AgRISTARS

SR-N1-04053
NAS9-15981

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

March 1981

Supporting Research

NEW OUTPUT IMPROVEMENTS FOR CLASSY

M. E. Rassbach

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

Lyndon B. Johnson Space Center
Houston, Texas 77058

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
SR-X1-04053		

4. Title and Subtitle	5. Report Date
New Output Improvements For CLASSY	6. Performing Organization Code
7. Author(s) M. E. Rassbach	8. Performing Organization Report No. NAS-811
9. Performing Organization Name and Address Elogic, Inc. 4242 S.W. Freeway, Suite 301 Houston, TX 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

Elogic has developed a number of new output data and formats for the CLASSY clustering algorithm. This report describes four such aids to the CLASSY user.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
CLASSY clustering	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	36	

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

SR-X1-04053
NAS9-15981

NEW OUTPUT IMPROVEMENTS FOR CLASSY

BY

M. E. RASSBACH

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

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

March 10, 1981

TABLE OF CONTENTS

NEW OUTPUT IMPROVEMENTS FOR CLASSY

Elogic has designed and implemented several improvements in both the form and content of CLASSY output. These improvements fall in four categories:

1) New statistical measures
2) Special map types
3) New formats for standard output
4) Special cluster display method

New Statistical Measures:

Coding has been added to CLASSY to calculate some additional statistical measures. Some of these are simply calculations made by the output formatting routine from quantities already available. These new calculations are described in the New Formats section, rather than here.

The new statistical measures calculated by the program give information about the relative probability distribution of the data. We define r, the fraction of the probability for a point which is allocated to a particular cluster, as

$$r = \frac{P_{cluster}}{P_{total}}$$

where the P's refer to posterior probabilities. r will always be 1 for a well-separated cluster, and will become smaller as the cluster becomes more mixed with other clusters. The distribution of r thus becomes a measure of how directly the algorithm can converge, as well as a measure of the separability of the clusters.

The new measures give a histogram of r and a special statistic useful when r is near 1. For the histogram, the values of r are divided into five regions, 0.8-1., 0.6-0.8, 0.4-0.6, 0.2-0.4, and 0-0.2. The weighted numbers of points falling into each of these regions is accumulated during the normal CLASSY processing. (Actually, the first region is calculated from the other four and the total weight.) This

forms a five point histogram of the weights falling into each r-interval, which is printed out in the new cluster printout.

In addition, the program accumulates the average r in the 0.8-1.0 interval, which is useful when the probabilities are generally close to 1. This statistic is also displayed in the cluster printout. The formats for both of these statistics are given in the New Formats section of this report.

Special Map Types

The map normally produced by CLASSY is a maximum likelihood classification of each pixel into the clusters generated by CLASSY. As part of the additional output designed by Elogic, an additional type of map is now available. This second map type is auxiliary to the first, and gives an indication of the degree of doubt in which the classifications are held. The posterior probability of the best class for each pixel is pigeon-holed into 5% intervals, p=95-100%, 90-95%, etc. These are then printed as a map, which may be compared to the original map type.

This second map type is useful in discovering the reasons for any misclassifications with CLASSY. For example, if most of the misclassification in some region is due to problems with boundary pixels, then the high uncertainty points will be along the boundaries. Similarly, isolated errors may be on points with high uncertainty.

To access the second map type, the user uses the MAPTYPE command in the control card input.

MAPTYPE 1 means ordinary map

MAPTYPE 2 means special map only

MAPTYPE 3 means both maps

The symbol set is different from that used in the ordinary map, and is given at the bottom of the map, along with counts for each interval. The SYMB instruction has been altered so that the first character of the parameter string denotes the maptype to which the symbols apply, and the remainder are the map symbols.

New Formats for Standard Output

The standard output from CLASSY was generally reformatted to improve organization and readability, and to eliminate extraneous output. The new formats are labeled and extensively described in Appendix A.

The new printouts reformat the cluster printout entirely, restructure the ADJUST printout to occupy only one line, and omit a great deal of extra printout in the main listing. Additional changes were made in the initial printout, in the timing printout, and in the cluster tree printout. Each group of printout has been labeled with the total number of points processed when it was printed. This gives an effective "clock" for relating the different events of a CLASSY run.

Special Cluster Display Method

During the course of this contract, Elogic has also developed a special method for displaying a multidimensional cluster in two dimensions.

A typical problem is the display of brightness-greenness covariances for several acquisitions. The brightness-greenness information for any one pass is easily displayed in two dimensions using the covariance ellipse. (Actually, it is better to use an ellipse with radius $\sqrt{2}$ times that of the covariance ellipse, to equalize the number of points inside and outside the ellipse.) For several acquisitions, the brightness-greenness ellipse can be projected to display each acquistion separately, giving one ellipse per pass. However, this form of display still ignores any relation between the passes. For three acquisitions (six channels) only 9 elements of the 21 in the overall covariance are displayed.

The correlation matrices usually used to generate the display acquisition as independent. Elogic has derived an additional type of projected correlation matrix which treats each acquisition as dependent. Displaying this with the standard covariance enlarges the number of items visually available in the 6-channel case from 9 to 18 (out of 21). It produces ellipses lying inside the covariance ellipses, which give the covariance of a point, assuming the other channels are given fixed values.

Suppose we let P be one of the six-channel to two-

channel projection operators. Then the ordinary two-channel covariance is $P^t \Sigma P$, where Σ is the six-channel covariance matrix. The additional (dependent) covariance displays $(P^t \Sigma^{-1} P)^{-1}$ as well, which is the covariance of one pass for fixed spectral values of the other passes. Since a lot of the variation in one pass may be correlated to variations in the other passes, holding those latter passes at a fixed value may considerably reduce the range of variation of the first pass.

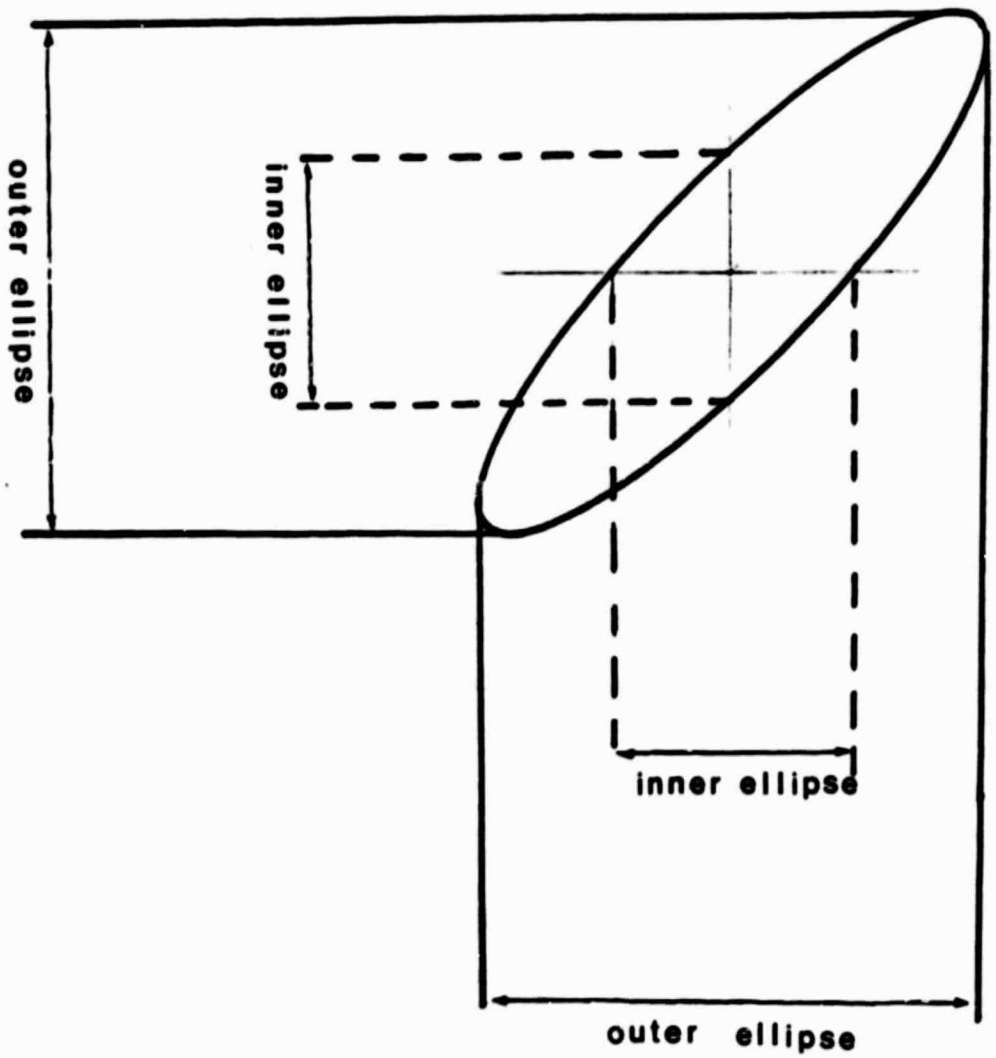
Figure 1 shows the situation for a two-channel to one-channel reduction. In this highly correlated example the inner "ellipse" is much smaller than the outer ellipse. This is because the dependent ellipse is a cross-section of the covariance rather than the full range as given by the independent ellipse.

Figure 2 shows some examples of these ellipses for a CLASSY run on real Landsat data (the two-ellipse system was coded by Lockheed electronics.) As can be seen, not a great deal of the variation is removed by using the dependent matrices. This means that little of the variation in one pair of channels can be explained or predicted by the variation in the other channels. To the extent that the passes are uncorrelated, improved classification becomes a statistical accumulation of data from several passes, rather than a coherent picture of ground color variation.

Figure 1.

Projection from

2 Dimensions to 1 Dimension



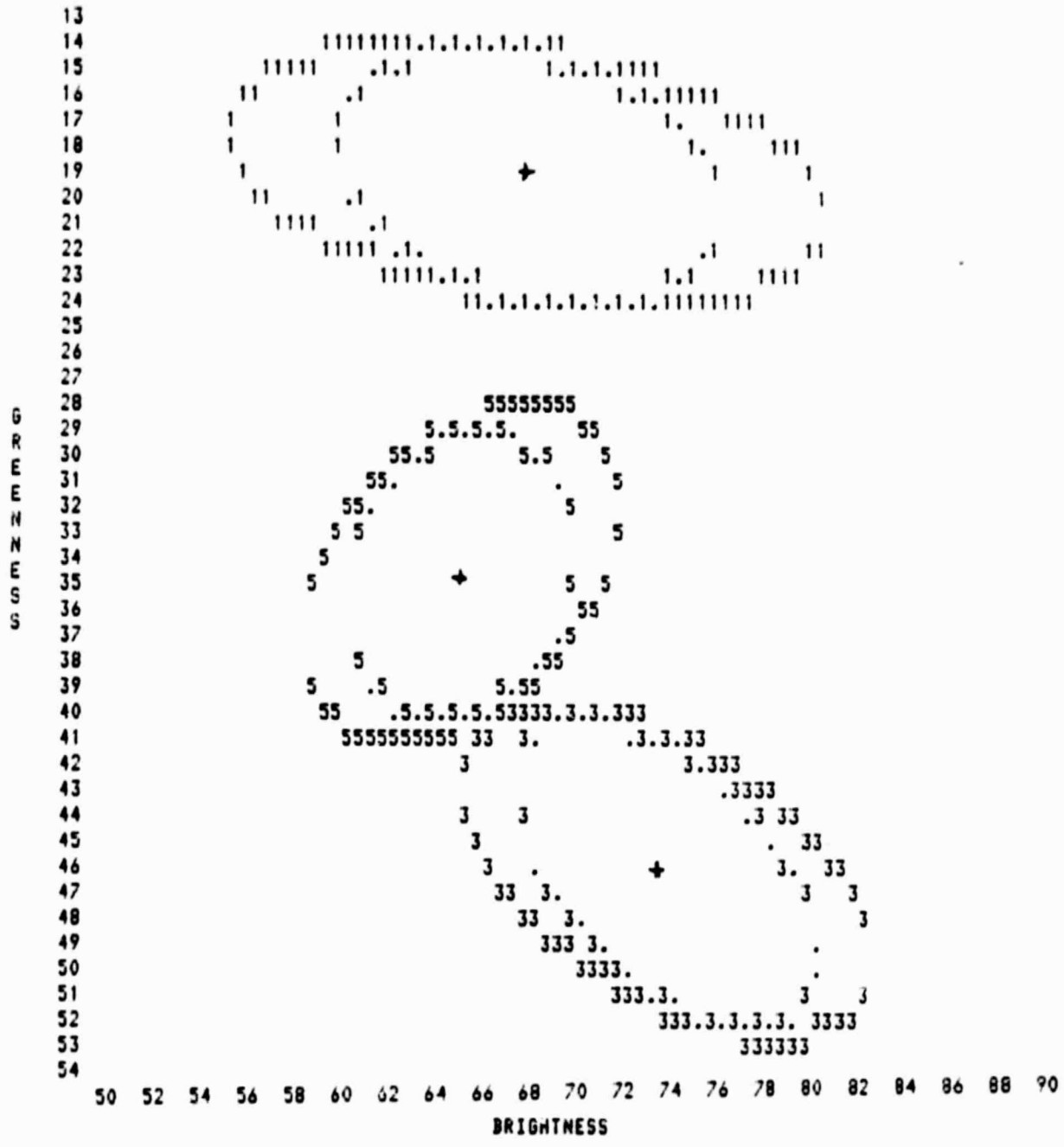


Figure 2. Brightness-greenness plot for segment 1805 (3 acquistitions). The seperation between inner and outer ellipses is relatively large in this example (cluster 12), so that most of the variation can be considered to be acquisition independent.

APPENDIX A

DESCRIPTION OF PRINTOUT ITEMS

The names correspond to those on the annotated printout.

1) Cluster printout--printed occasionally when a cluster is adjusted, and for each cluster at the end of an iteration. This printout is controlled by several switches, which delete certain portions of the output, or control line length (80 or 132 columns). The full printout is given here, but usually the minimum output is used.

The printout is described here for the 132 column format; the 80 column format is labeled alike and can be understood from this description.

Cluster printout
(132 column format,
no substitutions)

```
0  8 / ( 7692)  PROP 0.24129   VOL    151.35 *   4.05  DEV 1.40  SENS  4.46   LEV1292 SUPER   0( 3)   ILE (0.926) 78 12  7 -1  2
     ADJ: W    543.9     CURRENT: P    0.00 =CUM    0.00  COND    0.00    PROP: REL 0.241  CIN  481.4  DEN 1992.3 =(W PAR  7692.0
        ID  11408          (R**2)       2.49    PRIOR 1.000  PASS    0.94       (OLD)  0.251       224.0      908.1   -CTOT  5699.7)
   TRACE: OW     261.9  VOLIN 0.3720E-20  RT 0.6099E-10  DCON 57.080  KL  647  LSUPER  119  LSUBS     0  LINK  881(  9)  NSYMB  1
    MEAN  CHAN   COVAR:          ////  W=      544.1      ------  NO SUBS  ------
   25.82     1    1.59  0.16  1.41  2.13
   23.55     7    0.16  2.30  0.61  0.67
   20.11    10    1.41  0.61  6.81  7.73
   20.46    14    2.13  0.67  7.73  9.75
     OLD MEAN    OLD COVAR:
   25.86     1    1.58  0.25  1.16  1.84
   23.59     7    0.25  2.56  0.00  0.07
   19.89    10    1.16  0.00  7.37  8.37
   20.25    14    1.84  0.07  8.37 10.34
     SKEW  CHAN  KURTOSIS - (DIM+2)*COVAR:
   -0.76     1    0.51 -1.75  2.39  2.70
    0.49     7   -1.75 -2.47  0.23 -1.10
   -0.13    10    2.39  0.23 -0.02 -1.21
   -0.89    14    2.70 -1.10 -1.21 -1.53
```

a) Main line (always printed).

Cluster number--the number assigned by CLASSY to the cluster. This number is unique to the cluster and consistent throughout a run.

Cluster symbol--the one-character symbol assigned to the cluster for mapping, etc. It changes from iteration to iteration.

Total points clock--total number of points so far processed during this CLASSY run.

Absolute proportion--the proportion assigned to the cluster relative to the top of the tree.

Volume--the cluster volume $((2\pi)^{d/2}(\det \Sigma)^{1/2})$.

Volume factor--the ratio of the cluster volume to its diagonal part (i.e., to the volume with all correlations set to 0) $((\det \operatorname{diag} \Sigma)^{1/2}/(\det \Sigma)^{1/2})$.

Typical deviation--typical standard deviation in one channel (actually the geometric mean of all the deviations) $((\det \Sigma)^{1/d})$.

Sensitivity--the sensitivity of the volume of the cluster to the addition of a small component of the unit matrix $(\partial \det(\Sigma+\alpha 1)/\partial \alpha|_{\alpha=0} = \operatorname{tr} \Sigma^{-1} \det \Sigma)$.

Level--in cluster tree printout, the level of the cluster (0 for the nominal top node);

in other cluster printout, a marker for the source of the output.

Super cluster--the cluster number of the cluster above this one.

Which in super list--the position of this cluster in the list of subclusters of its super cluster.

Average of good probabilities--the average of the a posteriori probability of all points which assign an a posteriori probability of 80% or greater to this cluster.

Pentile percents--(5 numbers). The percentage of the weight of the cluster coming from points which give a posteriori probabilities to this cluster in the five ranges (80-100%, 60-80%, 40-60%, 20-40%, 0-20%). The points in the first range are those used in calculating the average of good probabilities above. These measures give an indication of how strongly this cluster may overlap others. The percentages may not add up to 100 due to rounding error.

```
0
6 .  (    cluster  number
          cluster symbol

4898)  PROP 0.12459    absolute proportion

VOL    54.48 *         volume

       4.85            volume factor

DEV 1.08               typical deviation

SENS 7.87              sensitivity

LEV  1                 level

SUPER  0 ( 2)          super cluster
                       which in super list

ILE (0.904) 29 30 27   average of good probabilities
            9          pentile percents
            4
```

.8-1.
.6-.8
.4-.6
.2-.4
0-.2

b) Internal values (printout controlled by INTPR):

First line:

Adjust weight--(WADJ). Weight at which cluster will be adjusted.

Current probability--(PST). Last probability assigned to this cluster and its subclusters.

Cumulated probability--(CUM). The last (total) probability assigned to this clusters' subclusters.

Conditional probability--(PCOND). The last probability assigned to this cluster itself.

Relative proportion--(PROP). The proportion of this cluster relative to its sibling clusters (or to the proportion of the parent cluster).

Proportion numerator--(CIN). The proportions are kept in the form A/B. This is the A factor.

Proportion denominator--this is the B portion (see above).

Weight of parent--the current weight of the parent cluster.

ADJ: W 420.4 CURRENT: P 0.00 =CUM 0.01 COND 0.01 PROP: REL 0.125 CIN 291.5 DEN 2334.2 =(W PAR 4898.0

adjust weight

current probability

cumulated probability

conditional probability

relative proportion

proportion numerator

proportion denominator

weight of parent

Second line:

Full cycle adjust--(IDADJ). The point at which the cluster will have "seen" all the data and thus require adjustment.

Last R**2--(DISS). The last distance squared of a point from the cluster.

Cumulated prior probability--(PRIRCM). The last sum of the subcluster prior probabilities assigned to this cluster. (Near 1.)

Passed probability--(PPASS). The probability "passed" to this cluster from its parent cluster. A temporary.

Old proportion--(OPROP). The value of the relative proportion (c.f.) at the last ADJUST.

Old proportion numerator--(OCIN). The value of the proportion numerator (c.f.) at the last ADJUST.

Old proportion denominator--(ODEN). The proportion denominator (c.f.) at the last ADJUST.

Weight not entering denominator--(CTOT). The part of the weight of parent (c.f.) not entering the proportion denominator (c.f.).

ID	8345	full cycle adjust
(R**2)	1.69	last R**2
PRIOR	0.996	cumulated prior probability
PASS	0.81	passed probability
(OLD)	0.126	old proportion
	146.2	old proportion numerator
	1142.0	old proportion denominator
-CIOT	2563.0)	weight not entering denominator

c) Trace quantities (controlled by TRACPR).

Old weight--(OW). The cluster weight (c.f.) at its last ADJUST.

Volin--(VOLIN). Cluster volume with extra factors.

Volrt--(VOLRT). The square root of VOLIN.

Dcon--(DCON). A term used to offset various volume factors; effective volumetric factor for the normal distribution is VOLRT * EXP (DCON/2).

Cluster index--(KL). The actual position of the cluster in memory.

Super index--(LSUPER). The actual position of the super cluster in memory.

Subcluster index--(LSUBS). The actual position of the first subcluster in memory.

Sibling index--(LINK). The actual position in memory of the next cluster on the sibling list.

Sibling number--the cluster number of the next sibling.

Symbol number--(NSYMB). The number of the graphic (see "cluster symbol") used for this cluster.

TRACE: OU 200.2 VOLIN 0.3189E-21 RT 0.1786E-10 DCON 57.493 KL 881 LSUPER 119 LSUBS 959 LINK 803(7) NSYMB 2

old weight

volin

volrt

dcon

cluster index

super index

subcluster index

sibling index

sibling number

symbol number

d) Subcluster data (always printed). (This line also includes the captions for the mean and covariance.)

Weight--(W). The weight assigned to this cluster.

Variant 1--"NO SUBS"--means that this cluster has no subclusters.

Variant 2--(subcluster information):

Splitting--(SPFAC). The current value of the likelihood ratio between this cluster and its subclusters, including the bias from the priors. A sufficiently postive value will cause the cluster to be separated, a sufficiently negative value will cause the subclusters to be eliminated.

Specific splitting--(SPFAC/DW). The average contribution of each point to the splitting.

Difference in models--(from PQRAT). The average (RMS) difference between this cluster and its subclusters. When this difference is small, it is unlikely that the subclusters will ever give a higher likelihood than this parent cluster.

Number of subclusters--the number of subclusters this cluster has.

Subcluster list--the cluster number of each subcluster of this cluster.

Subcluster data
variant 2-subclusters exist
(see first sheet for an
example of variant 1.)

MEAN CHAN COVAR: //// W= 374.9 SPLIT -31.73 (-.130) RMS 0.231 2 SUBS 10 11

weight

splitting

specific splitting

difference in models

number of subclusters

subcluster list.

e) Mean and covariance (controlled by COVPR).

Mean--the cluster mean vector.

Chan--the alphabetic or numeric label for each channel.

Covar--the cluster covariance matrix.

f) Old mean and old covariance (controlled by OLDPR).

The old values of the mean and covariance, as described above.

g) Skewness and Kurtosis (controlled by KRTPR).

Skew--the skewness vector for the cluster.

Chan--as in mean and covariance.

Kurtosis-(d+2)* covar--the kurtosis matrix for the cluster with part of the covariance matrix deducted to make the expectation value zero for a true normal distribution.

2) Adjust printout--printed each time a cluster is adjusted.

Total points clock--total number of points so far processed during this CLASSY run.

Cluster--the number assigned by CLASSY to the cluster being adjusted. This number is unique to the cluster and consistent throughout a run.

Absolute proportion--the proportion assigned to the cluster relative to the top of the tree.

Current weight--(W(KL)). The current weight assigned to the cluster (before adjustment).

Old weight--(OW(KL)). The weight assigned to the cluster at its last adjustment.

Motion of mean in standard deviations--the total motion of the mean measured in terms of the variance; the number of standard deviations the mean has moved since the last iteration.

Square of the change in covariance--($tr(\Sigma^{-1}\Delta\Sigma\Sigma^{-1}\Delta\Sigma)$). The total squared change in the covariance of the cluster since the last iteration.

Variant 1 (Used for non-split clusters):

Trace test (statistic)--the value of the trace of the kurtosis test at this adjustment ($Tr(\Sigma^{-}K)$).

Trace test (test compare)--the difference between the trace test value and the test success threshold; minus means split the cluster.

Skewness test (statistic)--the square of the skewness at this adjustment ($S^t\Sigma^-S$).

Skewness test (test compare)--the difference between the skewness test value and the test success threshold; minus means split the cluster.

Kurtosis test (statistic)--the square of the traceless part of the kurtosis at this adjustment ($Tr(\Sigma^-K_o\Sigma^-K_o)$, where $K_o = K-1^\#(trK/d)$).

Kurtosis test (test compare)--the difference between the kurtosis test value and the test success threshold; minus means split the cluster.

Adjust printout
(variant 1)

```
( 6240)ADJ  7: 0.1780 W  503.5/  241.4  V 0.02931  0.48322 TR  -17.4:  5030.4 S  174.6:  565.5 K  548.7:  2455.7
( 6377)ADJ  5: 0.1848 W  544.3/  261.9  V 0.03673  0.20979 TR   21.2:  4715.7 S  313.9:  403.2 K 1542.3:  1368.6
( 6510)ADJ  8: 0.2408 W  503.6/  241.6  V 0.06189  7.02241 TR   -29.0:  4489.5 S  103.1:  637.1 K  518.2:  2486.9
( 6659)ADJ  9: 0.1463 W  461.2/  220.3  V 0.13217  2.02191 $ -17.520  RMS    0.1
( 6942)ADJ 10: 0.1313 W  421.3/  200.6  V 0.03255  2.74866 TR   -22.8:  5244.7 S  234.7:  565.5 K   964.4:  2283.9
( 6950)ADJ 12: 0.1536 W  240.4/   40.0  V 0.12180  3.18965 TR   -20.8:  5610.1 S  495.8:  342.8 K  1154.0:  2250.5
( 7692)ADJ  8: 0.2413 W  544.1/  261.9  V 0.05876  4.27041 TR    -3.9:  5152.3 S  222.6:  494.7 K  2630.8:   281.3
```

total points clock

cluster

absolute proportion

current weight

old weight

motion of the mean in
standard deviations

square of the change in
covariance

trace test (statistic)

trace test (test compare)

showness test (statistic)

showness test (test compare)

kurtosis test (statistic)

kurtosis test (test compare)

Variant 2 (Used for split clusters):

Splitting--(SPFAC). The current value of the likelihood ratio between this cluster and its subclusters, including the bias from the priors. A sufficiently positive value will cause the cluster to be separated (from PQRAT).

Difference in models--(from PQRAT). The average (RMS) difference between this cluster and its subclusters.

3) Cluster tree printout (PRTREE).

Under several circumstances CLASSY prints out the current cluster tree. The tree consists of one item per cluster, with subcluster items one line below the parent cluster, directly underneath it and proceeding to the right (see figure).

Each item consists of a two-digit cluster number followed by a code character "-" or "/", followed by two digits representing the cluster's percentage proportion. The code character represents a cluster which is not ready to be separated ("-"), and one which is ready ("/").

Examples:

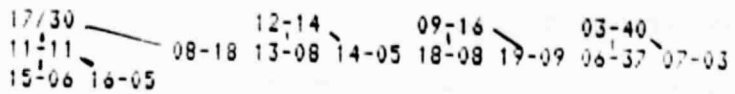
01-17 cluster 1, not split, 17% of total proportion

Adjust variant 2
(split clusters)

(9645)ADJ 9: 0.1788 W 543.6/ 261.5 V 0.12575 0.05238 $ -15.617 RMS 0.1

splitting

difference in models.



Tree printout example. Lines have been drawn on the printout to show how the tree structure is expressed. The top left cluster should be seperated.

```
75/32                   32-12       63/07       23-06       69/13
03-27       26-04       61-05 62-07 22-05 14-02 64-04 65-02 50-06 19-07
47-11 48-16 55-02 56-02                                           71-01

------------------------
            51-07 27-05 33-04 29-09       16-05
                              73-05 74-04
72-05
77-04 78-01
```

Extended example. (Four level tree). The items below the dashed line belong to the right of the upper part of the tree.

13/85 cluster 13, should be split, 85% of total proportion

4) Iteration printout line.

Total points clock--total number of points processed by CLASSY this run.

Number of iterations--the number of times CLASSY has processed the complete data sample.

Time used--the total time used by CLASSY since program start (including some time before iteration 0). The time is as provided by the system, converted to seconds.

5) Structural changes lines.

These are lines printed when a structural change is made in the cluster tree, such as splitting, separating, or joining clusters. Each of these lines is started by the total number of points so far processed, followed by an indication of the line time (see below). The cluster tree may be printed out following one of these structural changes.

a) HAVE SPLIT--when a cluster is hypothetically split.

--cluster number.

Weight--current weight of the split cluster.

Subs--the cluster numbers of the two new subclusters.

Iter--the number of iterations required by SPLIT estimating routine.

b) SEPARATE--when a cluster is to be separated (eliminated in favor of its subclusters).

--cluster number.

Super--cluster number of the parent cluster.

Subs--cluster number of the first subcluster.

Spfac--the splitting (c.f.) of the cluster.

Weight--the current weight of the cluster being split.

c) JOINING--when two clusters are joined as a hypothesis - AND - the cluster numbers of the two clusters being joined.

TO GET--the newly created join cluster.

d) ELIMINATE-when a single cluster is being removed from the cluster tree.

--the cluster to be eliminated.

Link--the first sibling of this cluster.

Lsubs--the first subcluster of this cluster.

Lsuper--the parent cluster of this cluster.

e) SUB ELIM--when all the subclusters of a cluster are to be eliminated.

--the cluster whose subclusters are to be eliminated.

Splitting--the splitting (c.f.) of the current cluster.

+ -the threshold for subeliminating a cluster from splitting alone.

Pqrat--the RMS difference between this cluster and the sum of its subclusters (a low PQRAT is the usual reason for SUB ELIMing a cluster).