May 31, 1991

A New Option for CLUST Allowing Retrieval of an Optimal Statistics File

by Michael Bellow and Martin Ozga

The following is a description of a proposed addition to CLUST, the PEDITOR program that invokes the ISODATA clustering algorithm to create cover signatures. With the current version, the user usually accepts the statistics file created at the end of program execution. That statistics file represents the final clustering of the data, where the conditions for program termination have been satisfied. However, the user can also obtain statistics files created after each iteration of the program run. It may be the case that a clustering formed at an earlier stage is better than the final clustering, according to some efficiency measure. Currently, the user could examine a number of intermediate statistics files and select the one that seems "best", but that can be a tedious process and is rarely done. The proposed modification would give the user the option of retrieving a statistics file associated with a specific intermediate clustering, found by the program to optimize a certain criterion.

The criterion to be optimized is the Calinski-Harabasz index, also known as the F-statistic. This measure is currently computed by CLUST after each cluster merge or split, but has no effect on the output of the program. The formula for the F-statistic is as follows:

$$F = [(n-c)/(c-1)] \sum_{i=1}^{c} [n_i |\bar{\mathbf{z}}_i - \bar{\mathbf{z}}|^2] / \sum_{i=1}^{c} \sum_{j=1}^{n_i} |\mathbf{z}_{ij} - \bar{\mathbf{z}}_i|^2]$$

where:

n = number of pixels in data set

c = number of clusters

n_i = number of pixels in cluster i ($i=1,..,c$)

$\mathbf{z}_{ij}$ = vector of spectral values for cluster i, pixel j ($j=1,...,n_i$)

$\bar{\mathbf{z}}_i$ = mean vector for pixels in cluster i

$\bar{\mathbf{z}}$ = mean vector for all pixels in data set

The F-statistic can be regarded as a ratio of between-cluster variability to within-cluster variability. It has been found to be an excellent measure of clustering effectiveness.

In the proposed new version of CLUST, F would still be computed after each cluster merge or split. However, the program would now keep track of which clustering produced the highest value of F thus far. The mean vectors and covariance matrices corresponding to that clustering would be stored, and only replaced if a subsequent clustering produced a higher value of F. Thus, a statistics file corresponding to the clustering that resulted in the highest value of F would be available at the conclusion of the program run, in addition to any other statistics files requested by the user. This special file will be referred to as the F-optimal statistics file. The user would be prompted at the outset as to whether the F-optimal statistics file should be created, and if so, the filename to assign to it. At the conclusion, the program would tell the user which iteration gave the highest value of F. If F was maximized at the final iteration, then the F-optimal and final statistics files would be identical, and the user would be notified of that fact.

This enhancement could be available for use in this year's Delta Remote Sensing Project. The F-optimal statistics file will be tested on various data sets to determine whether or not it can produce better crop acreage estimates than the final statistics file.