THE EVALUATION OF SAMPLING AND ANALYTICAL VARIATION IN REGIONAL GEOCHEMICAL SURVEYS

Robert G. Garrett and Thomas I. Goss
Geochemistry Section
Resource Geophysics and Geochemistry Div.
Geological Survey of Canada
Ottawa, Ontario
and
Systems Approach Ltd.
350 Sparks Street
Ottawa, Ontario

Published by permission of the Director-General, Geological Survey of Canada, Ottawa

ABSTRACT

An improved method of regional survey data evaluation is presented which is based on inverted nested sampling designs followed by analysis of variance (ANOVA). The proposed method allows the entire survey data set to be evaluated simultaneously, so eliminating the two stage approach formerly used by exploration geochemists. The ANOVA partitions the variability of the data into regional, sampling cell, lake and analytical components for the centre-lake bottom sediment surveys of Canada's National Geochemical Reconnaissance Program. The individual variance components are tested to see if they are significantly different from zero, and when expressed as percentages may be readily used to compare between data sets. Confidence and prediction intervals at the 95% level are developed for means at the regional, sampling cell and lake levels. These may be used to select appropriate data groupings for contour or symbol map preparation and to detect geographic areas of increased heterogeneity in the search for mineral resources.

INTRODUCTION

It is a common requirement of regional geochemical surveys to evaluate the data set in terms of variability at regional, local and analytical levels prior to carrying out any detailed interpretation. Only if a significant proportion of the data variability is at the regional level can one be confident that the substantive variability observed in the data is due to regional, or broad scale, geological and geochemical features and not a consequence of local variability or analytical error.

Analysis of variance has been used in geochemical studies for 20 years, mainly to estimate the magnitude of the various sources of variation in the data, an early landmark paper being that of Krumbein and Slack (1956). From the mid-1960's to the early 1970's a

number of papers were published where analysis of variance methods were applied to regional and exploration geochemical problems; Miesch (1964, 1967), Garrett (1969, 1973) and Michie (1973), Howarth and Lowenstein (1971), and Bolviken and Sinding-Larsen (1973). In many of these studies only a subset of the regional sample sites were sampled in replicate due to considerations of cost, since full replication would have made the surveys prohibitively expensive. This procedure led to two stage analyses, one to determine if local and analytical variability were sufficiently low, and a second to see if the subset of sites sampled in replicate was a valid subset of the whole (Hornbrook and Garrett, 1976). This procedure was far from ideal as, although it did allow certain hypotheses regarding levels of variability to be tested, a full components of variance study was impossible. An example of the application of a full study made possible by total replication is the work of Chork (1977) on stream sediment variability. During the 1970's, sampling design work at the United States Geological Survey in regional geochemical surveys has produced a number of relevant papers; Miesch (1976), Ebens and McNeal (1976), and Tidball and Severson (1976). The latter two papers describe nested sampling designs that are partially replicated or unbalanced, known as staggered designs (Bainbridge, 1963). This design has two desirable features. Firstly, the degrees of freedom for the analysis of variance are more evenly spread over all sampling levels than with a balanced design, hence, more information on the regional variability is gained at the expense of the lowest design stage without increasing the overall size of the survey, and secondly, it is quite easy to administer and analyze.

PREVIOUS STUDY

In 1974, the first lake sediment regional geochemical reconnaissance survey of a systematic National Geochemical Reconnaissance of Canada was undertaken in Saskatchewan by the Geological Survey of Canada (Hornbrook and Garrett, 1976). The basic sample density of 1 sample per 5 sq. miles (13 sq. km) was chosen in order that targets of various preconceived sizes and shapes could be detected at stated confidence levels (Garrett, 1977). The targets in these surveys are not individual mineral occurrences but the geochemical dispersion haloes enclosing one, or more, of them. Once these target haloes, or zones, have been recognized more detailed work may reveal the individual occurrence(s). The major consideration in the design of the sampling plan was to ensure the detection of such mineral exploration targets; a secondary aspect was the preparation of regional geochemical maps for the recognition of major regional geochemical trends.

The basic sampling plan chosen was very similar to a nested stratified design employing random selection procedures. In order to assess the data for the purpose of regional geochemical map preparation, replicate samples were each taken from lakes and from prepared split samples at a 5% frequency. Every block of 20 samples submitted for analysis contained one lake duplicate and one analytical duplicate. However, this sample design did not readily allow full statistical analysis because it did not allow estimation of the variability within the basic sampling unit, i.e. the 5 sq. mile cell. In order to overcome these shortcomings a modified sampling procedure was developed during the winter of 1976/1977 and applied in the 1977 field sampling program.

PRESENT STUDY

To ensure adequate coverage for mineral exploration purposes, all 5 sq. mile cells containing suitable lakes are still sampled. The design modification was made in the manner of replication, while maintaining at least a 5% replicate incidence, by selecting three replicate samples in each block of data for 16 cells. Specifically, once every 16 cells a second lake within a cell was sampled, wherein two samples were drawn from this lake, one of which was split to yield the analytical duplicate. The design is shown in Figure 1 together with the remaining 15 cells also sampled in the data block; the 20th position is reserved for one of

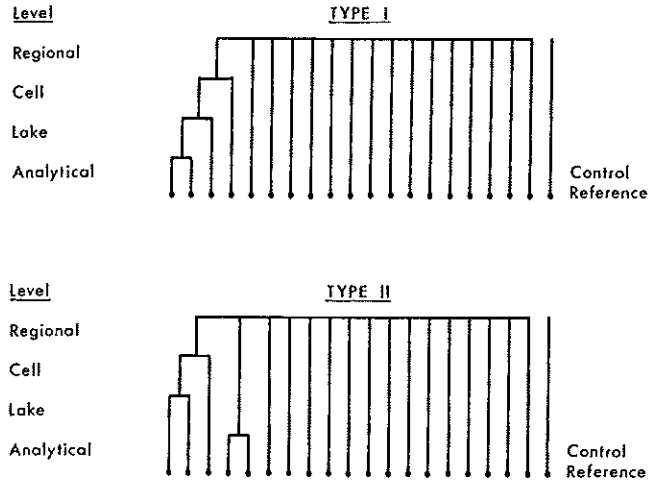


FIGURE 1. Data structures.

several control reference samples used to monitor long-term analytical drift which are not utilized in the subsequent analysis. This sample structure is referred to as Type I. In some cases, insufficient sample material was available to generate the analytical duplicate from one of the lake duplicates. In these instances, which were usually less than 20% of the total blocks in a survey area in the southern Canadian Shield, the analytical duplicate was split from one of the single samples; this led to a Type II sample structure as illustrated in Figure 1. Because of statistical considerations (to be described later), it is planned never to have the Type II structure occur; however, sometimes they cannot be avoided.

In practice, the cell for replicate sampling is chosen in the field by the sampling crew in a statistically haphazard fashion; the location of the control reference sample in the block of 20 is selected by a formal randomization procedure. The blocks of 19 field samples cannot be ground into square or rectangular super-cells as an additional level in the analysis of variance because the samples are collected sequentially along helicopter traverses controlled by the exigencies of sampling (i.e. wind, location of fuel caches, which cells have already been sampled, etc.). Figure 2 illustrates a typical field sampling grid.

The resulting sampling scheme, an unbalanced design akin to the aforementioned staggered designs, is known as an inverted nested design (Bainbridge, 1963). Bainbridge considered inverted sampling designs as statistically efficient (i.e. potentially the best designs to reliably estimate variance components), but knew of no practical example of their application. The application to regional geochemical reconnaissance such as Canada's National Geochemical Reconnaissance is, in fact, a natural choice considering the constraints due to sampling density considerations and the costs of sample collection and analysis. The term "inverted" is due to the larger number of degrees of freedom used to estimate variation at the top levels of the design, the inverse of the usual situation. In the ensuing analysis of variance, the concentration of the degrees of freedom at the highest sampling level will be an asset in determining the variability (i.e., component of variance) more precisely at the regional level.

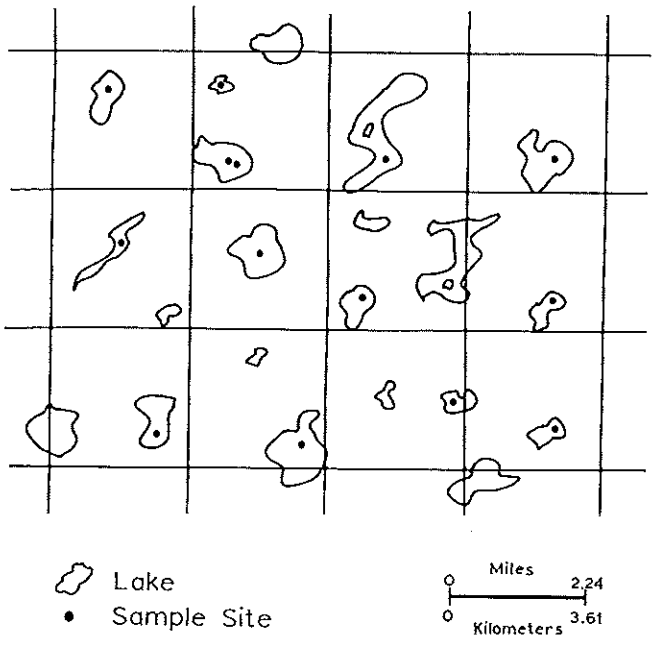


FIGURE 2. Typical field sampling grid.

The objective of the analysis of variance was to determine the components of variance and to test a series of hypotheses regarding the total variability in the data set. This will also allow the partitioning of the data variability into four parts, that at the regional, 5 sq. mile, lake, and analytical levels, which can be expressed as percentages. In addition, it is possible to determine if any of the components are not significantly different from zero; if such is the case, sampling and analysis at that level may be a wasteful exercise and should be further examined in this light—for example, dropping that level may be in order, to maximize overall survey efficiency. Lastly, the partitioned variances can be used to establish confidence and prediction limits on means, the latter, for instance, expressed as a single coefficient usable by the exploration geochemist to estimate the range within which a single future sample would lie. The following three sections describe the statistical methodology and findings.

STATISTICAL MODEL

The general linear model for a 4-level unbalanced nesting design has been previously described by Searle (1971), Graybill (1961), and Anderson and Bancroft (1952) amongst others. Explicitly, it may be expressed as the following:

$$x_{ijkm} = \mu + a_i + b_{j(i)} + c_{k(ij)} + e_{m(ijk)}$$

where

x_{ijkm} represents an individual analysis (i.e., the m^{th} analytical determination on the k^{th} sample from the j^{th} lake of the i^{th} cell)

μ is the true overall mean concentration of the element in all cells, lakes, and potential samples of the region

a_i represents the deviations between the mean value of the i^{th} *cell* and μ, ($i = 1,2,\ldots n$; n = no. of cells)

$b_{j(i)}$ represents the deviations of the mean of the j^{th} *lake* of cell i from the i^{th} cell mean, ($j = 1,2,\ldots, n_i$; n_i = no. of lakes in the i^{th} cell)

$c_{k(ij)}$ represents the deviations of the mean of the k^{th} *sample* of lake j of cell i from the mean of the j^{th} lake of cell i, ($k = 1,2,\ldots, n_{ij}$; n_{ij} = no. of samples in i^{th} cell)

$e_{m(ijk)}$ represents the deviations of the individual analysis or measurement of sample k of lake j of cell i from the mean of the k^{th} sample of lake j of cell i, ($m = 1,2,\ldots, n_{ijk}$; n_{ijk} = no. of analyses from k^{th} sample in j^{th} lake in i^{th} cell).

Note that n_i, n_{ij}, n_{ijk} only take the values 1 or 2, as shown in Figure 1.

All the contributions a_i, $b_{j(i)}$, $c_{k(ij)}$, $e_{m(ijk)}$ are considered random effects (i.e. all the factor classes for cells, lakes, samples, and analyses are assumed to be random samples from a conceptually very large population), each mutually independent and normally distributed random variables with zero mean and constant variance for the corresponding sampling level. Notationally, we have the abbreviated forms:

$$a_i \simeq \text{NID}(o, \sigma_a^2)$$
$$b_j \simeq \text{NID}(o, \sigma_b^2)$$
$$c_k \simeq \text{NID}(o, \sigma_c^2)$$
$$e_m \simeq \text{NID}(o, \sigma_e^2).$$

The e term represents the variability in laboratory measurements, while the remaining random effects are attributed to sampling or geochemical variation. The assumption of normality, with respect to all the random effects, is needed only in formal significance testing of hypotheses, but is not required when calculating the sum of squares, mean squares or variance components.

ANALYSIS OF VARIANCE AND COMPUTATIONAL METHODS

(i) COMPONENTS OF VARIANCE ESTIMATION

Although there are several techniques available for estimating variance components from unbalanced data (Searle, 1971; p. 421), the traditional analysis of variance (ANOVA) method was chosen due to its simplicity (e.g. direct analogue of the ANOVA for balanced data), and its successful and complementary usage in geochemical surveys at the United States Geological Survey.

Generally, ANOVA is used to determine the magnitude of variation associated with each sampling level. By separating the total data variability into mean squares associated with each level and equating the expected or theoretical mean squares under the given model to their observed or calculated values, the resulting set of linear equations are then solved to give the desired estimates of the components of variance. Table 1 summarizes the general 4-level nested design ANOVA calculations. The expected mean squares column illustrates how contributions from variabilities at lower levels accumulate towards the top level.

The calculation of the variance components for the inverted nested design followed the traditional procedure of Ganguli (1941), as described by Anderson and Bancroft (1952), and involved pooling of all sums of squares for each level of sampling for Type I and Type II sampling structures. The estimated variance components—which are simply the "average" estimates of the variation between the x's (the individual analyses of the model outlined in the previous section) at the levels a_i, $b_{j(i)}$, $c_{k(ij)}$, and $e_{m(ijk)}$—provide a basis for interpreting the geochemical trends across the study area, while simultaneously allowing evaluation of the adequacy of the method of laboratory analysis. This information can also be used to design more efficient sampling plans to improve the stability or reproducibility of geochemical maps of the area (Miesch, 1976). Illustrations of these points will be discussed in the next section.

More theoretical work on multi-level (>3) inverted designs is needed, particularly on the use of multiple sampling structures and their effect on design efficiency, variance component estimation, the distributional assumptions, and on the testing of hypotheses. Recent studies of optimum (in the sense of minimum variation of the variance component estimators) combinations of sampling structures for 2 and 3-level nested designs suggest that increasing the number of fundamental structures in a design beyond two or three is *not* the direction to go (Anderson and Crump, 1967; Goldsmith and Gaylor, 1970, pp. 496–497). Type I of our model employs two different sampling patterns while Type II uses three. More-than-one structure designs might be better analyzed (e.g., calculations would involve the "between structure" mean squares) using maximum likelihood estimation of variance components (Anderson, 1975; p. 11). The method of

TABLE 1.

ANALYSIS OF VARIANCE OF GENERAL 4-LEVEL NESTED DESIGN*

Variation Between	Sum of Squares	Degrees of Freedom (DF)	Mean Squares	Expected Mean Squares[2]
Cells (a)	$\sum_{i} n_{i..}(\bar{x}_{i...}-\bar{x}_{....})^2=A$	n-1	A/DF(A)	$\sigma_e^2 + (\sum_{ijk} n_{ijk}^2 f_i)\sigma_c^2 + (\sum_{ij} n_{ij.}^2 f_i)\sigma_b^2 + (\sum_{i} n_{i..}^2 f_i)\sigma_a^2$
Lakes (b) in Cells	$\sum_{ij} n_{ij.}(\bar{x}_{ij..}-\bar{x}_{i...})^2=B$	$\sum_{i} n_i - n$	B/DF(B)	$\sigma_e^2 + (\sum_{ijk} n_{ijk}^2 f_{ij})\sigma_c^2 + (\sum_{ij} n_{ij.}^2 f_{ij})\sigma_b^2$
Samples (c) in lakes	$\sum_{ijk} n_{ijk}(\bar{x}_{ijk.}-\bar{x}_{ij..})^2=C$	$\sum_{ij} n_{ij} - \sum_{i} n_i$	C/DF(C)	$\sigma_e^2 + (\sum_{ijk} n_{ijk}^2 f_{ijk})\sigma_c^2$
Analyses (e) in Samples	$\sum_{ijkm} (x_{ijkm}-\bar{x}_{ijk.})^2=E$	$\sum_{ijk} n_{ijk} - \sum_{ij} n_{ij}$	E/DF(E)	σ_e^2
Total	$\sum_{ijkm} (x_{ijkm}-\bar{x}_{....})^2$	$\sum_{ijk} n_{ijk} - 1$		

[1] Notation is as follows:
summation limits for subscripts (i,j,k,m) are 1 to $(n, n_{i.}, n_{ij}, n_{ijk})$ respectively;
the dot notation is used to indicate summation over an index:

$n_{...} = \sum_{ijk} n_{ijk}$ = total no. of analyses

$n_{i..} = \sum_{jk} n_{ijk}$ = total no. of analyses in i^{th} cell

$n_{ij.} = \sum_{k} n_{ijk}$ = total no. of analyses in j^{th} lake in i^{th} cell

$\bar{x}_{....} = \sum_{ijkm} x_{ijkm}/n_{...}$ = overall element mean (estimate of μ)

etc.

VARIANCE COMPONENTS $(\sigma_a^2, \sigma_b^2, \sigma_c^2, \sigma_e^2)$ ESTIMATION

- equate the "Mean Squares" and "Expected Mean Squares" columns then solve the linear equations

e.g., $s_e^2 = E/DF(E)$

$s_c^2 = (C/DF(C) - s_e^2)/(\sum_{ijk} n_{ijk}^2 f_{ijk})$

etc.

[2] $f_i = (1/n_{i..} - 1/n_{...})/DF(A)$; $f_{ij} = (1/n_{ij.} - 1/n_{i..})/DF(B)$; $f_{ijk} = (1/n_{ijk} - 1/n_{ij.})/DF(C)$
Note: these formulas assume random effects model for infinite populations, see text.

pooling used in the ANOVA approach avoids such complex methods at the expense of losing some information about the design effects. Swallow and Searle (1978) present an alternative method of variance component estimation for a two-level unbalanced design. Cummings and Gaylor (1974) study the ramifications of mixed data structures of three-level unbalanced designs on the testing of hypotheses on variance components. An empirical study of the sampling distributions of variance components for an inverted 4-level design (Leone *et al,* 1968) with four replicated sampling structures showed quite reasonable and stable distributions of the variance estimates under normality conditions.

As noted earlier, the inverted design uses most of the degrees of freedom in estimating σ_a^2, the regional component of variation. This leads to narrow confidence interval bounds for interval estimation of σ_a^2. However, even where the design is balanced, the confidence intervals for variance components (except for σ_e^2) are only approximate (Anderson and Bancroft, 1952, p. 321; and Boardman, 1974, both describe various methods). The methods of computation, employed as second approximations to the unbalanced situation, are unproven and their discussion will be postponed until further research leads to more exact methods.

(ii) STATISTICAL TESTS OF SIGNIFICANCE)

To complement the calculation of variance component estimates, appropriate F-tests may be constructed to test whether or not these estimates are significantly different from zero. For the balanced situation, in general, if σ_x^2 is the variance component association with the x^{th} level of a nested design, where the levels are numbered in ascending order beginning at the analytical level. Then testing the hypothesis $\sigma_x^2=0$ consists of forming the ratio of the mean squares (MS) for the x^{th} and $(x-1)^{th}$ levels (that is, $F=MS_x/MS_{x-1}$ which has an F distribution with x and $(x-1)$ degrees of freedom), and secondly, comparing this with the critical value of $F(x, x-1)$ at the desired significance (α). This exact test results in rejection, or acceptance, of the hypothesis $\sigma_x^2=0$.

However, in the unbalanced case only approximate F tests are available. Of the tests possible, synthesizing the mean squares for the denominator of the F-test (denoted the "error mean squares") and applying Satterthwaite's (1946) formula to calculate the corresponding degrees of freedom seems most appropriate (Tietjen and Moore, 1968). For example, to test the hypothesis $\alpha_a^2=0$, the numerator of the approximate F-test is given by the observed mean square at level *a*, which has $(n-1)$ degrees of freedom. The denominator is formed from the linear combination of

expected mean squares given in Table 1 for level a by setting σ_a^2 to zero and substituting the variance component estimates s_i^2 for σ_i^2 (i=e,c,b) respectively. Equivalently, the synthesized denominator (L) is expressed as $L = r_1MS_e+r_2MS_c+r_3MS_b$, where the s_i^2 are now expressed in terms of the observed mean squares with the r's the resulting coefficients. The degrees of freedom for L are calculated as $L^2/(\Sigma_i(r_iMS_i)^2/DF_i)$, where DF_i are the degrees of freedom associated with the i^{th} observed mean square (MS_i). Finally, the hypothesis $\sigma_a^2=0$ is tested by comparing the ratio MS_a/L with the critical value of F(n−1, DF(L)) at the desired significance. Snee (1974), besides providing a tutorial discussion of the above using the more efficient matrix notation, summarizes some of the properties of this procedure; it is generally agreed the approximation is good.

(iii) USE OF LOGARITHMIC TRANSFORMATION

Prior to calculation all data were logarithmically transformed. This was carried out to eliminate (or at least minimize) the relationship between the mean and variance in the data; incidentally, the date for uranium in all survey areas spanned at least 2 orders of magnitude—uranium is the element used to demonstrate the application of the method. Independence of the variance and the mean is a basic assumption of analysis of variance and the logarithmic transform is well established as a means towards satisfying this assumption (Bartlett, 1947; Cochran, 1947). This transform has two additional benefits; firstly, it satisfies the general notion that large sets of exploration geochemical data appear to be lognormally distributed, and secondly, that geochemists tend to think in terms of ratios rather than absolute differences. Its many uses in treating geochemical data are well summarized in Miesch (1976). For a fundamental paper on the theory of error in geochemical data, see Miesch (1967).

(iv) EXAMPLE)

A computer program was built to calculate the survey area statistics (variance components, F-tests, etc.) using the preceding methodology for the 4-level inverted ANOVA model. For reasons related to the computer program, incomplete blocks of data (less than 1% of collected data), which occurred in the final stages of sampling within each NTS topographic map sheet, were rejected. As an illustration of the pertinent computer output, Table 2 gives the computations for uranium in the Northwestern Ontario survey area. Most of the table information is self-explanatory. The unit size column gives the total number of sampling units at each level in the design. The variance component (%) column of Table 1 gives the percentages of the total variance in uranium estimated to at the regional, cell, lake and analytical levels. For instance, 61% of the total variation is attributable to the regional component (σ_a^2), while only 3% occurs between samples from within lakes and between analyses of the same sample. About 36% of the variance is between lakes within the same 5 sq. mile cell. All variance components that can be tested (σ_a^2, σ_b^2, σ_c^2) are significantly different than zero. Further interpretations are discussed below along with other survey results.

(v) SUMMARY STATISTICS

A number of useful summary statistics were calculated for each survey area. The formulae are summarized below; interpretational insights and conclusions are discussed in the next section.

The *approximate 95% confidence interval for the*

TABLE 2

ANALYSIS OF VARIANCE TO DETERMINE SIGNIFICANCE OF VARIANCE COMPONENTS FOR URANIUM IN THE NORTHWESTERN ONTARIO AREA

Variation Between	Sums of Squares	df	Mean Squares	Unit Size	Variance Component	V.C. %	Error Mean Squares	Error df	Approx F	Signif
Cells	320.73478	1679	0.19103	1680	0.101481	61.05	0.07054[1]	108.47[1]	2.71	>.999
Lakes in Cells	9.73124	105	0.09268	1785	0.059743	35.94	0.00525[1]	118.55[1]	17.67	>.999
Samples in Lakes	0.57682	105	0.00549	1890	0.001907	1.15	0.00309	105.00	1.78	.998
Analyses in Samples	0.32447	105	0.00309	1995	0.003090	1.86				
Total	331.36731	1994			0.166221	100.00				

[1]Error mean square sythesized and degrees of freedom computed by Satterthwaites (1946) formula.

true overall or regional element mean (μ) is given as: (see Table 1 for notation)

$$x_{....} \pm t_{(.025,n-1)} \sqrt{\frac{(\sum_i n_{i..}^2)s_a^2 + (\sum_{ij} n_{ij}^2)s_b^2 + (\sum_{ijk} n_{ijk}^2)s_c^2 + n_{...}s_e^2}{n_{...}^2}}$$

where $t_{(.025,n-1)}$ is the value of a normal deviate of Student's t table corresponding to .025 confidence probability with $(n-1)$ degrees of freedom. The formula for the degrees of freedom is an approximation based on the suggestion of Anderson and Brancroft (1952). In effect, this interval estimate is probably conservative, but, as will, be seen, quite adequate for our purposes.

A useful empirical *variance ratio* (v), that may be used to evaluate the "effectiveness" of the sampling design is given as:

$$v = s_a^2 / (s_b^2 + s_c^2 + s_e^2)$$

If v is $>>1$, no additional sampling or analytical effort is usually required to describe the compositional differences among cells at the regional level.

Another similar variance ratio, v_m, has been proposed as a measure of "map stability" by Miesch (1976); several applications have been made to balanced and unbalanced sampling designs, e.g. Tidball and Severson (1976), and Ebens and McNeal (1976). However, it has yet to be applied to inverted designs and research is underway as to this ratios' applicability in the inverted case.

In exploration it is often desirable to set limits within which the true value will fall from small groups of data. As the initial data have been transformed to logarithms the traditional confidence intervals, which would have to be additively applied to logarithmic values, have been converted for convenience to confidence factors that are applied multiplicatively to the original data.

To determine the 95% *confidence factor* (CF) at the cell (b) and lake (c) sampling levels, we compute:

$$CF_b = \log_{10}^{-1}\left[t_{(.025,DF(B))}\sqrt{s_b^2 + \frac{s_c^2}{1.8} + \frac{s_e^2}{3.0}}\right]$$

$$CF_c = \log_{10}^{-1}\left[t_{(.025,DF(C))}\sqrt{s_c^2 + \frac{s_e^2}{2.0}}\right]$$

The formulae are derived, from Leone *et al's* (1968) formulation of variances for level means in the Appendix.

Similarly the predictability factors can be computed so that the limits can be determined within which a future single new sample of a group of cells, or lakes, can be expected to lie (Hahn, 1970).

To determine the 95% *predictability factor* (PF) at the cell (b) and lake (c) sampling levels, we compute:

$$PF_b = \log_{10}^{-1}\left[CF_b\sqrt{1 + \frac{1}{DF(B)}}\right]$$

$$PF_c = \log_{10}^{-1}\left[CF_c\sqrt{1 + \frac{1}{DF(C)}}\right]$$

The degrees of freedom used in the CF and PF calculations are approximate but conservative. The assumption of normally distributed observations (or the equivalent under an appropriate transformation) must be met approximately for these factors, CF and PF, to be valid estimates. Since it is assumed that the replicated cells represent a random sample from the survey area the sample formulations are used directly as estimates for the survey area as a whole.

DESCRIPTION OF RESULTS

National Geochemical Reconnaissance surveys were undertaken in four areas of the Canadian Shield during the 1977 field season, Figure 3. The four areas are in Labrador (L) where 28,000 sq. miles were covered; the Melville Peninsula (M) of the District of Franklin in the Northwest Territories, 11,500 sq. miles; Northwestern Ontario (O), 11,000 sq. miles; and Northeastern Saskatchewan (S), 4,600 sq. miles. The first three of these areas were covered at a density of 1 sample per 2.41 sq. miles; the last, Saskatchewan, was experimentally covered at a higher sampling density of 1 sample per 2.41 sq. miles. The reason for the higher sampling density was both to better define targets related to mineral exploration and to detect smaller targets. For example, at 1 sample per 5 sq. miles an elliptical target 5 times longer than wide must be 6 miles long to be found at the 95% confidence level; using 1 sample per 2.41 sq. miles, the ellipse need only be 4.2 miles long to be found with the same confidence (Garrett, 1977).

The four survey areas cover a range of physical and geological environments on the Canadian Shield. The Labrador area is of mixed high and low topographic relief and largely south of the limit of discontinuous

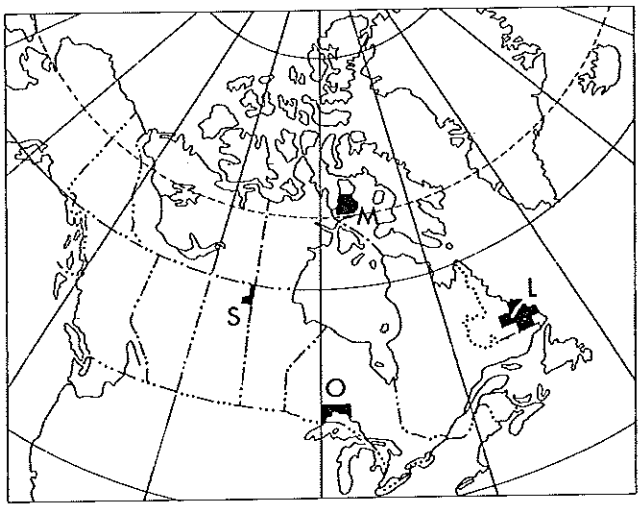


FIGURE 3. Location of survey areas.

permafrost, the high relief is found near the coast in the north of the survey area. In that same northern part the geology is more complex, the area being underlain by rocks of the Churchill, Grenville and Nutak structural provinces, whereas the southern part is underlain only by rocks of the Grenville province. The Melville area is permafrosted barren-ground country of mixed high and low relief well north of the tree line and is underlain by Churchill province rocks which here include some greenstones and sedimentary rocks of the Foxe fold belt. The Ontario survey area is in typical southern Canadian boreal forest south of any permafrost. The area is underlain by rocks of the Superior and Southern provinces, the latter being represented by units of the Port Arthur belted plain, the Nipigon and Lake Superior basins.The topography of the area is variable, and like the Labrador area ranges from rugged close to the "sea," Lake Superior, to gentle in the interior. Lastly, the Saskatchewan area is one of medium to gentle relief covered by boreal forest and just within the zone of discontinuous but widespread permafrost. The area is underlain by Churchill rocks again and includes parts of the Wollaston fold belt, a known uranium producing belt.

In order to assess the impact of mixing Types I and II structures a series of analyses of variance were carried out where only Type I data were used, and then again where Types I and II data were combined. The results of the components of variance study for uranium are shown in Table 3. In all instances, the variance components are significantly different from zero (99% level) and therefore, each component from the regional, cell and lake sampling levels contributes a significant amount to the total variability, the Total $Log_{10}S^2$ column indicates the total variance partitioned, c.f. Table 2. In all areas, except Melville, the differences between the Types I and I + II data sets are not considered to be geologically significant as a similar geological interpretation would be drawn from either data set. The Melville data are mostly data blocks of Type II structure. However, the general pattern of the results agree well with those of the other areas in spite of this fact. The types I + II Melville data are accepted for comparison with the other survey data sets.

The data in Table 3 can be graphically plotted to facilitate overall comparisons (Figure 4). The data sets exhibit generally similar characteristics, and in some instances, their characteristics are extremely similar. In all areas, except Melville, 3% or less of the total variability of the respective data sets is accounted for at the sampling within lake and analytical levels. This indicates that a very small amount of the total variability in the data is due to these sources of variation. At the same time, it confirms the point commonly made by geochemists who have undertaken studies along similar lines that analytical variation is rarely an important contributor to total variation and that the analytical techniques are adequate. For the 12 or so elements other than uranium, routinely analysed in the National Geochemical Reconnaissance, this is not always the case and the methods of ANOVA are used to identify those for which there are weaknesses in the analytical method relative to the problem under

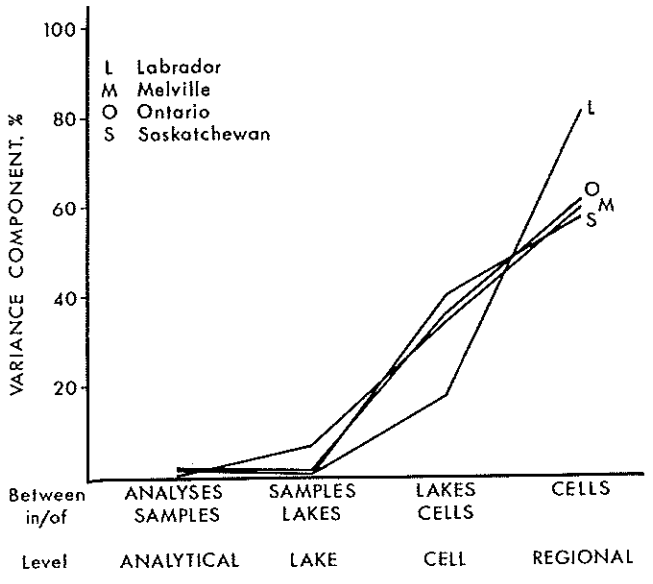


FIGURE 4. Components of variance plot for main survey areas.

TABLE 3
COMPARISON OF TYPE I AND TYPES I + II DATA FOR URANIUM

Area	Data Type	No. of Blocks	Total $Log_{10}S^2$	Variance Components as Percentages P_a	P_b*	P_c	P_e
Labrador	I	225	.308734	80.3	17.4	0.9	1.4
	I + II	239	.299526	80.1	17.6	1.0	1.3
Melville	I	14	.078902	79.9	13.8	6.2	0.1
	I + II	120	.095715	59.5	34.1	6.3	0.1
Ontario	I	82	.161095	57.5	39.2	1.7	1.6
	I + II	105	.166221	61.1	35.9	1.1	1.9
Saskatchewan	I	101	.166681	54.4	42.7	1.0	1.9
	I + II	115	.169208	57.6	39.9	0.8	1.7

*NOTE: Area of cell is 5 sq. miles for all areas except Saskatchewan which is 2.41 sq. miles.

study. In the instance of the Melville uranium data, a pronounced shift occurs with a six-fold increase in variation at the sampling within lake level, and a considerable drop at the analytical level. It is proposed that this is due to the lesser amounts of organic matter in the Melville lake sediments, with a commensurate increase in clastic components. Centre-lake bottom sediments are notably homogenous in the part of the Canadian Shield south of the tree line as a large part of their bulk is made up of organic precipitates and gels which are chemically similar across a centre-lake basin. In contrast, the Melville sediments are inherently less homogenous as a smaller proportion of their bulk is made up of chemically precipitated organic material.

The generally low variance at the sampling within lake level deserves some additional comment. The lake replicates are taken some 100 feet apart, in fact controlled by the distance drifted by the helicopter across the lake surface during sampling. As such the lake replicates may be likened to the "outcrop" replicates of a rock sampling program. In this respect the variance measured in this study may underestimate the true within lake variance. The result of this would be the inclusion of this missing variance at the cell level. Avoidance of this possible difficulty would be difficult and impractical due to the variable geometry of centre-lake basins and time possibly wasted in making several attempts to collect a second truly random centre-lake bottom sediment in smaller lakes.

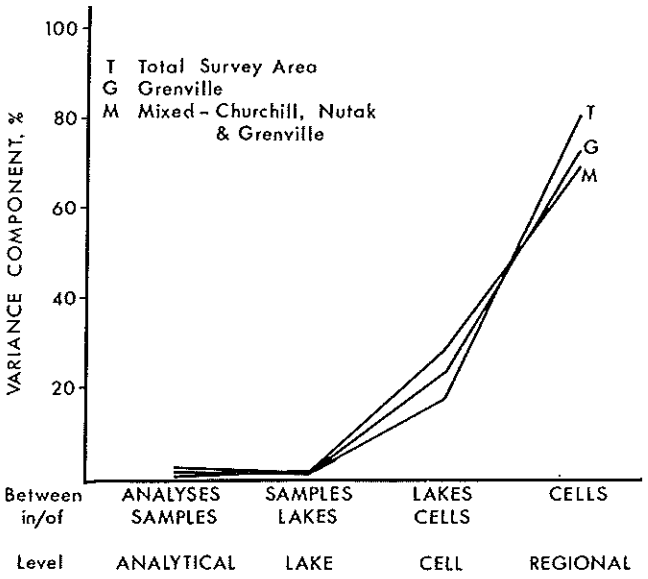


FIGURE 5. Components of variance plot for Labrador area.

The data sets may be split into two groups on the basis of variance at the cell and regional levels: Melville, Ontario and Saskatchewan on one hand, and Labrador on the other. In the former group the variation at the within cell level ranges from 34 to 40% and from 57 to 61% at the regional level. This consistency is somewhat surprising in view of the fact that two different sampling cell sizes were employed, 5 and 2.41 sq. miles. The inference is that on a relative scale

TABLE 4
COMPARISON OF URANIUM RESULTS IN LABRADOR SURVEY WITH SUB-AREAS

Area	No. of Blocks	Total $Log_{10}S^2$	Variance Components as Percentages P_a	P_b	P_c	P_e	$g\bar{x}$	95% bounds
Total Labrador Survey	239	.299526	80.1	17.6	1.0	1.3	1.52	1.45–1.59
Mixed Area	48	.223174	69.7	28.9	1.1	0.3	5.86	5.37–6.40
Grenville Province	191	.210745	72.6	23.7	1.4	2.3	1.08	1.04–1.13

NOTE: $g\bar{x}$ is geometric mean in ppm.

there is as much variation over 2.41 sq. miles of the Saskatchewan area as there is over 5 sq. miles of the Melville or Ontario areas. The Labrador data are characterized by only 18% of the variability at the 5 sq. mile cell level and 80% at the regional level. A detailed inspection of the Labrador data indicated that higher uranium values generally occurred north of the Grenville Front and therefore the data were divided into two sets, one consisting of the area underlain by Grenville rocks, and the other, to the north, underlain by a mixture of rocks from the Churchill, Nutak and Grenville provinces which contain the known uranium occurrences of the survey area. The percentage components of variance, regional means and their confidence bounds are given in Table 4 and presented graphically in Figure 5. The two data subsets, identified as Mixed area and Grenville province, are still not similar to the other three main survey areas but are intermediate between them and the total Labrador data set. The variability at the regional level has dropped by some 10% and has increased at the sampling cell level by 6 to 11%. The geometric means and their confidence bounds reveal that the elevated regional feature to the north of the Grenville Front in the Mixed area, relative to the Grenville province area, is both geochemically and statistically significant. It seems that the initial high percentage of the variation observed at the regional level in the Labrador data set was due in part to variation equivalent to the difference in the means between the two sub-areas being included at the regional level. In analysis of variance terms, if a more detailed study of the Labrador data was to be undertaken a mixed model with a fixed effect included to describe this difference in means would be appropriate. By treating the entire Labrador data set as a random effects model this "fixed effect" variation due to the geochemically distinct sub-areas has been accounted for at the regional level.

The partition of variances at the regional and cell levels so that regional variability is up to 2 to 3 times cell variability confirms some general observations on Shield geology. The inference from the results has to be that local primary and secondary geochemical environments are a half or a third as diverse as those across entire regions. However, it may also be that changes in secondary environment are more important at the cell level, whereas on a regional scale they are masked by geological variation. This does not seem unreasonable in the light of field experience which reveals the extremely heterogeneous, but limited lithological diversity of large areas of Shield rocks—a continuous variation on a limited theme. The implication from the analysis is that if one desired to make detailed geochemical maps in order to reflect bedrock geological features by the sampling of all lakes some 97% of all the variability would be at the between the lakes level. Such a program would be total overkill from the viewpoint of reconnaissance mineral exploration which utilizes the geochemical dispersion from locally anomalous, in terms of bedrock background geochemistry, sources into broad haloes.

As a result of the ANOVAs and the estimation of the components of variance a number of additional steps may be taken to compute empirical variance ratios and other factors. Table 5 contains one such ratio, v, and two factors to indicate confidence and prediction, CF and PF, for both cells and lakes. It should be remembered that the number of blocks refers to the number of times the data structure, and thus replication at all levels, is repeated in a survey.

In all the survey areas and sub-areas, v is in excess of unity which indicates that the sampling program is efficient (Miesch, 1976). In instances where v is low, one procedure that can be carried out is to increase the number of samples collected at lower levels in the design; in effect, the goal being the reduction of the variance terms in the denominator. Often geochemical data are contoured at some increased scale to provide regional trend information in page-size format. The use of computer programs which grid the irregularly spaced data prior to contouring is common. These gridding algorithms generate values from predefined

TABLE 5

EMPIRICAL VARIANCE RATIO AND 95% CONFIDENCE AND PREDICTABILITY FACTORS FOR URANIUM

Area	No. of Blocks	V	Cell Level		Lake Level	
			CF_b	PF_b	CF_c	PF_c
Labrador	239	4.02	2.92	2.93	1.37	1.37
Mixed	48	2.30	3.29	3.33	1.28	1.28
Grenville	191	2.66	2.85	2.86	1.39	1.40
Melville	120	1.47	2.38	2.39	1.43	1.43
Ontario	105	1.57	3.11	3.13	1.31	1.31
Saskatchewan	115	1.36	3.33	3.34	1.28	1.28

numbers of nearest neighbours at the sacrifice of resolution in the final map. In cases where v is low, it would be advisable to use larger numbers of nearest neighbours. When v is high, unless it is desired to produce a "smoothed" map, the number of nearest nieghbours can be kept to a minimum.

The ratio v described above allow statements to be made concerning the data as a whole; however, in exploration it is often necessary to study phenomena on a local scale. To aid in this type of study two factors have been used, the confidence factor (CF) and the predictability factor (PF). Factors are used as they evolve naturally from the treatment of logarithmically transformed data and they conform to the geochemists use of ratios. The confidence factors establish limits, here at the 95% level, that can be set up around any mean value for a small group of data by multiplication or division to define limits within which the true value should lie. The predicability factor similarly sets up limits within which a future single resample of a cell, or lake, can be expected to lie. With large data sets and small values of the factors it will be noted that there is little or no numeric difference between confidence and predictability factors (Table 5), However, the predictability factor is always larger than the confidence factor and so sets up wider limits.

The four main survey area confidence factors for lakes cluster close to 1.3, and for cells close to 2.9. The small range of the lake factors is of particular note considering the wide range of climatic and physical environments the surveys covered. The range of the lake confidence factors is only 0.15. The range for cell factors is larger, 0.41, and this is due to the varying geological environments and changes in sample density. Saskatchewan, with the smaller 2.41 sq. mile sample density, in fact, exhibits the highest cell confidence factor, even though the percentage of variability at the cell level is very similar between Melville, Ontario and Saskatchewan. However, in detail, the Saskatchewan data does have the highest percentage variability at the cell level and also the highest Total $\log_{10}$ Variance (see Table 3, Type I + II) of the three areas, these compound to give the observed result. Confidence or predictability factors can be used to obtain absolute measures, in comparison to the relative nature of v. The lake confidence factor (1.3) implies that with a local mean value of say, 10 ppm uranium, the true value can be expected to lie between 7.7 ppm and 13 ppm at the 95% level. With a cell mean of 10 ppm, representative of several 5 or 2.41 sq. mile sampling cells, and the confidence factor of 2.9 the true cell mean can be expected to lie between 3.4 ppm and 29 ppm. The lake limits do not seem out of order; however, the cell limits may seem large. However, many of the 5 or 2.41 sq. mile cells of Shield terrane cover a multitude of geological and secondary environments; in the light of this fact, the limits do not seem overly wide as the scale of lithological change is often far less than 1.2 miles, which is the average distance between two samples collected randomly within a 5 sq. mile square.

The major use of these factors is in aiding the choice of contouring or symbol intervals for the depiction of regional geochemical data. A regional geochemical map should reflect regional features and not local noise. In view of this, it may be concluded that plotting individual numerical values can be misleading as they can lead to an overconfidence in the data and potential over-interpretation, they do not reflect the fringes of grey that surround the numbers in reality. To aid graphic impact the maps of the National Geochemical Reconnaissance the data are depicted as symbols on a pseudo-logarithmic scale of base 10, i.e. 1, 2, 5, 10 etc. This scale somewhat over-smoothes for lakes but under-smooths for cells. Also, the intervals for contour maps should be chosen using the confidence factors as guides. Similar 1, 2, 5, 10 etc. intervals have been used on 1:1,000,000 scale maps produced by Hornbrook and Garrett (1976).

The use of predictability limits may be beneficial in follow-up studies. The prediction factors in any of the major survey areas are means based on at least 100 replications; therefore, they reflect average variability levels across the survey area and are only slightly, if at all, modified by the rare anomalous effects caused by mineral occurrences. Thus, if on resampling and analysis of a lake, or a cell, the new value falls outside the predicted limits, then it may be concluded that some abnormal source of local variation is effecting the sample. This variation could be due to a variety of heterogeneous geological and/or secondary environmental causes: nonetheless, one geological cause could be the presence of mineral occurrences and therefore the source of the locally high variability should be sought and identified.

CONCLUSIONS

It has been demonstrated that the earlier two-stage analyses of variance used in many regional geochemical surveys can be replaced by the single analysis of an inverted nested design intuitively chosen by geochemists on logistical grounds. This design has proved to be relatively easy to administer, reasonably

efficient statistically, and useful for generating empirical ratios and factors. The components of variance in the examples of uranium in centre-lake bottom sediments are all significantly different from zero at the 99% confidence level. Therefore no single level of sampling can be simply dismissed in order to save survey costs. However, if such an economy move was necessitated it is clear from the data presented that replication of samples within lakes is the least critical. The analytical replication cannot be sacrificed because it is used in the National Geochemical Reconnaissance to monitor analytical contractor performance.

The components of variance derived from four widely spaced geologically and environmentally dissimilar areas have provided data unavailable before the 1977 field season on average sampling variability across large regions at the sample cell level of 5 and 2.41 sq. miles. This variability comprises some 20% to 40% of the total variability. In general, variability within lakes and analytical variability account for less than 3%, and the approximate remaining 60% to 70% of the data variability is at the regional level, between the sampling cells.

The results of the components of variance study, coupled with the empirical ratio v, indicate that the National Geochemical Reconnaissance maps provide an efficient and fairly stable measure of the regional variability of uranium across the survey areas. The confidence factors at the cell and lake levels indicate the magnitude of uncertainty relating to survey observations and can be used in the objective selection of grouped data intervals for symbolic or contoured presentation of the data. Finally the use of prediction intervals is proposed as a useful adjunct to regional or local thresholds in the search for areas of increased heterogeneity which are often the geochemical reflection of belts of mineral occurrences.

ACKNOWLEDGEMENTS

The authors would like to express their thanks to their co-workers at the Geological Survey of Canada, especially E. M. Cameron and E. H. W. Hornbrook for their interest, encouragement and suggestions; and at Systems Approach Limited, particularly R. F. Stuart, T. Gucu and L. M. Campeau who carried out the computer program development. The senior author would like to acknowledge with thanks the time spent by Dr. A. T. Miesch of the USGS in discussing the embryo ideas behind this work during a workshop he held at the University of Calgary in the early fall of 1976. Lastly, and not least, the authors would like to express their thanks to Zita LeBlanc for her care and patience in preparing the manuscript.

REFERENCES

Anderson, R. L., 1975. Designs and estimators for variance components. *In* A Survey of Statistical Design and Linear Models, Ed. J. N. Srivastava, North-Holland Publishing Co., pp. 1–29.

Anderson, R. L. and Bancroft, T. A., 1952. Statistical theory in research. McGraw-Hill Book Co., New York, 399 p.

Anderson, R. L. and Crump, P. P., 1967. Comparisons of designs and estimation procedures for estimating parameters in a two-stage nested process. Technometrics, V. 9, No. 4, pp. 499–516.

Bainbridge, T. R., 1963. Staggered, nested designs for estimating variance components. Am. Soc. Quality Control. Ann. Conf. Trans. pp. 93–103.

Bartlett, M. S., 1947. The use of transformations. Biometrics, V. 3, No. 1, pp. 39–52.

Boardman, T. J., 1974. Confidence intervals for variance components—a comparative Monte Carlo study. Biometrics, V. 30, No. 2, pp. 251–262.

Bolviken, B. and Sinding-Larsen, R., 1973. Total error and ther criteria in the interpretation of stream sediment data. *In* Geochemical Exploration 1972. Ed. J. M. Jones. I.M.M. (London), pp. 285–295.

Chork, C. Y., 1977. Season, sampling and analytical variations in stream sediment surveys. Jour. Geochem. Expl., V. 7, No. 1, pp. 31–47.

Cochrane, W. G., 1947. Some consequences when the assumptions for analysis of variance are not satisfied. Biometrics, V. 3, No. 1, pp. 22–38.

Cummings, W. B. and Gaylor, D. W., 1974. Variance component testing in unbalanced nested designs. Jour. Am. Stat. Assoc., V. 69, No. 347, pp. 765–771.

Ebens, R. J. and McNeal, J. M., 1976. Geochemistry of the Fort Union Formation. U.S. Geol. Surv. Open File Rept. 76-729, pp. 94–111.

Ganguli, M., 1941. A note on nested sampling. Sankhya, V. 5, pp. 449–452.

Garrett, R. G., 1969. The determination of sampling and analytical errors in exploration geochemistry. Econ. Geol., V. 64, No. 5, pp. 568–569.

—— , 1973. The determination of sampling and analytical errors in exploration geochemistry—a reply. Econ. Geol., V. 68, No. 2, pp. 282–283.

—— , 1977. Sample density investigations in lake sediment geochemical surveys of Canada's Uranium Reconnaissance Program. Symposium on hydrogeochemistry and stream sediment reconnaissance for uranium in the United States. United States Dept. of Energy, Grand Junction, Colorado, pp. 173–185.

Goldsmith, C. H. and Gaylor, D. W., 1970. Three stage nested designs for estimating variance components. Technometrics, V. 12, No. 3, pp. 487–498.

Graybill, F. A., 1961. An introduction to linear statistical models, Vol. 1. McGraw-Hill Book Co., New York, 463 p.

Hahn, G. J., 1970. Additional factors for calculating prediction intervals for samples from a normal distribution. Jour. Am. Statis. Assoc., V. 65, pp. 1668–1676.
Hornbrook, E. H. W. and Garrett, R. G., 1976. Regional geochemical lake sediment survey, east-central Saskatchewan. Geol. Surv. Can. Paper 75–41, 20 p.
Howarth, R. J. and Lowenstein, P. L., 1971. Sampling variability of stream sediments in broad-scale regional geochemical reconnaissance. Trans. I.M.M., V. 80, pp. B363–372.
Krumbein, W. C. and Slack, H. A., 1956. Statistical analysis of low-level radioactivity of Pennsylvanian black fissile shale in Illinois. Bull. Geol. Soc. Am., V. 67, No. 6, pp. 739–761.
Leone, F. C., Nelson, L. S., Johnson, N. G. and Eisenstat, S., 1968. Sampling distributions of variance components II. Empirical studies of unbalanced nested designs. Technometrics, V. 10, No. 8, pp. 719–737.
Michie, U. McL., 1973. The determination of sampling and analytical errors in exploration geochemistry—discussion. Econ. Geol., V. 68, No. 2, pp. 281–282.
Miesch, A. T., 1964. Effects of sampling and analytical error in geochemical prospecting. Computers in the Mineral Industries (Pt. 1), Ed. G. A. Parks. Stanford Univ. Publ. Geol. Sci., V. 9, No. 1, pp. 156–170.
Miesch, A. T., 1967. Theory of error in geochemical data. U.S. Geol. Surv. Prof. Paper 574-A, 17 p.
Miesch, A. T., 1976. Geochemical survey of Missouri-methods of sampling, laboratory analysis and statistical reduction of data. U.S. Geol. Surv. Prof. Paper 954-A, 39 p.
Satterthwaite, F. E., 1946. An approximate distribution of estimates of variance components. Biometrics, V. 2, No. 2., pp. 110–114.
Searle, S. R., 1971. Linear Models. John Wiley and Sons Inc., New York, 532 p.
Snee, R. D., 1974. Computation and use of expected mean squares in analysis of variance. Jour. Qual. Tech., V. 6, No. 3, pp. 128–137.
Swallow, W. H. and Searle, S. R., 1978. Minimum Variance Quadratic Unbiased Estimation (MIVQUE) of variance components. Technometrics, V. 20, No. 3, pp. 265–272.
Tidball, R. R. and Severson, R. C., 1976. Chemistry of northern Great Plains soils. U.S. Geol. Surv. Open File Rept. 76-729, pp. 57–81.
Tietjen, G. L. and Moore, R. H., 1968. On testing significance of components of variance in the unbalanced nested analysis of variance. Biometrics, V. 24, No. 2, pp. 423–429.

APPENDIX

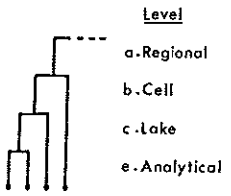
Computation of Constants Used in the Confidence Factor (CF) Formula.

An integral part of the CF is the standard error of the chosen level mean. Following Leone et al. (1968), the variance of the mean for the b (or 2nd) level of the design is written: (refer to table 1 for notation)

$$\text{Var}(\bar{x}_{ij..}) = \underbrace{\left(\frac{\sum_k n^2_{ijk}}{n^2_{ij.}}\right)}_{a_3} s^2_c + \underbrace{\left(\frac{n_{ij.}}{n^2_{ij.}}\right)}_{a_4} s^2_e$$

The constants or coefficients (a_3 and a_4) of the respective variance component estimates represent the average number of analyses at the c (3rd) and e (4th) levels of the design accordingly.

For the Type I sub-sampling structure illustrated below:



The a's are calculated as follows:

$$a_3 = \frac{\sum (\text{number in c-level})^2}{(\text{number in b-level})^2} = \frac{2^2 + 1^2}{3^2} = 0.555$$

$$a_4 = \frac{\sum (\text{number in e-level})}{(\text{number in b-level})^2} = \frac{2 + 1}{3^2} = 0.333$$

The recipricals of a_3 (1.8) and a_4 (3) are used in the text format of the CF calculations for the cell level.

Similarly, from the variance of the mean for the c (or 3rd) level of the design, that is,

$$\text{Var}(\bar{x}_{ijk.}) = \underbrace{\left(\frac{1}{n_{ijk}}\right)}_{a_4} s^2_e$$

$$a_4 = \frac{1}{(\text{number in c-level})} = 0.5$$

Again, the reciprocal of a_4(2) is used in the text format of the CF calculation for the lake level.

Note that the computation of these constants (a's) was based strictly on the use of the Type I sampling structure. In effect, the approximations introduced (particularly for Melville data) are deemed minimal on the resulting CF calculations, such as displayed in Table 5 of the text.