1-1-1976

Number of Signatures Necessary for Accurate Classification

W. Richardson

A. Pentland

R. Crane

H. Horwitz

Reprinted from

Symposium on

Machine Processing of

Remotely Sensed Data

June 29 - July 1, 1976

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
76CH1103-1 MPRSD

NUMBER OF SIGNATURES NECESSARY FOR ACCURATE CLASSIFICATION*

W. Richardson, A. Pentland, R. Crane and H. Horwitz

Environmental Research Institute of Michigan
Ann Arbor, Michigan

ABSTRACT

This paper presents a procedure for determining the number of signatures to use in classifying multispectral scanner data. A large initial set of signatures is obtained by clustering the training points within each category (such as "wheat" or "other") to be recognized. These clusters are then combined into broader signatures by a program that considers each pair of signatures within a category, combines the best pair in the light of certain criteria, saves the combined signature and repeats the procedure until there is one signature for each category. The result is a collection of sets of signatures, one set for each number between the number of initial clusters and the number of categories. With the aid of statistics such as an estimate of the probability of misclassification between categories, the user can choose the smallest set satisfying his requirements for classification accuracy.

INTRODUCTION

Computer processing of multispectral scanner data as a means for measuring the earth's resources depends for its success on the definition of spectral classes, i.e. signatures, corresponding to materials to be recognized and backgrounds in the scene. Clustering techniques for defining these classes have been used with success, but have left unresolved the question of how many signatures to define. When classes are too few, they are so broad they overlap, resulting in unnecessarily large classification errors, while too many classes increase classification costs and cause difficulty in matching spectral classes with materials in the scene.†

A procedure at ERIM is to cluster the points into small spectral classes by a processing module CLUSTR and then to combine the clusters into larger signatures by a program GROUP. CLUSTR uses a relatively simple algorithm because it is applied to every data point. The number of small clusters it produces is an upper bound on the number of significant modes in the data space. GROUP, working on the set of clusters, much fewer in number than the data points, can take time to be careful. It uses covariance information and before each step of combining a pair of clusters, considers all possible pairs in the light of certain criteria. At the end of a run of GROUP, the analyst has a choice of sets of combined signatures, each set being the best choice given the number of signatures. He also is provided tables and graphs to help decide how many signatures to use.

DESCRIPTION OF A TECHNIQUE FOR DETERMINING THE NUMBER OF SIGNATURES

Our procedure for reducing the number of signatures combines signatures within categories. In principle, the procedure can be applied to any number of categories from one on up. The present implementation, program GROUP, requires two, which we name for definiteness "wheat" and "other". Both categories are treated the same way.

The procedure is summarized by the following steps:

A. Compute for each pair of signatures (clusters) within each category up to five measures of inter-signature distance.
   1. Distance based on a combined covariance matrix.
   2. Determinant of the combined covariance matrix.
   3. Trace of the combined covariance matrix.
   4. Probability of misclassification between the pair.

*Support for this research was provided by NASA contract NAS9-14123 with the Earth Observation Division, Johnson Space Center.

†When clustering is unsupervised, the difficulty of identifying spectral classes increases with the number of classes and with the smallness of the classes. When clustering is supervised and recognition is extended from training to test areas, test classes may appear between training modes and thus be recognized better by broader signatures.

5. Increase in the probability of misclassification between categories (we describe these measures more fully below).

B. For each distance criterion selected, rank every pair of signatures and then combine the pair with the smallest weighted sum of ranks. Punch or otherwise save this combined signature.

C. Compute descriptive statistics such as the following:

1. The average pairwise probability of misclassification between categories.
2. The maximum determinant scaled to compare with distance measurement.
3. The maximum trace scaled to compare with distance measurement.

D. Compute the observed probability of misclassification by classifying the training data from which the signatures were extracted. The classification uses the current set of signatures.

E. Repeat steps (A)-(E) until only one signature per category remains.

F. Display the statistics computed in (C) and (D) in a table and graphs.

From these displays, the user decides how many signatures are right for the multispectral recognition problem being attacked. The procedure has minimized the use of qualitative judgement by selecting from the myriad of possible signature combinations a few likely candidates and providing information to aid in the qualitative choice among the few. When the user has made his choice, he assembles the chosen set of signatures from among those saved.

The input to the program GROUP is a number of "wheat" and "other" signatures. Each signature is in the form of a mean vector and a covariance matrix, parameters that are assumed to specify a multivariate normal distribution of data vectors from the material the signature represents. Signatures: computed from fewer then five points are not accepted by the program.

The program provides five criteria for combining groups. Any of these criteria or any subset of them may be used. If two or more criteria are chosen, then the possible pairs of signatures to be combined are ranked according to each criterion and the pair with the smallest weighted sum of ranks is chosen. In that way the pair of signatures combined is the one most generally in harmony with the criteria selected. The five criteria are as follows:

1. An average covariance matrix A_W for the wheat signatures and one A_O for the other are calculated. The pair of signatures combined is the one with the smallest squared distance.

$$(\mu_2-\mu_1)^T A_W^{-1} (\mu_2-\mu_1)$$

or

$$(\mu_2-\mu_1)^T A_O^{-1} (\mu_2-\mu_1)$$

depending on whether the pair is wheat or other. It is essentially the square of the usual distance between the means but with the scale modified by the inverse of the average covariance matrix.

2. The determinant of the combined covariance matrix. The combined covariance matrix of the training set is the covariance matrix of the union of the two sets except that each set may be given an arbitrary weight. If the weights are proportional to the number of pixels used in calculating the signature, then the combined signature is identical to the signature calculated from all points of the two sets. If the two sets have circular signatures far apart, for example, the combined covariance matrix is long and thin whereas the average covariance matrix is circular. The determinant is the product of the eigenvalues, in other words the product of the variances in the axial directions of the ellipsoidal distribution. The bigger the determinant, the more spread out the distribution.

3. The trace of the combined covariance matrix. The trace is the sum of the diagonal elements, namely the variances, and is also the sum of the eigenvalues. It is invariant under a rotation of the space. Like the determinant, it is a measure of how spread out the combined distribution is.

4. The squared Mahalanobis distance

$$D_{ij}^2 = (\mu_j-\mu_i)^T \left(\frac{R_i+R_j}{2}\right)^{-1} (\mu_j-\mu_i)$$

This is the same distance as criterion 1 except that the covariance matrix modifying the distance is the average of the two covariance matrices of the pair rather than the average of all the covariance matrices in the category. The difficulty with this criterion is that the more spread out a signature is, the smaller is its distance to any other signature. The criterion thus tends to encourage large variances rather than to hold them down. This criterion is included in the program largely by tradition. Our former method of combining signatures was to make a table of the probability of misclassification (p. of m.) defined for each pair of signatures as

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{1}{2}D_{ij}} e^{-\frac{t^2}{2}}\, dt \qquad (1)$$

and then to group the signatures intuitively as suggested by the table. Expression (1) is an estimate of the probability of deciding on signature j, given that the distribution is really represented by signature i or vice versa -- an estimate that becomes exact[1] if the covariance matrices of signature i and signature j are both equal to $(R_i+R_j)/2$.

5. The average pairwise wheat-other p. of m. For each wheat-other pair, the Mahalanobis distance D computed and from that the p. of m. as in criterion 4. The criterion is a weighted average of these

pairwise p. of m.'s.

The wheat signatures start out with weights α_i that add to 1 and the other signatures with weights β_j that add to 1. The weights are initially equal but may be set in the control input. When two signatures are combined, their weights are added. The average pairwise wheat-other p. of m. is

$$\sum_{\text{wheat } i} \sum_{\text{other } j} \alpha_i \beta_j \text{ p. of m. } (i,j)$$

This number is printed at every step of the program and is one of the ways the user decides when the combining has gone far enough.

There is a case to be made for using only criterion 5 for combining. After all, is not the ultimate goal to minimize the probability of misclassification? The reason the distance criteria are also included is because experience shows that the training data seldom fully represent the data to be processed. If two distant signatures are combined because such a combination does not adversely affect the p. of m. of the training data, the combination might swallow up competing signatures in the test data. The safest plan is to use one or more distance criteria along with criterion 5 so that the two signatures to be combined will be a good choice both from the standpoint of distance and p. of m.

The criteria can be weighted so that the p. of m. criterion 5 gets half the weight and the distance criteria divide the other half. At the end of the run, a summary table is printed, each row of which corresponds to the number of signatures, so that the rows go from 2 to the original number of signatures. The columns refer to the criteria for the signature that was combined at that step and to other useful information. Digital plots of any requested colums of the table are given. The columns of the table we have found most useful are

1. Criterion 5, the average pairwise wheat-other p. of m.

2. The (2n)th root of the maximum covariance determinant. The determinant is the product of the eigenvalues. Hence, the nth root of the determinant is the geometric mean of the eigenvalues. An eigenvalue is the variance of the distribution in the direction of an axis of the ellipsoid. The variance is a squared quantity. Its square root, the standard deviation, is in units of Enchidean distance. Thus the (2n)th root of the covariance determinant is an average standard deviation of the distribution, a measure of how spread out the distribution is. The maximum of these values shows how spread out the combined signatures are getting.

3. The square root of 1/n(maximum covariance trace). The trace of a covariance matrix is the sum of the diagonal terms (the variances) and is also the sum of the eigenvalues. Thus the trace/n is the arithmetic mean of the eigenvalues, an average variance, and its square root is therefore an average standard deviation of the distribution. It is also a measure of how spread out the distribution is. The only difference between this measure and the previous one is that the arithmetic rather than the geometric mean of the eigenvalues is taken.

4. The average pairwise p. of m. (as in column 1) multiplied by one half the number of signatures in the set. The purpose of the multiplication is to make the average pairwise p. of m. more closely approximate the overall p. of m. Suppose for example there are three "other" signatures and one wheat signature. There are three wheat-other pairwise p. of m.'s, $p(W_1O_1)$, $p(W_1O_2)$, and $p(W_1O_3)$. Prob{other|wheat} is more closely approximated by $p(W_1O_1)+p(W_1O_2)+p(W_1O_3)$ than by 1/3 this amount. But prob{wheat|other} = $1/3\ p(W_1O_1)+1/3\ p(W_1O_2)+1/3\ p(W_1O_3)$ because the probability of choosing O_1 is 1/3 and the subsequent probability of deciding on wheat is $p(W_1O_1)$, and similarly for O_2 and O_3. Thus, the average of prob{other|wheat} and prob{wheat|other} is approximated by

$$\frac{2}{3}\left[p(W_1O_1)+p(W_1O_2)+p(W_1O_3)\right]$$

which is the average pairwise p. of m. times one half the number of signatures in the set. The figure we have calculated is an overestimate of the p. of m. just as the average pairwise p. of m. is an underestimate so columns 1 and 4 bound the true theoretical p. of m. between categories.

5. The observed p. of m. calculated by classifying the training points using the current set of signatures. This empirical measure of performance of the signature set complements the theoretical measures.

APPLICATION OF THE TECHNIQUE

This process of clustering and GROUPing has been carried out on LANDSAT MSS data drawn from five agricultural sites in Kansas and Texas. For each site, training fields were selected at random and then divided into the two categories "wheat" and "other". CLUSTR was then run in a supervised mode to provide several signatures (clusters) for each category, and these signatures were used as input to GROUP. The statistics produced by GROUP as the number of signatures was reduced to one per category were displayed in digital plots such as those in Figures 1 - 6.

The first four figures typify the plots of maximum determinant, maximum trace, average pairwise p. of m. and this last measure multiplied by one-half the number of signatures. These measures tend to behave as expected, decreasing rapidly at first as the number of signatures increases and then flattening out. The typical backward slant of the curve for pairwise p. of m. times factor (Figure 4) probably indicates that the factor overcompensates in its task of making pairwise p. of m. a better estimate of the overall p. of m. Possibly a factor half as large would be a good compromise between the two bounds.

The observed p. of m. on occasion follows the pattern of the other measures (Figure 5) but when the number of points misclassified is small, the observed p. of m. jumps about randomly. Figure 6 shows a case where a maximum of 8 points were

misclassified. These misclassified points may reflect the unpredictable behavior of clusters too small to be accepted by GROUP or weakness in the original definition of the clusters.

CONCLUSIONS

Starting with either field-by-field signatures or clusters, the question of how many and which signatures to use is often decided by guesswork. The GROUP procedure attempts to solve this problem by providing the analyst with the most likely sets of combined signatures and the information needed to choose from among them.

The rule used by GROUP in choosing which signatures to combine is constructed according to two principles: first, signatures chosen to be combined should be as close to each other as possible; second, the combining of these signatures should keep the probability of misclassification between categories as small as possible. GROUP then provides the analyst with sufficient information about its combining activities to allow him to choose from among the sets of signatures the one set which he believes represents the best compromise between cost and classification accuracy.

The GROUP procedure may also be used for investigating both practical and theoretical questions. Some of the investigations which might profitably employ GROUP include the relationship between theoretical and empirical measures of the probability of misclassification, the robustness of various schemes for signature selection; and the number of signatures normally needed to maintain accurate classification.

REFERENCE

1. W. A. Malila, R. B. Crane and W. Richardson, Discrimination Techniques Employing Both Reflective and Thermal Multispectral Signals, Report 31650-75-T, Environmental Research Institute of Michigan, Ann Arbor, Michigan, 1973, page 42, equation (9).

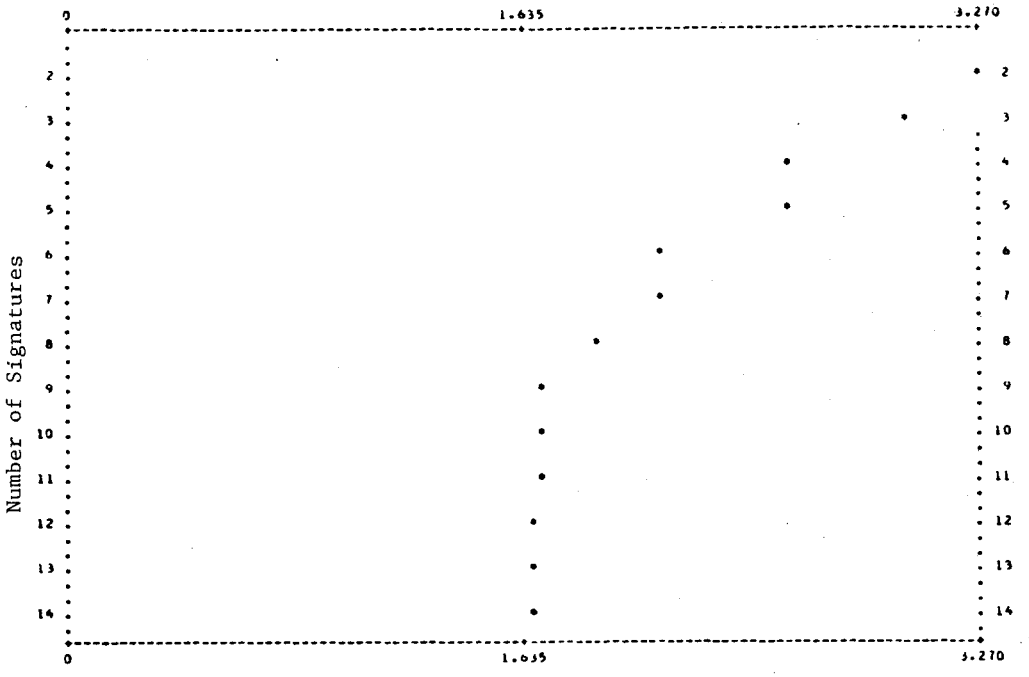

Figure 1. Maximum Determinant (Saline Site)

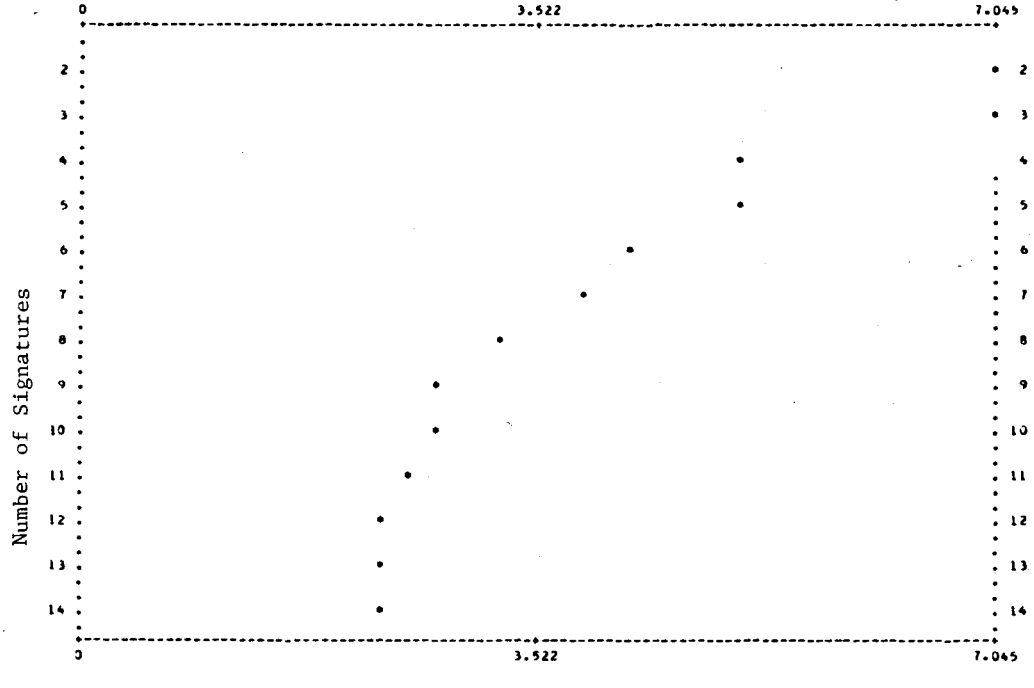

Figure 2. Maximum Trace (Saline Site)

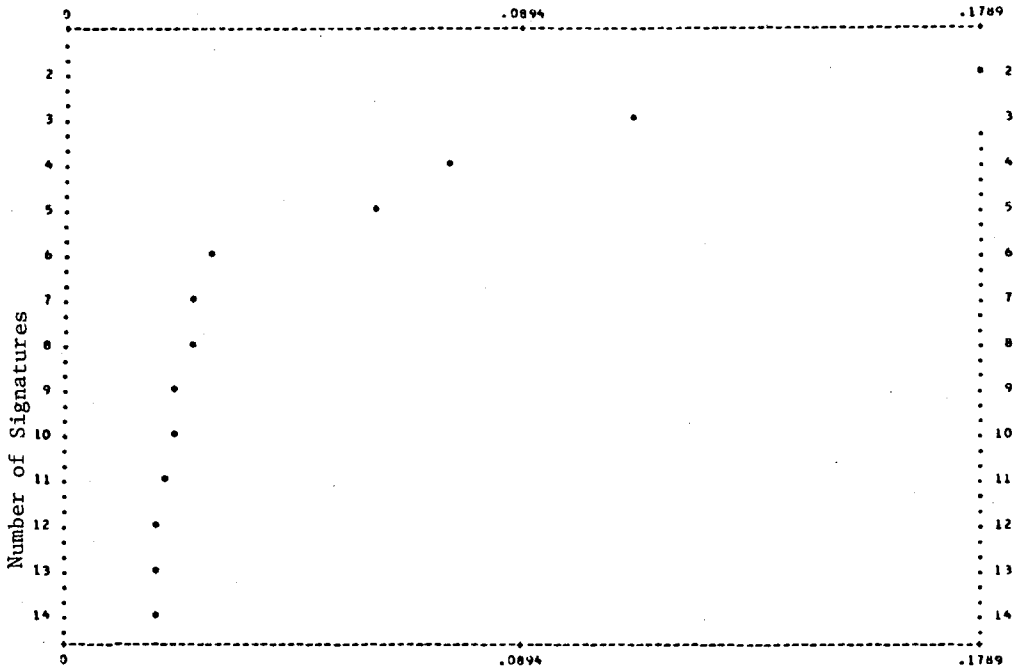

Figure 3. Average Pairwise Probability of Misclassification (Saline Site)

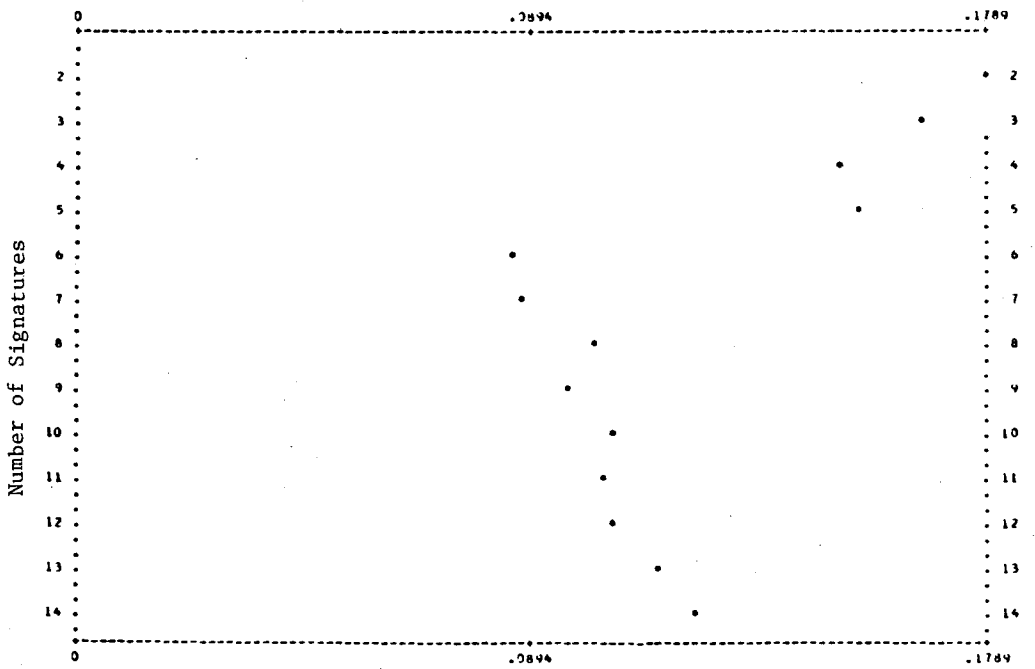

Figure 4. Pairwise Probability of Misclassification Times Factor (Saline Site)

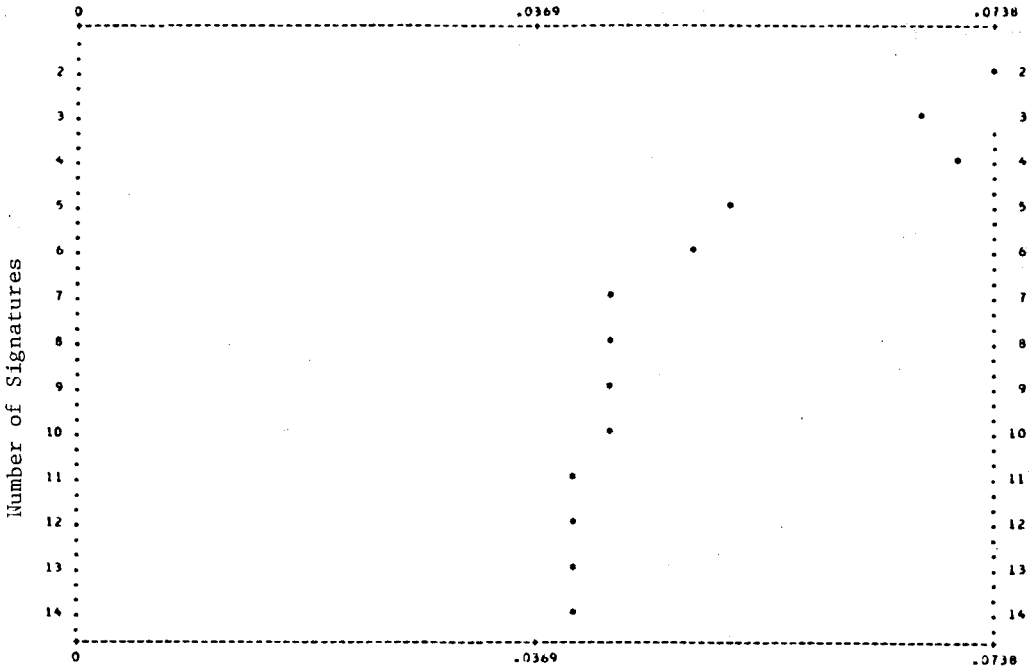

Figure 5. Observed Probability of Misclassification (Saline Site)

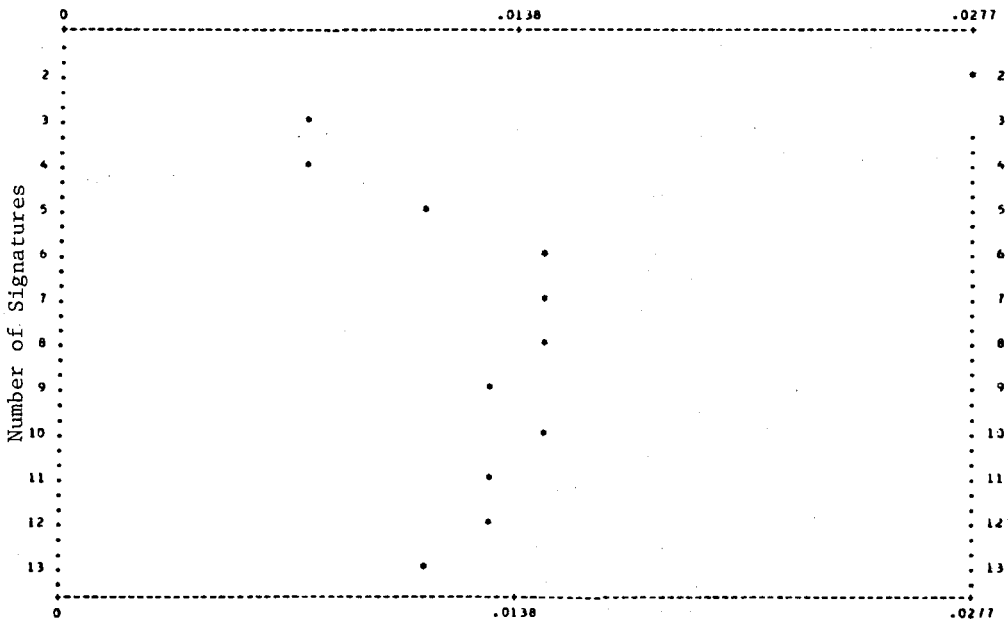

Figure 6. Observed Probability of Misclassification (Finney Site)