably be withdrawn. Brown and Fisher (p. 664) do not require this independence, but both papers arrive at essentially the same variance formulas.

An assumption used in both papers is that the mixing proportions are independent of the concentrations Y_j of samples making up the composite. This assumption would be violated, e.g., if those samples with the highest concentrations tend to be systematically under or over-represented in aliquots withdrawn from the composite. For example, if samples of different types of soils were composited, and certain soil types had higher Pu concentrations than other types, it is conceivable that these higher activity soil types would not be ground (by ball milling, say) to the same size soil particle as the low concentration soils. Then the high concentration soils could be systematically excluded (or included) to an extent different than that expected based on their proportion of the composite by weight.

For applications where it is not reasonable to assume the mixing proportions are independent of the concentrations Y_j , variance formulas for $\text{Var}(\bar{Z})$ could be developed, but they would be more complex and would require estimating the magnitude of the presumed correlation. This is an area that needs statistical input.

We have mentioned previously that Brown and Fisher, Rohde, and Kussmaul and Anderson all assume that observations contain no measurement error. The first two papers deal with applications where this assumption is more justified than for transuranic field studies. Kussmaul and Anderson point out that if measurement error is present, their procedures for estimating variance components in a 2 stage nested design will lead to an inflated estimate of the variance component for the second stage. That is, the estimated variance between samples (within locations) will contain measurement error, and hence will overestimate the true variance between samples.

SECTION 7. Comments on a Statistic Related to Compositing of Samples

Eberhardt (1976), Eberhardt, et al. (1976) and Eberhardt and Gilbert (1973) note that a statistic denoted by z may be useful in "accessing the effects of compositing samples". In the notation used above, the z statistic is

$$z = \log_e Z - \overline{\log_e Y} \quad , \quad (21)$$

where $\log_e Z$ is the logarithm of the concentration of an aliquot taken from a composite sample which is made up of independent field samples Y_j , ($j = 1, 2, \dots, n$); and $\overline{\log_e Y}$ is the arithmetic mean of the logarithms of these same n field samples Y_j . Now, using Equation (1) we may write z as

$$z = \log_e \left(\sum_{j=1}^n \alpha_j Y_j \right) - \frac{1}{n} \sum_{j=1}^n \log_e Y_j \quad (22)$$

where $\sum \alpha_j Y_j$ can be interpreted as a weighted mean of the Y 's, the weights being the mixing proportions α_j . If perfect mixing has occurred and each sample being composited has the same number of Pu particles, then each α_j should (on the average) equal $1/n$, in which case $\sum \alpha_j Y_j$ becomes the usual arithmetic mean of the Y 's. Z in Equation (21) could also be the concentration obtained by analyzing the entire composite (no aliquoting) in which case

$\sum_j \alpha_j Y_j$ is exactly the arithmetic mean $\bar{Y}$ since each α_j is exactly $1/n$ (assuming equal numbers of Pu particles in each sample and no measurement error).

Now, concentrations are not usually reported in log scale. Transforming z back to arithmetic scale we obtain

$$e^z = e^{\log Z / e^{\log \bar{Y}}} = Z/GM, \quad (23)$$

where GM stands for the geometric mean of the Y 's, and Z is either $\bar{Y}$ (for the case when the entire aliquot is analyzed) or a weighted mean (when only an aliquot from the composite is analyzed). Hence, e^z is an approximate measure of the ratio between the arithmetic mean and GM of the Y 's. This ratio will always be greater than 1 except for the unlikely case where all Y 's are identical. (We note that an approximation to the variance of e^z is given by Norris, 1938, Equation (11), when the distribution of the Y 's is gamma). The arithmetic mean will be substantially larger than the GM when the data are highly skewed, a frequent occurrence with Pu field data.

The above suggests that if perfect mixing has occurred and each sample has the same number of Pu particles, the single arithmetic value ($Z = \bar{Y}$) obtained from a composite sample will always in practice be larger than the geometric mean of the individual Y 's (i.e. if compositing had not been done).

An approximation of the ratio Z/GM can be obtained if an estimate of the coefficient of variation c (standard deviation divided by the mean) of the Y_j 's is available. Eberhardt and Gilbert (1973) showed that the average magnitude of z is given approximately by $c^2/2$ for samples drawn from log-normal or gamma distributions. Hence, approximations of Z/GM are given by e^z for these distributions. Some examples are:

c :	0.2	0.5	1.0	2.0	3.0
$z \approx c^2/2$:	0.02	0.125	0.50	2.0	4.5
$e^z = Z/GM$:	1.02	1.13	1.65	7.4	9.0

It would be of interest to know the statistical distribution of z under various mixing and compositing schemes. One way of obtaining such information would be computer simulations. From Equation (22) it is clear that the ratio e^z will be affected if the composite is not adequately mixed, i.e., if the α_j systematically deviate from $1/n$. This suggests that there might be some way to use e^z to statistically test for inadequate mixing if c were known. We have not given the matter sufficient thought to offer any specific advice at this point in time.

SECTION 8: Notation

Below we list the mathematical notation used by Brown and Fisher, and Rohde for those who may want to read these papers.

Description	Notation	
	Brown and Fisher	Rohde
Concentration of a sample going into a composite.	X	Y
Number of samples going into a composite at a given location or plot	K	n or n_j
Number of locations or plots for which composites are obtained	J	J
Concentration of an aliquot withdrawn from the composite sample	y	Z or T
Mixing proportion	α	α
Component of variation between samples at a given location or plot	σ_c^2	σ^2
Component of variation between locations or plots	σ_b^2	σ_t^2

REFERENCES

- Brown, G. H. and N. I. Fisher. 1972. Subsampling a Mixture of Sampled Material, Technometrics 14, pp. 663-668.
- Brown, G. H., and D. S. Robson. 1975. "The Estimation of Mixing Proportions by Double Sampling Using Bulk Measurements", Technometrics 17: 119-126.
- Eberhardt, L. L. and R. O. Gilbert. 1973. Gamma and Lognormal Distributions as Models in Studying Food-Chain Kinetics. BNWL-1747, Battelle, Pacific Northwest Laboratories, Richland, WA, 95 pages.
- Eberhardt, L. L. 1976. Sampling for Radionuclides and Other Trace Substances, Radioecology and Energy Resources (C. E. Cushing, Jr., ed.). Dowden, Hutchinson and Ross, Inc. Stroudsburg, PA, pp. 199-208.
- Eberhardt, L. L., R. O. Gilbert, H. L. Hollister, and J. M. Thomas. 1976. Sampling for Contaminants in Ecological Systems. Environmental Science and Technology 10, pp. 917-925.

Johnson, N. L. and S. Kotz. 1969. Discrete Distributions, Houghton Mifflin Company, Boston, 328 pages.

Kussmaul, K. and R. L. Anderson. 1967. "Estimation of Variance Components in Two-Stage Nested Designs With Composite Samples", Technometrics 9, pp. 373-389.

Norris, N. 1938. Some Efficient Measures of Relative Dispersion, Annals of Mathematical Statistics 9, pp. 214-220.

Rohde, C. A. 1976. Composite Sampling. Biometrics 32, pp. 273-282.

Wilks, S. S. 1962. Mathematical Statistics. John Wiley and Sons, New York, 644 pages.

Snedecor, G. W. and W. G. Cochran. 1967. Statistical Methods, Iowa State University Press, Ames, Iowa, 593 pages.

Sobel, M. and R. M. Elashoff. 1975. "Group Testing with a New Goal, Estimation", Biometrika 62: 181-194.