E83-10308 CR-171644

# AgRISTARS

SR-X1-04046
NAS9-15981

Supporting Research

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

March 1981

# IMPROVED VERSION OF THE SPLIT ROUTINE FOR CLASSY

M. E. Rassbach

Elogic, Inc.
4242 S.W. Freeway, Suite 304
Houston, Texas 77027

UNITED STATES DEPARTMENT OF AGRICULTURE
NASA
DEPARTMENT OF COMMERCE UNITED STATES OF AMERICA
U.S. DEPARTMENT OF THE INTERIOR March 3, 1849

Lyndon B. Johnson Space Center
Houston, Texas 77058

Cp13

1. Report No.
SR-X1-04046

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Improved Version of the SPLIT Routine for CLASSY

5. Report Date
March 3, 1981

6. Performing Organization Code

7. Author(s)
M. E. Rassbach

8. Performing Organization Report No.

9. Performing Organization Name and Address
Elogic, Inc.
4242 S.W. Freeway, Suite 304
Houston, Texas 77027

10. Work Unit No.

11. Contract or Grant No.
NAS9-15981

12. Sponsoring Agency Name and Address
National Aeronautics and Space Administration
Lyndon B. Johnson Space Center
Houston, Texas 77058

13. Type of Report and Period Covered
Technical Report

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This report develops the theoretical basis for a new version of the SPLIT routine for CLASSY. (The SPLIT routine is used to guess the positions of the parts of a cluster which is not normally distributed.) The proposed routine would be a faster, more accurate replacement for the existing SPLIT routine, which was developed on an ad hoc basis.

17. Key Words (Suggested by Author(s))
CLASSY algorithm
skewness
kurtosis

18. Distribution Statement

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
15

22. Price*

*For sale by the National Technical Information Service, Springfield, Virginia 22161

SR-X1-04046
NAS9-15981

# IMPROVED VERSION OF THE SPLIT ROUTINE FOR CLASSY

BY

M. E. RASSBACH

This report describes Classification activities of the Supporting Research project of the AgRISTARS program.

Elogic, inc.
4242 S.W. Freeway, Suite 304
Houston, Texas, 77027

March 3, 1981

# TABLE OF CONTENTS

# IMPROVED VERSION OF THE SPLIT ROUTINE FOR CLASSY

## Introduction

The SPLIT routine originally implemented for CLASSY was intended to be a quickly written, temporary routine to get the system running. In fact, another version had been designed and coded earlier, but was judged too difficult to debug quickly.

The original version of SPLIT attempted to minimize a quadratic form of the difference between the observed skewness and kurtosis and the skewness and kurtosis produced by the model. The model was then adjusted by a kind of steepest descents approach until further adjustment would not prove practical. This frequently requires 50 or more iterations, which can consume a fairly large amount of machine time. In fact, with other speedups of CLASSY being implemented by Elogic, SPLIT could well become the primary consumer of machine time.

In addition, analyses done since the original SPLIT was encoded indicate that it may not give the best solution for common situations. Thus a new SPLIT routine would have to produce a better set of solutions than the original version.

For both of the above reasons, Elogic designed a new SPLIT routine which is being implemented by Lockheed.

## Calculation Procedure

We wish to find approximate values for the parameters (mean, covariance, and proportion) of two distributions, given the same parameters plus the traces of skewness and kurtosis for the mixture distribution. The calculation is done in a coordinate frame where the mixture covariance is the unit matrix, which allows us to easily calculate inner products with respect to that covariance, etc.

The two clusters to be found are defined by

| | cluster | |
|---|---|---|
| proportion | $\alpha$ | $\beta$ |
| mean | $\mu$ | $\nu$ |
| covariance | $\sigma^2$ | $\tau^2$ |

We are given S (the skewness vector $S_k=(x^2x_k)$ ); and the input kurtosis matrix $K_1$, which is immediately used to derive the practical kurtosis matrix, $K=K_1-(d+2)\Sigma$, where d is the dimensionality of the space.

The initial variables are obtained as in the present SPLIT routine, and are used to set up the new clusters as in that routine. The reader is referred to the documentation for SPLIT Version I for the details of these calculations. In addition the rotation to the frame with unit covariance and the initialization of the iterations is done as in Version I. Version II differs from Version I only in the form of the iterative step and convergence calculation, and in the handling of certain special conditions.

We wish to solve the equations

$$\frac{1-\gamma^2}{4}\,\delta\delta^t + \frac{1+\gamma}{2}\,\sigma^2 + \frac{1-\gamma}{2}\,\tau^2 = 1$$ (covariance--a symmetric matrix, 1 means the unit matrix)

(1a)

$$\frac{1-\gamma^2}{4}\,(2D^2\delta + \delta\,\mathrm{tr}D^2 - \gamma\delta^2\delta) = S$$ (skewness--a vector)

(1b)

$$\frac{1-\gamma^2}{4}\,(2D^4 + D^2\,\mathrm{tr}D^2 + (\frac{3\gamma^2-1}{2})\,\delta^2\delta\,\delta^t -$$ (1c)

$$2\gamma\,\left[(D^2\delta)\,\delta^t + \delta\,(D^2\delta)^t\right] - \gamma\delta^2D^2 - \gamma\,\mathrm{tr}D^2\delta\,\delta)^t = K$$

(kurtosis, $(d+2)\Sigma^4$ removed; a symmetric matrix)

where

$-1\leq\gamma\leq 1$ is the mixing parameter

$\delta$ is the displacement vector between the clusters $(=\mu-\nu)$

(2a)

$D^2 = \sigma^2-\tau^2$ (2b)

The unknowns are $\sigma^2$, $\tau^2$, $\delta$, and $\gamma$.

This system has one more unknown than equations, requiring an additional assumption. The simplest such assumption is $\gamma=0$; other possibilities are $\mathrm{tr}D^2=0$, etc. We will use $\gamma=0$ (equal weight clusters), here, but some hueristic allowance must be made at the end of the calculation to avoid wild solutions. Some checking of solutions must be done in any event to avoid problems caused by multiple clusters, etc.

The equations to be solved become

$$\frac{1}{4}\,(2D^2\delta \;+\; \delta\;\mathrm{tr}D^2) = S \tag{3a}$$

$$\frac{1}{4}\,(2D^4 + D^2\mathrm{tr}D^2 - \frac{1}{2}\delta^2\;\delta\delta^t) = K \text{ (where } \delta^2 = \sum_i \delta_i{}^2) \tag{3b}$$

which are to be solved for $\delta$ and $D^2$. (Note that $\sigma^2$ and $\tau^2$ appear here only via $D^2$.) The definition of $D^2$, along with the equation

$$\frac{1}{4}\,\delta\delta^t + \frac{1}{2}\,\sigma^2 + \frac{1}{2}\,\tau^2 = 1, \tag{3c}$$

can be treated separately and later than (3a) and (3b) except for the positivity constraint on $\sigma^2$ and $\tau^2$. This requirement is handled as a special constraint during the solution of the system (2b) and (3c); if no solution is possible, the error solution is taken.

Under normal circumstances, the matrix K will have exactly one negative eigenvalue. If K has no negative eigenvalues, this indicates that the two hypothesized clusters differ primarily in covariance (a condition which might reasonably be flagged). A situation where K has multiple negative eigenvalues can be contrived with only two clusters, but this generally indicates multiple clusters, requiring a more even division of the space for later additional SPLITting.

## Additional Conditions

Besides the actual numeric solution, the following special conditions must be observed:

1) If there are excess negative eigenvalues below a certain threshold, probably indicating multiple subclusters, the "basic" (initial) solution is used. A message condition is also raised.

2) If the solution discovered has covariance matrices which would divide the space too unevenly, then the "basic" solution is also used, with a message condition.

3) The "basic" solution and a third message condition are used if the numerical iteration fails to converge.

The "basic" solution is a hueristically calculated solution to the equations which is used as the initial solution and as a last resort solution in the case of errors. It is in fact the initial solution for the original SPLIT routine.

## Iteration Procedure for Version II of SPLIT

Write $$\Delta S = \frac{1}{4} (2D^2\delta + \delta \operatorname{tr}D^2) - S \qquad (4a)$$

$$\Delta K = \frac{1}{4} (2D^4 + D^2 \operatorname{tr}D^2 - \frac{1}{2} \delta^2\delta\delta^t) - K \qquad (4b)$$

for the solution errors. (These will be 0 if $\delta$ and $D^2$ are correct.) The initial values are determined by the "basic" procedure, described above. The iterations are done in a frame where $D^2$ is constantly forced to be a diagonal matrix--that is, if $D^2$ is corrected by an off-diagonal term, the coordinate frame is rotated to make $D^2$ be diagonal again. (This requires rotating $D^2$, $\delta$, K, and S.)

Thus the frame in which the problem is stated rotates from iteration to iteration, overall from one in which K is diagonal to one in which the final solution $D^2$ is diagonal. These rotations must be stored in a separate rotation matrix which gives the net rotation between the transformed frame and the original one. When making the rotations, $\delta$ and $D^2$ are both rotated incrementally to the new frame, and S and K are rotated from their original values to the new frame by the accumulated transformation before $\Delta S$ and $\Delta K$ are calculated.

The variables are separated into two groups:

1) $\delta$ and the diagonal elements of $D^2$

2) the off-diagonal elements of $D^2$

The variables of Group 1 are modified during iteration using a matrix version of Newton's method; those in Group 2 are modified by direct substitution (Aitken iteration), implemented by frame rotation.

Iteration on $\delta$ and the diagonal elements of $D^2$:

We take $\Delta S$ and the diagonals of $\Delta K$ and use them to correct $\delta$ and the diagonals of $D^2$, via

$$\begin{pmatrix} \delta_{\text{new}} \\ \\ \text{diag } D^2_{\text{new}} \end{pmatrix} = \begin{pmatrix} \delta_{\text{old}} \\ \\ \text{diag } D^2_{\text{old}} \end{pmatrix} - \underbrace{\begin{pmatrix} \frac{\partial \Delta S}{\partial \delta} & \frac{\partial \Delta S}{\partial D^2_{\text{diag}}} \\ \frac{\partial \Delta K}{\partial \delta} & \frac{\partial \Delta K}{\delta D^2_{\text{diag}}} \end{pmatrix}^{-1}}_{\text{(2dx2d mtx)}} \begin{pmatrix} \Delta S \\ \\ \text{diag } \Delta K \end{pmatrix} \quad (5)$$

Since this is a Newton's method technique, it is subject to bad overshoots, etc. Therefore, a test must be made that the various error terms $\Delta S$ and diag $\Delta K$ decrease due to the iteration, if they do not, the increment given to $\delta$ and diag $D^2$ can be decreased. (For small enough increment, they are guaranteed to decrease.) This test may be applied to the whole iteration, not just the Newton's method part; this would save the need to make duplicate calculations of $\Delta S$ and $\Delta K$. In the worst case of non-convergence, the "basic" solution could be taken, with the raising of an error flag.

The various derivatives are:

$$\frac{\partial \Delta S_i}{\partial \delta_k} = \frac{1}{2} D^2{}_{ik} + \frac{1}{4} \delta_{ik} \mathrm{tr} D^2 \qquad \text{(5a)}$$

($\delta_{ik}$ is a Kronecker delta function, =1 if i=k, else 0; not to be confused with $\delta_i$ the vector, which has 1 or no subscripts, instead of 2)

$$\frac{\partial \Delta S_i}{\partial D^2{}_{k\ell}} = \frac{1}{4} \delta_\ell \delta_{ki} + \frac{1}{4} \delta_k \delta_{\ell i} + \frac{1}{4} \delta_i \delta_{k\ell} \qquad \text{(5b)}$$

$$\frac{\partial \Delta K_{ij}}{\partial \delta_k} = \frac{1}{4} \delta_i \delta_k \delta_j - \frac{1}{8} \delta_{ik} \delta_j \delta^2 - \frac{1}{8} \delta_{jk} \delta_i \delta^2 \qquad (\delta^2 = \sum_i \delta_i \delta_i) \qquad \text{(5c)}$$

$$\frac{\partial \Delta K_{ij}}{\partial D^2{}_{k\ell}} = \frac{1}{2} (D^2{}_{ik} \delta_{j\ell} + D^2{}_{i\ell} \delta_{jk}) + \frac{1}{4} (\delta_{ik} \delta_{j\ell} + \delta_{i\ell} \delta_{jk}) \mathrm{tr}(D^2)$$

$$+ \frac{1}{2} D^2{}_{ij} \delta_{k\ell} \qquad \text{(5d)}$$

Although these equations include the off-diagonal terms of $\Delta K$ and $D^2$, only the diagonal terms are used in the Newton's method matrix.

Off diagonal elements

The off diagonal elements of $D^2$ are adjusted using direct substitution. The off diagonal components of $\Delta K$ are used to determine the change in $D^2$ required. The amount of change required can be determined from $\partial \Delta K_{ij}/\partial D^2{}_{k\ell}$. For $i \neq j$, $k \neq \ell$,

$$\frac{\partial \Delta K_{ij}}{\partial D^2{}_{k\ell}} = \frac{1}{4} \left[ \left( 2D^2{}_{ik} + \delta_{ik} \mathrm{tr} D^2 \right) \delta_{j\ell} + \left( 2D^2{}_{i\ell} + \delta_{i\ell} \mathrm{tr} D^2 \right) \delta_{jk} \right] \qquad (6)$$

Note that $D^2$ is diagonal (in the current coordinate frame) so that either $i=k$ and $j=\ell$ or $i=\ell$ and $j=k$ for the derivative to be non-zero. By symmetry conditions on $\Delta K$ and $D^2$, we may assume $i>j$ and $k>\ell$, $i=k$ and $j=\ell$ so that we calculate

$$\left( \underset{\text{old}}{D^2{}_{k\ell}} - \underset{\text{new}}{D^2{}_{k\ell}} \right)_{\text{off diagonal}} = -2 \left[ 2D^2{}_{kk} + \mathrm{tr} D^2 \right]^{-1} \Delta K_{k\ell}$$

The two-part calculation of $\delta$ and $D^2$ is basically a Newton's method calculation with the derivative terms between the parts treated as 0.

## Final Solution

After $D^2$ and $\delta$ have been calculated, the program must check that the variances $\sigma^2$ and $\tau^2$ are satisfactory. This is done via the relations

$$D^2 = \sigma^2 - \tau^2$$

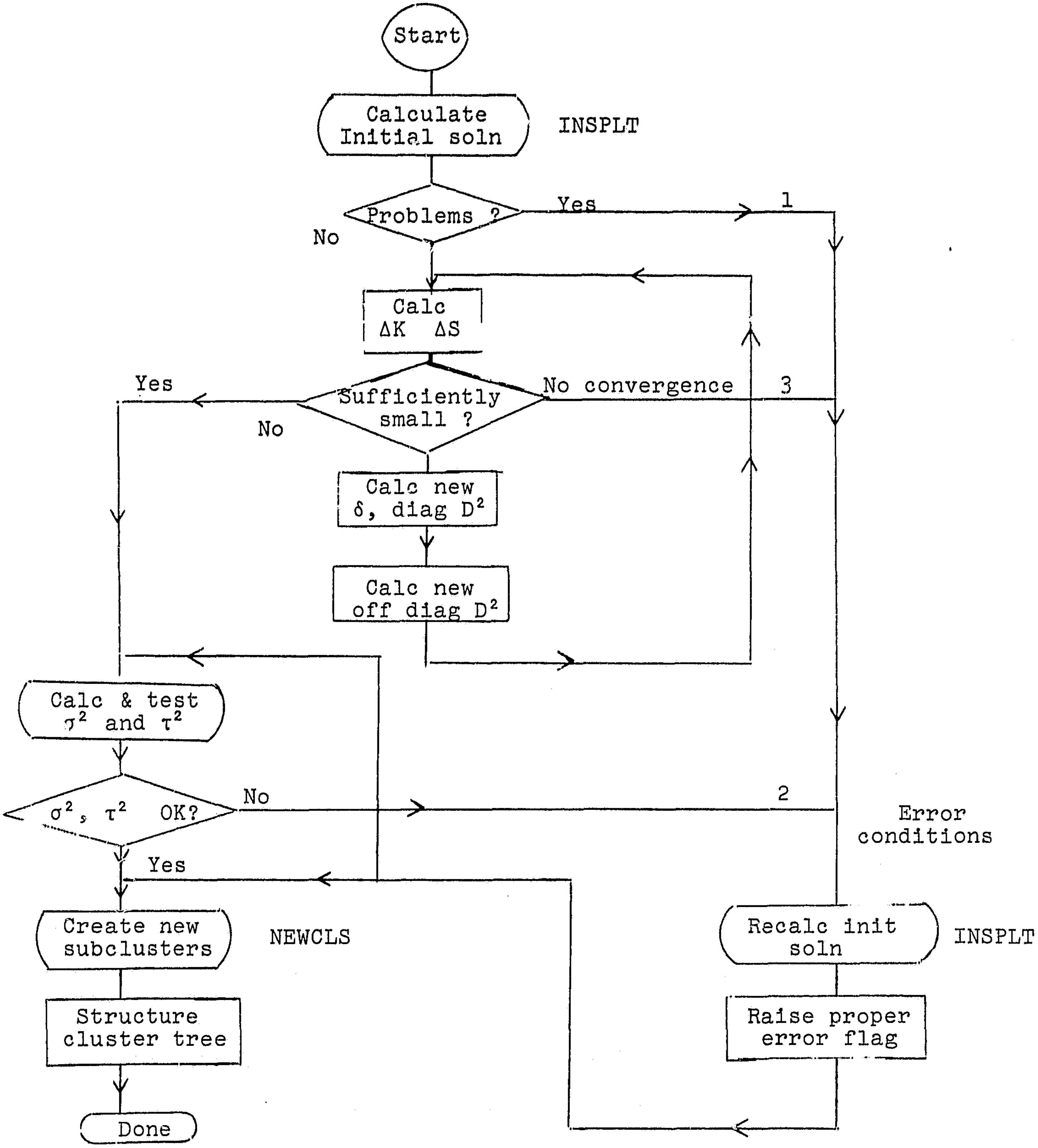
$$1 = \delta^t + \frac{1}{2}\sigma^2 + \frac{1}{2}\tau^2$$

$\sigma^2$ and $\tau^2$ may be calculated from these, and checked for positive definiteness. Actually, $\sigma^2$ and $\tau^2$ must be held fairly far away from zero eigenvalues and must not have too large or small a volume ratio. The actual threshold for these tests is under the control of parameters. If these tests fail, we revert to the "basic" solution (this is error 2 above).

After checking $\sigma^2$ and $\tau^2$, the system must be rotated back to the original frame of reference by the inverse of the accumulated transforms of $D^2$. These back-transformed $\delta$, $\sigma^2$, and $\tau^2$ can then be used to generate the pair of new clusters.

SPLIT VERSION 2 FLOW CHART

(includes subroutine names)



## Summary

This report has developed the theory for a new, improved version of the SPLIT routine. This version is intended to replace the old version which was put tcgether quickly, without using any fundamental understanding of the problem.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1981—771-014/779