NASA TECHNICAL
MEMORANDUM

NASA TM X-64762

NOTES FOR THE IMPROVEMENT OF A REMOTE SENSING MULTISPECTRAL DATA NON-SUPERVISED CLASSIFICATION AND MAPPING TECHNIQUE

By Charles C. Dalton
Aero-Astrodynamics Laboratory

July 27, 1973

NASA

*George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama*

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO.
NASA TM X-64762

2. GOVERNMENT ACCESSION NO.

3. RECIPIENT'S CATALOG NO.

4. TITLE AND SUBTITLE
Notes for the Improvement of a Remote Sensing Multispectral Data Non-Supervised Classification and Mapping Technique

5. REPORT DATE
July 27, 1973

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
Charles C. Dalton

8. PERFORMING ORGANIZATION REPORT #

9. PERFORMING ORGANIZATION NAME AND ADDRESS
George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama 35812

10. WORK UNIT NO.

11. CONTRACT OR GRANT NO.

12. SPONSORING AGENCY NAME AND ADDRESS
National Aeronautics and Space Administration
Washington, D. C. 20546

13. TYPE OF REPORT & PERIOD COVERED
Technical Memorandum

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The report examines the "Sequential Clustering" technique for the unsupervised automatic classification and mapping of earth resources satellite data, makes theoretical analysis of the tests which were used, and derives an alternative set of tests and their necessary algorithm.

17. KEY WORDS
Unsupervised Classification
Earth Resources Mapping
ERTS Algorithm
Pattern Recognition

18. DISTRIBUTION STATEMENT

for
E. D. Geissler
Director, Aero-Astrodynamics Laboratory

19. SECURITY CLASSIF. (of this report)	20. SECURITY CLASSIF. (of this page)	21. NO. OF PAGES	22. PRICE
Unclassified	Unclassified	33	NTIS

TABLE OF CONTENTS

Notes for the Improvement of a Remote Sensing Multispectral Data Non-Supervised Classification and Mapping Technique

I. INTRODUCTION

A. Background

Any adequate analysis of flight data statistics from remote sensing multispectral scanners leads toward a computational burden so extensive for repeated and extended scenes of the earth from orbit that we endeavor to adapt or develop and perfect an effective algorithm for the automatic articulation of a scene. To be adequate, such a method must both be effective, or thorough in that it articulates a scene accurately, and be efficient, or economical in that its computational requirements are reasonably within the state-of-the-art. The sought technique is a non-supervised classification and mapping technique to the extent that it should achieve articulation of the scene independently of any other information or training area. The interpretation of the variously articulated and correspondingly mapped characteristics of a scene of interest would be obvious only to a limited extent, and adequate identification would largely require comparison with ground truth information for their complete identification. However, the advantages of being able to complete automatically so much of the analysis and a considerable compression of data should be obvious. Consequently, it has not gone without notice that mappings of, per se, the spectrally dependent articulations and their easily followed deriatives may play the most fundamental role in any analysis for change detection. Probably the retrieval of automatic change detection from multispectral scanner data must presuppose an adequate algorithm for the non-supervised automatic articulation of the undoubtedly obscure spectral functions of the data from repetitive overflights of a scene of interest or from scene to scene as the case may be. This follows from the mentioned functions being obscure and inconstant even in the absence of any changes in the characteristics of interest when other conditions change. This means that the signature of an item of interest may be somewhat variable inadvertently, seemingly necessitating insatiable demands for ground truth data in order to re-calibrate the signatures before data classification can be continued in those techniques which require supervision. Contrariwise, the techniques which we pursue, the unsupervised techniques, adjust automatically to any changes in the signatures.

B. Present Situation

During recent months there has been documented two different algorithms for the subject technique: (1) Su's[1] model, called "Sequential Clustering," and Jayroe's[2] model, "Spatial and Spectral Clustering." Each of the authors[1,2], using samples of data, gave sufficient results to prove that his model separately constitutes a major accomplishment. Each model works; yet, the two models are quite different. Therefore, any immediate attempt to combine the two models before they are fully developed and better understood might be deleterious to their collective potential.

C. Opinion and Purpose

It does not seem prudent at this time to favor one of the models, in their present forms, over the other model or to decide which one of them has the best potential. Consequently, to minimize comparison between the two models at this time, this note will not further review Reference 2.

The purpose of this critical review is to try to find any parts of the model[1] for which there may be a theoretical basis for a revision which might improve its effectiveness without sacrificing computational efficiency. The present model[1] had the benefit of adjustments after experience with data. Similarly, the considerable further revision based on theoretical considerations given in this note should benefit if parameter adjustments will be fine tuned through experimentation with data.

II. DISCUSSION

A. Description

Anyone wanting a general account of the "Sequential Clustering" model more briefly than its developer[1] gave will find a very helpful brief coverage of its principles and operation given by Krause and Frederick[3]. They[3] identify the sequential variance analysis as the key to Su's[1] work, and they note that it was originally developed by Krause, Jones, and Fisher[4] to detect periods of stationary behavior in time series. Howsoever, the least-squares derivations of the sequential variance formulas based on modes of chi-square are those which were given by Su and Krause[5]. Possible improvements to those key formulations will be suggested in this note Section II. B.

The sequential variance analysis is used in Su's[1] algorithm to test whether scan line segments are homogeneous and to test which line segments should be merged in the initial spectroscopic classification. Preprocessing depends on the type of data and the objectives of the analysis, may be necessary for higher accuracy. The first pass with the data establishes the measures of the classes into which the data will be classified on the second pass. Iteration may be necessary because after the data are put into the classes they, the data, do themselves give better measures of the classes than those measures which could be found in the first pass, etc. Anything which can be done to increase the accuracy of the sequential statistical tests should both: (1) reduce the amount of iterative computation necessary to give the best results, and (2) ultimately give better results.

B. Statistical Sequential Clustering

1. Establishing New Classes

Equation (2-4) of Reference 1 shows that a set of $M \geq 6$ resolution elements are considered to be a homogeneous set of samples from a new population or class when the M points, which represent them in the hyperspace for K spectral channels, are such that the squares of the ratios of their distances from their mean and the distance from the origin to their mean are all $\leq T^2$, where T^2 "is some threshold value to be given." No reasons were offered and no discussion was given to show whether or not the value to be used for T^2 should depend on M. Also, no reasons were given for using the mean distance to normalize the distances from the mean. The criterion seems somewhat discordant, opposite from what one would have expected; e.g., haze increases the albedo of the atmosphere while lowering the ground level illumination, and both effects reduce contrast in ground level images. Yet, the cited criterion says that when the reflected illumination is high, then the difference between different classes must also be higher in order for such difference to be accepted as meaningful.

Instead of normalizing by the mean, one should normalize each of the K components of the deviation from the mean by the sample estimate of the standard deviation s_k in that same dimension, where, following equation (2-1) of Reference 1,

$$s_k = \left[\frac{1}{M} \sum_{i=1}^{M} (x_{ki} - \bar{x}_k)^2 \right]^{1/2} , \qquad (1)$$

the unbiased estimate $(\frac{M}{M-1})^{1/2} s_k$ is not used herein for the further derivations. Then, equation (2-2) of Reference 1 would be replaced by

$$\Delta x_i^2 = \sum_{k=1}^{K} \left(\frac{x_{ki} - \bar{x}_k}{s_k} \right)^2 . \qquad (2)$$

Besides deleting the denominator in equation (2-4) in Reference 1 there is a further consideration of the relation between T^2 and $M \geq 6$. Because the coordinates x_{ki} are proportional to spectral radiant intensity in channel k (see page 2-4 of Reference 1) they can have only positive values and therefore cannot quite have normal distributions. Nevertheless, from a hypothetical spherical joint normal distribution in the K dimensions one can approximate roughly the relation which one might reasonably expect between T^2 and M. First, consider the case where the population mean μ and variance σ^2, per dimension, are known or where M is large enough for their accurate determination. Then, for a K-dimensional spherical distribution (meaning zero covariances), the ratio of the square of the resultant distance d^2 from the mean and the variance σ^2 per dimension has a chi-square distribution with K degrees of freedom; the expected value is K.

Let P be the right side area for χ_P^2; i.e., for an individual sample the probability is 1 - P that

$$d^2/\sigma^2 \leq \chi_P^2 \ . \qquad (3)$$

For the set of M samples, from the hypothetical spherical distribution, (where one tests the hypothesis that the consecutive samples are random observations of the same population) the probability P_o that not any of them fail to satisfy equation (3) is

$$P_o = (1 - P)^M \ . \qquad (4)$$

If one or more of them fail inadvertently to satisfy equation (3), then the first sample is discarded, the probability P_1 for which is

$$P_1 = 1 - (1 - P)^M \ . \qquad (5)$$

However, because each discarded sample is replaced with another sample, one finds essentially that each sample has the same probability of being discarded; that P_1 in equation (5) is also the fraction of samples which are discarded. When PM is numerically much smaller than unity, then the right side of equation (5) is approximated very well by PM, giving for the area (right side) index P in equation (3)

$$P \simeq P_1/M \ . \qquad (6)$$

Consider now another case of a hypothetical distribution which is known to be normal in K statistically independent dimensions. Let it be just coincidental that the distribution is spherical, and let the coordinate means and variances be estimated from M random samples. One wants, then, to establish population probabilities for Δx^2, the square of the displacement from the centroid of the M samples when its sample values Δx_i^2 are given by equation (2). It was given that x_k is normally distributed with mean μ_k and variance σ_k^2. Then consider

$$\Delta x^2 = \sum_{k=1}^{K} r_k^2 \tag{7}$$

where

$$r_k = \frac{x_k - \bar{x}_k}{s_k} \quad . \tag{8}$$

The variable r_k in equations (7) and (8) has a one-dimensional r distribution with M - 2 degrees of freedom, for which tables are given in Reference 6. With summations for the M samples one finds for r_k that the sample estimates of the mean and variance are zero and unity, respectively. The table of areas of $|r_k|$ show, for example, that the 99 percentile of $|r_k|$ is an increasing function of M, giving, to the limit of the table,

$$\left.\begin{array}{c} 2.051 \leq |r| \leq 2.556 \\ 6 \leq M \leq 122 \end{array}\right\} \quad . \tag{9}$$

The same percentile for the standard normal distribution is 2.576. The parameter r^2 will be used in the derivations in Section II. C. for the sum of F parameters, but in the present section one will not try to develope the distribution of Δx^2 in equation (7) as a K-dimensional F distribution; instead, one approximates $(x_k - \bar{x}_k)/s_k$ by a normal variable $(x_k - \mu_k)/\sigma_k$ and hopes that agreement is sufficient to support the approximation that each of the k variables r_k in equation (7) is normally distributed with zero mean and unit variance. This is more particularly tempting because the approximation is used only to support the χ^2 distribution Δx^2 in equation (7). So far as one can assume that the data in any of the k channels is statistically independent of the data in any other channel, it follows that the distance square Δx^2 in equation (7) (normalized for individual components as in equation (8)) is χ^2 distributed with K degrees of freedom. Because, in this example, approximately the same value was found for the variance in any channel it follows that Δx^2 in equation (7) is the same ratio as in equation (3); then equations (3) through (6) apply in this case also. More rigorous tests are developed in Section II. C. considering correlation, etc.

The examples just considered are unnecessarily restrictive; it still follows that Δx^2 in equation (7) has the χ^2 distribution with K degrees of freedom regardless of whether or not the variances in the different channels are quite different. This follows because the r distribution for the r_k in equations (7) and (8) is independent of the mean and variance of any x_k. It is anticipated that the extent to which the multispectral scanner data will be non-normal will have only negligible effect on the end results just given. One does anticipate, though, that non-vanishing cross-channel correlation may have a practical effect in that the number of degrees of freedom in the χ^2 distribution of Δx^2 in euqation (7) may be effectively somewhat less than K. Otherwise, instead of the equation (2-4) in Reference 1, one would require all the Δx_i^2 calculated for the set of M samples by equation (2) to satisfy the following criterion, with K degrees of freedom for χ^2:

$$\Delta x_i^2 < \chi^2_{(P_1/M)} \tag{10}$$

where (P_1/M) is the right side area for χ^2 and P_1 is the average fraction of samples which one is willing to discard inadvertently before deciding that a homogeneous population is being sampled (new class). For example, if K is 4 and M is 6, and if one prefers not to discard more than six percent of the samples when they are homogeneous, then all of the Δx_i^2 must be less than 13.3; the average or expected value would be K or 4 and the mode or most frequently occurring value would be K - 2 or 2.

The fraction of inadvertent rejects P_1 in equations (5), (6), and (10) is one type of risk, say "producer's risk." There is also a "consumer's risk," the fraction P_2 of samples which should have been rejected but which are inadvertently included in a new class. The three parameters P_1, P_2, and M are approximately related not only by

$$\frac{\partial^{P_1}}{\partial M} \simeq P_1/M \tag{11}$$

which follows from equations (6) and (10), but also by

$$\frac{dP_2}{d(P_1/M)} < 0 \tag{12}$$

which follows from geometrical considerations of neighboring populations. Operationally, though, one should ignore P_2; should consider empirically a parameter 0 as a function of the two independent parameters M and P_1, where 0 is a judicious measure of the quality and computational efficiency of the analysis. Ideally one would like to have iso-0 contour curves plotted on a graph of M versus P_1 which would be generated by practice with typical data. The results would be used to perfect the model expressed by equation (10).

2. Merging Excessive Classes

When the number of established classes exceeds the prescribed maximum allowable number W_{max} it is necessary to combine the two classes which are most similar. Reference 1 used the Euclidian distance between the means of two classes as the measure of similarity for this purpose. Instead, it is more pertinent and almost as easy to use a distance measure in which difference between the means in each of the K spectral dimensions is normalized by the two-class estimate of its standard deviation. One should, by assuming statistical independence between class i and class j, replace equation (2-8) of Reference 1 by

$$D^2_{i,j} = \sum_{k=1}^{K} \left(\frac{\bar{x}_{i,k} - \bar{x}_{j,k}}{s_{i,j,k}} \right)^2 \tag{13}$$

where $s^2_{i,j,k} = \frac{s^2_{i,k}}{m_i} + \frac{s^2_{j,k}}{m_j}$

where $s^2_{i,k}$ is the variance in dimension k for the m_i samples in class i and $s^2_{j,k}$ is the variance in dimension k for the m_j samples in class j. One can test the hypothesis that the populations which the two classes represent are not different beyond some level of significance or probability P_c. Then, to the extent that the normalized differences in euqation (13) are approximately normally distributed with zero mean and variance one, and to the further extent that the components in the K different dimensions are statistically indepedent, the squared difference $D^2_{i,j}$ from equation (13) has a χ^2 distribution with K degrees of freedom; i.e., the probability is P_c that

$$D^2_{i,j} \leq \chi^2_{(1 - P_c)} \tag{14}$$

For example, K is 4 for a multispectral scanner with four spectral channels; then, without interchannel correlation, it follows by equation (14) that the expected value of $D^2_{i,j}$ from equation (13) is 4, there is only a 10 percent chance that $D^2_{i,j}$ would be as small as 1.06, and there is even a 10 percent chance that it would be larger than 7.78.

3. Classifying New Samples Into Established Classes

a. Tests Being Used

In the statistical sequential clustering method which is used in Reference 1 each new sample is checked (to see to which one, if any, of the established classes it should belong) by a series of two tests.

The first test is a sequential test of the variance, to restrict its increase or decrease. Among the classes which are compatible with the new sample by the sequential variance test, the second test assigns the sample to the class for which the normalized distance from the mean is the least. Regardless of other changes which seem to be needed, reversing the order of two such tests would seem to be an improvement. When the apriori assumption is that the different classes represent populations which may have equally likely membership, then, one might test to see which classes, if any, are such that the normalized distances from the new sample to the means of the classes have reasonable values. Then, instead of choosing the smallest one of those values, one might prefer to consider, say, the smallest three values and use either a sequential variance test or a sequential mean test to find which one of the three classes would most nearly continue its sequence in the way which is in best agreement with the particular order of the compilation of the class.

b. Mean Versus Mode Estimators

In his sequential variance test, Su[1] continued as Su and Krause[5] had done by, beginning with equation (2-15), using the mode of χ^2 ("most probable value") instead of the mean (expected value); thus, the factor $(m_i - 3)/(m_i - 1)$ in equations (2-16) and (2-18) of Reference 1 and in equations (5) and (7) of Reference 5 is spurious and undoubtedly must bias the result considerably. Also, if the mean had been used instead of the mode, then the sequence (see equations (2-15), (2-17), and (2-18) of Reference 1) could have started with the second sample instead of the fourth.

c. Variance Versus Standard Deviation

Another discrepancy of unknown consequence in the sequential variance test in Reference 1 was continued as Su and Krause[5] had previously done by assuming that an appropriate estimator for standard deviation is the square root of the corresponding estimator for variance. This may be a reason for their having used the mode instead of the mean as a basis for the sequential analysis. Howsoever, equation (2-14) in Reference 1 is a correct beginning for the derivation of a sequential variance test:

$$\chi^2 = \frac{ms^2}{\sigma^2} \tag{15}$$

where m is the number of samples in a class being checked, including the prospective member as the last member where the sequence of compilation is preserved, and where some subscripts for channel number k, etc. are temporarily dropped for brevity. Instead of equation (2-15), the mean of χ^2 of m - 1 degrees of freedom is

$$E[\chi^2] = m - 1 \tag{16}$$

and, instead of equation (2-16), the mean of s_j^2 is, by equations (15) and (16),

$$\overline{s_j^2} = \left(\frac{j-1}{j}\right) \sigma^2 \quad \text{for } j = 2, 3, \ldots, m \,. \tag{17}$$

d. Normalization: Standard Deviation Versus Mean

Let $\sigma_{s_j^2}$ be the standard deviation of s_j^2; then, by Reference 6, it is

$$\sigma_{s_j^2} = \sigma^2 \; [2(j - 1)]^{1/2}/j \; . \tag{18}$$

The question at this point is whether, in the least squares summation as in equation (2-17) of Reference 1, the deviations from the mean should be normalized or not, and if so, with what? In References 1 and 5 the deviations from the mean were normalized by the mean, in equation (2-17) similarly as in equation (2-4) of Reference 1. It seems necessary to normalize the differences by the standard deviation in equation (18) so that the least squares determination is not dominated by a few of the most uncertain values. Then equation (2-17) in Reference 1 should be replaced by

$$\frac{\partial F(\sigma^2)}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2} \sum_{j=2}^{m} \left(\frac{s_j^2 - \overline{s_j^2}}{\sigma_{s_j^2}} \right)^2$$

$$= \frac{\partial}{\partial \sigma^2} \sum_{j=2}^{m} \left(\frac{j^2}{j-1} \right) \left[\frac{s_j^2}{\sigma^2} - \left(\frac{j-1}{j} \right) \right]^2 \Big/ 2 \tag{19}$$

$$= \frac{1}{\sigma^6} \left[\sigma^2 \sum_{j=2}^{m} j s_j^2 - \sum_{j=2}^{m} \left(\frac{j^2}{j-1} \right) s_j^4 \right] . \tag{20}$$

Then, where $\hat{\sigma}^2$ is the sequential estimator of σ^2 which makes equation (20) vanish, equation (2-18) of Reference 1 should be replaced by

$$\hat{\sigma}^2 = \sum_{j=2}^{m} \left(\frac{j^2}{j-1} \right) s_j^4 \Big/ \sum_{j=2}^{m} j\, s_j^2 \ . \tag{21}$$

e. The F Distribution Versus Chi-Square

Su and Krause[5] gave the same distribution, χ^2 with m - 1 degrees of freedom, for $m\ s^2/\hat{\sigma}^2$ as they had correctly given for $m\ s^2/\sigma^2$, and Reference 1 continued that presumption in its equation (2-19). It would be difficult to establish the distribution of $\hat{\sigma}^2$ in equation (21) from basic principles. That sequential estimator of the variance is the ratio of the sums of two series, but the corresponding terms in the two series not only are not statistically independent of each other but also are not statistically independent of preceding terms. The relations are very involved; however, it seems likely that the distribution of $m\ \hat{\sigma}^2/\sigma^2$ might not be appreciably different from that of ms^2/σ^2. Then, although it follows from the normal assumption for x that $m\ s^2/\sigma^2$ has a χ^2 distribution with m - 1 degrees of freedom it reasonably can be suspected that $m\ \hat{\sigma}^2/\sigma^2$ may have also approximately the χ^2 distribution with m -1 degrees of freedom. Therefore, their ratio $s^2/\hat{\sigma}^2$ probably could have nearly an F distribution with m - 1 and m - 1 degrees of freedom or not, depending on whether the correlation is low enough. Therefore, while the correlation has not been evaluated either theoretically or by Monte Carlo experiment, the distribution of the ratio is quite problematical, and using the chi-square limits in equations (2-19) and (2-20) of Reference 1 (and in equation (8) of Reference 5) is quite arbitrary and is not known to relate to the stated percentage of significance. Some further analysis to illustrate the nature of sequential tests is given in Section II. B. 4. herein.

f. Replacing Several Tests with Similar Tests

It will be shown in Section II. B. 4. that the kind of sequential test which is used in Reference 1 (to give a least-squares estimator of the variance) gives an estimator which differs from the one commonly used, the maximum likelihood estimator, in that the weight given to a member of a given sequence depends on its position in the sequence. In looking for ways to reduce the burden of computations, which sometimes increase as refinements are added, it will be shown in Section II. B. 4. that it is prudent tentatively to abandon sequential tests, for their use is not likely to be a reason for the effectiveness of the method which has been demonstrated in Reference 1. It seems likely, too, that the number of tests should be reduced. Instead of having two separate tests to classify a new sample, a sequential test of the variance and a non-sequential test of the mean (called χ^2-test and N-test in Reference 1), it seems preferable to replace those two tests with one non-sequential test of the deviation from the population mean. This test will be developed in Section II. C. from student's t distribution. Reasons why the same test, or a similar one, should also be used not only to replace the one to establish new classes but also to replace the one to merge excessive classes will also be given in Section II. D.

4. Nature of Least-Squares Sequential Tests

In a class of m samples, including the prospective member of the class, which are considered to be random observations x_j from a homogeneous normal population of observations of a characteristic scene in a given spectral channel, the maximum likelihood estimator $\bar{x}$ for the unknown mean μ of the population is

$$\bar{x} = (1/m) \sum_{j=1}^{m} x_j \, . \tag{22}$$

Equation (22) shows that $\bar{x}$ is a random variable, a function of the random observations x_j, with a value $\bar{x}_j$ for each serially-increasing sub-set j of m. The expected value and standard deviation of $\bar{x}_j$ are μ and $\sigma/\sqrt{j}$, respectively, where σ is the unknown standard deviation of the population x being sampled. The sum F of the squares of the normalized differences between $\bar{x}_j$ and μ is

$$F(\mu, \sigma^2) = \sum_{j=1}^{m} j(\bar{x}_j - \mu)^2/\sigma^2 \tag{23}$$

for which the partial derivative with respect to μ is

$$\begin{aligned} \frac{\partial F}{\partial \mu} &= (2/\sigma^2) \sum_{j=1}^{m} - j(\bar{x}_j - \mu) \\ &= (2/\sigma^2) \left(\mu \sum_{j=1}^{m} j - \sum_{j=1}^{m} j\bar{x}_j \right). \end{aligned} \tag{24}$$

Let $\hat{\mu}$ be the sequential estimator of μ such that its value for μ makes equation (24) vanish; then

$$\begin{aligned} \hat{\mu} &= \sum_{j=1}^{m} j\bar{x}_j / \sum_{j=1}^{m} j \\ &= \frac{x_1 + (x_1+x_2) + (x_1+x_2+x_3) + \ldots + (x_1+x_2+\ldots+x_m)}{\sum_{j=1}^{m} j} \end{aligned} \tag{25}$$

$$= \frac{x_m + 2x_{m-1} + 3x_{m-2} + \dots + mx_1}{\sum\limits_{j=1}^{m} j} . \tag{26}$$

Because of the statistical independence of the observations, it follows from equation (26) that the mean of $\hat{\mu}$ is the population mean μ and that the variance $\sigma^2_{\hat{\mu}}$ is

$$\sigma^2_{\hat{\mu}} = \sigma^2 \sum_{j=1}^{m} j^2 / \left(\sum_{j=1}^{m} j \right)^2$$

$$= 2(2m + 1)\ \sigma^2 / 3m(m + 1) . \tag{27}$$

Thus, $\bar{x}$ and $\hat{\mu}$, the two estimators of μ, have the same mean, and the ratio of their variances is

$$\sigma^2_{\hat{\mu}} / \sigma^2_{\bar{x}} = 2(2m + 1) / 3(m + 1) \tag{28}$$

which increases asymptotically from one toward 4/3 as m increases from one. In considering $\hat{\mu}$ in equation (26) as a random variable, one does, of course, imply that the specific observations x_j are to be replaced by not-yet-made observations, that they are a set of statistically independent normal variables, each with the same mean μ and variance σ^2. Thus, both $\bar{x}$ in equation (22) and $\hat{\mu}$ in equation (26) are linear functions of the same set of statistically independent normal variables, so they are also both normal and somewhat correlated.

The correlation coefficient ρ of $\bar{x}$ and $\hat{\mu}$ is related to the covariance λ, involving expected values E[ ], by

$$\rho = \lambda/\sigma_{\bar{x}}\sigma_{\hat{\mu}} \qquad (29)$$

$$\lambda = E[(\bar{x} - \mu)(\hat{\mu} - \mu)]$$

$$= E\left[\frac{1}{m}\left\{(x_1-\mu) + (x_2-\mu) + ..\right\}\left(\sum_{j=1}^{m} j\right)^{-1}\left\{(x_m-\mu) + 2(x_{m-1}-\mu) + \ldots\right\}\right]$$

$$= E[(x_m-\mu)^2 + 2(x_{m-1}-\mu)^2 + \ldots]/m \sum_{j=1}^{m} j$$

$$= E[(x-\mu)^2]/m = \sigma^2/m \qquad (30)$$

$$\sigma_{\bar{x}} = \sigma/\sqrt{m} \quad , \qquad (31)$$

Then, by equations (27), (30), and (31), it follows that the correlation coefficient ρ in equation (29) is

$$\rho = [3(m + 1)/2(2m + 1)]^{1/2} \qquad (32)$$

which decreased from a maximum value one to an asymtotic value 0.87 as m increases from one.

No way is evident whereby these results for $\hat{\mu}$ could be used to construct a criterion for classification. The purpose which is served, instead, is heuristic, to examine an estimator $\hat{\mu}$ which is simple enough for its properties to be shown and which belongs to the least-squares-sequential family in which $\hat{\sigma}^2$ in equation (21) is too difficult to analyze very well. Equation (26) shows that $\hat{\mu}$ involves weighting the members of the class in an arithmetic progression from the last to the first, and is therefore very insensitive to the last member or prospective member. It is difficult to see what advantage, if any, this might have. Actually the mean and variance of $\hat{\mu}$ and the correlation between $\hat{\mu}$ and $\bar{x}$ are all invariant to reversing the order of the weighting progression. The correlation 0.87 by equation (32) is even higher than one might have guessed: it probably is a good indication that all such estimators may be highly correlated with their corresponding unbiased or maximum likelihood counterparts. If so, then both the F distribution discussed in Section II. B. 3. e. and the χ^2 distribution, which was used, are quite inappropriate for equations (2-19) and (2-20) of Reference 1 and for equation (8) of reference 5.

C. Classification With F Distributions

1. F Distributions for Each Channel

Because the analysis so far in this note shows that the techniques which were used in Reference 1 to classify a new sample, to decide whether or not it should be put in an established class, are seriously deficient of any firm statistical theory basis, one now returns to develope further the technique of equations (7) and (8) of Section II. B. 1. in order to have not only a valid test which will serve to decide the addition of subsequent members but also a similar test to establish a new class. After the formulation has been developed in this section for zero correlation between channels, it will be revised in Section II. C. 3. for correlation.

The expedient by which the same test, or a similar test, can be used both for establishing a class and for deciding further membership in the calss is as follows: (1) a class with m members infers a population for which Δx^2 in equation (7) has a consequent distribution with limits which the prospective next member is required to satisfy, but (2) in checking for a new class each of the M prospective members is checked against possibly other limits for the same distribution which they collectively infer for a further prospective member. Thus, the two tests are different only because on the one hand the variance of $x - \bar{x}$ in equation (8) is different because x and $\bar{x}$ are statistically independent only for assignment of x to a class for which the mean has already been assessed as $\bar{X}$ and on the other hand the two tests may be different because different fiducial limits may be used for deciding to accept the hypothesis being tested.

The procedure which will be followed in the derivation is that r^2 in equation (7) is proportional to a variable which has an F distribution, etc.

The variance of the numerator in equation (8) is

$$\sigma^2_{x-\bar{x}} = \sigma^2(m\mp 1)/m \tag{33}$$

where the minus sign is used in testing for a new class and the plus sign is used in testing a new sample for membership in an established class. So, when the numerator of equation (8) is normalized with its standard deviation its square has a χ^2 distribution with one degree of freedom. Also, the square of the denominator time m/σ^2 has a χ^2 distribution with $m - 1$ degrees of freedom. The ratio of those two χ^2 variables with each divided by its own degrees of freedom has an F distribution with 1 and $m - 1$ degrees of freedom; i.e.,

$$\left(\frac{m-1}{m\mp 1}\right) r^2 = \left(\frac{m-1}{m\mp 1}\right) \quad [(x-\bar{x})/s]^2 \tag{34}$$

$$= \frac{[(x-\bar{x})/\sigma\sqrt{(m\mp 1)/m}]^2/1}{(ms^2/\sigma^2)/(m-1)}$$

$$= F(1,\ m-1), \tag{35}$$

where the mean and variance of F are

$$\mu_F = (m-1)/(m-3), \quad m > 3 \tag{36}$$

$$\sigma_F^2 = 2(m-1)^2(m-2)/(m-3)^2(m-5), \quad m > 5\ . \tag{37}$$

2. All Channels Without Correlation

The parameter which must be within limits for classification is, by equations (7), (34), and (35),

$$\left(\frac{m-1}{m\mp 1}\right) \sum_{k=1}^{K} \left(\frac{x_k - \bar{x}_k}{s_k}\right)^2 = \sum_{k=1}^{K} F_k(1,\ m-1) \tag{38}$$

where the choice of sign has the significance which was stated for equation (33). In the unlikely event that correlation between channels is small enough to be neglected, then the mean of the sum is the sum of the equal means and is $K\mu_F$, see equation (36), and the variance of the sum is the sum of the equal variances and is $K\sigma_F^2$, see equation (37). Of course, these results presuppose the equal weighting for the data from all channels as per equation (38), but if unequal weighting $w_k / \sum_{k=1}^{K} w_k$ or $w_k / \P_{k=1}^{K} w_k$ is wanted it has only to be inserted in both sides of equation (38).

3. With Inter-Channel Correlation

Regardless of how the K parameters F_k in equation (38) are correlated, the mean of the sum is the sum of the means,

$$\mu_{\Sigma F} = K \mu_F \tag{39}$$

and the means and variances of all of the F_k are invariant of k. Because all of the first partial derivatives of the right side of equation (38) with respect to the F_k are one, it follows exactly by the propagation of error (e.g., Reference 7) that the variance is

$$\sigma^2_{\Sigma F_k} = K\sigma_F^2 + 2 \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \lambda_{kl} \tag{40}$$

where the covariance λ_{kl} between F_k and F_l is, by equation (34),

$$\begin{aligned} \lambda_{kl} &= E[(F_k - \mu_F)(F_l - \mu_F)] \\ &= E[F_k F_l] - \mu_F^2 \qquad (41) \\ &= \frac{1}{m} \sum_{a=1}^{m} F_{ka} F_{la} - \mu_F^2 \end{aligned}$$

$$\simeq \left(\frac{m-1}{m \mp 1}\right)^2 \left(\frac{1}{s_k^2 s_l^2}\right) \frac{1}{m} \sum_{a=1}^{m} [(x_{ka} - \bar{x}_k)(x_{la} - \bar{x}_l)]^2 \tag{42}$$

and where σ_F^2 in equation (40) and μ_F in equation (39) are functions of m alone in euqations (36) and (37). Thus, the right side of equation (38) can be replaced by the sum of its mean from equation (39) and some constant A times the standard deviation from taking the square root of equation (40); i.e., the criterion is

$$\left| \left(\frac{m-1}{m+1}\right) \sum_{k=1}^{K} \left(\frac{x_k - \bar{x}_k}{s_k}\right)^2 - \mu_{\Sigma F} \right| / \sigma_{\Sigma F} \leq A . \tag{43}$$

The choice of sign in equations (42) and (43), again as in equations (35) and (38), is the same as that which was stated for equation (33).

It must not go without notice that the main computational burden is imposed by the necessity to compute the covariance in equation (40).

D. Merging by F Distributions

In Section II. B. 2. the squared distance between the empirical centroids of two classes, equation (13), was found by ignoring any interchannel correlation and by using the normal approximation to the components for which the sum of the squares is χ^2 distributed. It will now be shown, without making those approximations, how to transform equation (13) into a sum of parameters which have each an F distribution.

First consider only one component and temporarily drop the channel subscript k. Let μ_i and μ_j be the means of the populations for classes i and j which have size m_i and m_j, etc. as was stated for equation (13). Then, t_{ij} as shown in Reference 8,

$$t_{ij} = \frac{(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)}{\sqrt{m_i s_i^2 + m_j s_j^2}} \left[\frac{m_i m_j \ (m_i + m_j - 2)}{m_i + m_j} \right]^{1/2} \tag{44}$$

will have Student's t distribution with $m_i + m_j - 2$ degrees of freedom. Then, by Reference 9, t_{ij}^2 has an F distribution with one and $m_i + m_j - 2$ degrees of freedom; i.e.,

$$\left[\frac{m_i m_j (m_i + m_j - 2)}{m_i + m_j} \right] \frac{[(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)]^2}{(m_i s_i^2 + m_j s_j^2)} = F_{k,\ ij} \tag{45}$$

where the mean $\mu_{F_{k,\ ij}}$ and variance $\sigma^2_{F_{k,\ ij}}$ of $F_{k,\ ij}$ are, by Reference 6,

$$\mu_{F_{k,\ ij}} = \frac{m_i + m_j - 2}{m_i + m_j - 4} \ , \quad m_i + m_j > 4 \tag{46}$$

$$\sigma^2_{F_{k,\ ij}} = \left(\frac{m_i + m_j - 2}{m_i + m_j - 4} \right)^2 2 \left(\frac{m_i + m_j - 3}{m_i + m_j - 6} \right) , \quad m_i + m_j > 6 \ . \tag{47}$$

The K channels could be considered collectively by summing equation (45) just as equation (38) was given by summing equation (35); then, equations corresponding to equations (39) through (41) would follow by changing the notation

$$\lambda_{kl,\ ij} = E[F_{k,\ ij} F_{l,\ ij}] - \mu_{F_k}^{\ 2} \tag{48}$$

$$= \left[\frac{m_i m_j (m_i + m_j - 2)}{m_i + m_j}\right]^2 \quad E\left[\left\{\frac{[(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)]^2}{m_i s_i^2 + m_j s_j^2}\right\}_k \left\{\frac{[(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)]^2}{m_i s_i^2 + m_j s_j^2}\right\}_1\right] - \mu_{F_k}^2 \tag{49}$$

$$= \left[\frac{m_i m_j (m_i + m_j - 2)}{m_i + m_j}\right]^2 \quad \frac{E(k,1)}{(m_i s_{ki}^2 + m_j s_{kj}^2)(m_i s_{1i}^2 + m_j s_{1j}^2)} \tag{50}$$

$$E(k,1) = E\left[\left\{(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)\right\}_k^2 \quad \left\{(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)\right\}_1^2\right] \tag{51}$$

$$= E\left[\left\{(\bar{x}_{ki} - \mu_{ki})(\bar{x}_{1i} - \mu_{1i})\right\}^2 + \left\{(\bar{x}_{kj} - \mu_{kj})(\bar{x}_{1j} - \mu_{1j})\right\}^2\right] \tag{52}$$

$$= \frac{1}{m_i} \sum_{a=1}^{m_i} [(x_{kia} - \bar{x}_{ki})(x_{1ia} - \bar{x}_{1i})]^2 + \frac{1}{m_j} \sum_{b=1}^{m_j} [(x_{kjb} - \bar{x}_{kj})(x_{1jb} - \bar{x}_{1j})]^2 \tag{53}$$

where equation (52) follows from equation (51) because correlation is assumed to be appreciable only between channels within a class and not between a given channel and given class and a different channel and different class. Whether or not such correlations might be sufficiently small to support elegantly the computational expedient by which equation (51) is replaced by (52) and in turn by (53) could be established by analysis of representative data, but only relative results are needed in the test for merging excessive classes because it is only a question of which two classes to merge and not a question of whether or not to merge any classes.

It will be seen that the summed terms in equation (53) are the same as that in equation (42) when they are converted to the same notation; thus, the criterion for merging two classes does not require a separate computation of such summations which are already used in the criteria for forming new classes and classifying new samples into established populations.

III. ALGORITHM FOR UNSUPERVISED CLASSIFICATION USING F DISTRIBUTIONS

For each class or prospective class one needs values for the following parameters:

m = number of members in the class, $m \geq 6$

$\bar{x}_k = \frac{1}{m} \sum_{a=1}^{m} x_{ka}$, class mean in each channel $k = 1, 2, \ldots, K$

$s_k^2 = \frac{1}{m} \sum_{a=1}^{m} (x_{ka} - \bar{x}_k)^2$, class variance in each channel

$Q_{kl} = \frac{1}{m} \sum_{a=1}^{m} [(x_{ka} - \bar{x}_k)(x_{la} - \bar{x}_l)]^2$ each pair of channels k and l

$\lambda_{kl} = \left(\frac{m-1}{m+1}\right)^2 Q_{kl}/s_k^2\, s_l^2$ " " " " " " "

$\mu_F = \left(\frac{m-1}{m-3}\right)$

$\sigma_F^2 = 2\mu_F^2 \left(\frac{m-2}{m-5}\right)$

$\mu_{\Sigma F} = K\mu_F$

$\sigma_{\Sigma F}^2 = K\sigma_F^2 + 2 \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \lambda_{kl}$.

Also, for each pair of established classes i and j containing m_i and m_j members one needs values for the following parameters:

$$\mu_{F_{ij}} = (m_i+m_j-2)/(m_i+m_j-4)$$

$$\sigma^2_{F_{ij}} = 2\mu^2_{F_{ij}} \; (m_i+m_j-3)/(m_i+m_j-6)$$

$$\Sigma F_{k,ij} = \frac{m_i m_j (m_i+m_j-2)}{m_i+m_j} \sum_{k=1}^{K} \frac{(\overline{x}_{ki}-\overline{x}_{kj})^2}{m_i s^2_{ki} + m_j s^2_{kj}}$$

$$\lambda_{kl,ij} = \left[\frac{m_i m_j (m_i+m_j-2)}{m_i+m_j}\right]^2 \frac{(Q_{kl,i}+Q_{kl,j})}{(m_i s^2_{ki}+ m_j s^2_{kj})(m_i s^2_{li}+ m_j s^2_{lj})}$$

$$\mu_{\Sigma F_{ij}} = K\mu_{F_{ij}}$$

$$\sigma^2_{\Sigma F_{ij}} = K\sigma^2_{F_{ij}} + 2 \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \lambda_{kl,ij}$$

$$A_{ij} = \left| (\Sigma F_{ij} - \mu_{\Sigma F_{ij}})/\sigma_{\Sigma F_{ij}} \right| .$$

The two other formulas which are always used together, with a purpose which depends on what datum is substituted for the parameter x_k, are

$$\left.\begin{aligned} \Sigma F &= \left(\frac{m-1}{m+1}\right) \sum_{k=1}^{K} (x_k - \bar{x}_k)^2 / s_k^2 \\[2em] A &= \left| (\Sigma F - \mu_{\Sigma F}) / \sigma_{\Sigma F} \right| \end{aligned}\right\} \qquad (54)$$

The number W of retained classes must not exceed an allowable number W_{max}.

Step 1. Read control parameters A_o, A_1, $M \geq 6$, and W_{max}.

Step 2. Read the first M samples.

Step 3. Calculate parameters for prospective class.

Step 4. With the $\bar{x}_k$ and s_k^2 from step 3, calculate a value of A in equation (54) for each of the M samples by using the values of x_k for that particular sample in equation (54) with the minus sign. Does the largest value of A satisfy $A < A_o$? Yes: go to step 7. No: go to step 5.

Step 5. Discard the first sample accumulated.

Step 6. Read a new sample, then go to step 3 (recursion formulas may be helpful).

Step 7. Designate a new class having the parameters extant.

Step 8. Does the program reach the end of the sample sequence? Yes: go to step 9. No: go to step 11.

Step 9. Print out any parameters and classification map which are required by the Flight Data Statistics Office.

Step 10. Stop.

Step 11. Does the number of classes W satisfy $W \leq W_{max}$? Yes: go to step 14. No: go to step 12.

Step 12. Calculate class-pair parameters for all combinations of classes in pairs (recursion formulas may be helpful).

Step 13. Combine the two classes i and j which give the smallest pair-parameter A_{ij} and compute the single-class parameters of the resulting class.

Step 14. Read a new sample.

Step 15. By using the values of x_k from the new sample in equation (54) with the plus sign, calculate a value of A for each of the W established classes according to their given values of $\bar{x}_k$, σ_k^2, etc. Does the smallest one of the m values of A satisfy $A \leq A_1$? Yes: add the sample to that class, revise the parameters of that class and go to step 8. No: put the sample in hold and go to step 16.

Step 16. Has the number of samples in hold reached M? No: go to step 14. Yes: go to step 17.

Step 17. Calculate parameters for prospective class.

Step 18. With $\bar{x}_k$ and s_k^2 from step 17, calculate a value of A in equation (54) for each of the M samples by using the values of x_k for the particular sample in equation (54) with the minus sign. Does the largest value of A satisfy $A < A_o$? Yes: go to step 19. No: discard the first one of the M samples held for step 17 and go to step 14.

Step 19. Designate a new calss with the parameter values which are extant (from step 17).

Step 20. Empty the hold from step 16 and go to step 8.

REFERENCES

1. Su, M. Y., "The Composite Sequential Clustering Technique for Analysis of Multispectral Scanner Data," NASA Contractor Report CR-128999, Contract NAS8-27364, October 1972.

2. Jayroe, R. R., "Unsupervised Spectral Clustering with Spectral Discrimination," NASA Technical Note D-7312, May 1973.

3. Krause, F. R., and Frederick, L. D., "Opportunities for Space Surveys of Moisture Anomalies," Proceedings of the Second "Remote Sensing of Earth Resources Conference," University of Tennessee Space Institute, Tullahoma, Tenn., March 26-27, 1973.

4. Krause, F. R., Jones, J. A., and Fisher, M. J., "Digital Analysis of Random Data Records by Piecewise Accumulation of Time Averages," NASA Technical Note D-6073, December 1970.

5. Su, M. Y., and Krause, F. R., "Automatic Processing of Multispectral Observations," AIAA Paper No. 71-234, "AIAA Integrated Information Systems Conference," Palo Alto, February 17-19, 1971.

6. Burington, R. S., and May, D. C., Jr., "Handbook of Probability and Statistics with Tables," Handbook Publishers, Inc., Sandusky, Ohio, 1953.

7. Deming, E. W., "Statistical Adjustment of Data," John Wiley and Sons, Inc., New York, 1943.

8. Hoel, P. G., "Introduction to Mathematical Statistics," John Wiley and Sons, Inc., New York, 1947.

9. Mood, A. M., "Introduction to the Theory of Statistics," McGraw-Hill Book Co., Inc., New York, 1950.