80-10132

AgRISTARS

SR-L0-00430
JSC-16345

NASA CR-
160584

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

Supporting Research

March 1980

TECHNICAL REPORT

LABEL IDENTIFICATION FROM STATISTICAL TABULATION (LIST) APPLICATION OF RIDIT ANALYSIS

T. B. Dennis

(E80-10132) LABEL IDENTIFICATION FROM
STATISTICAL TABULATION (LIST) (Lockheed
Engineering and Management) 18 p
HC A02/MF A01 CSCL 02C

N80-27764

Unclas
G3/43 00132

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Lyndon B. Johnson Space Center Houston, Texas 77058

1. Report No. JSC-16345; SR-L0-00430	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Label Identification from Statistical Tabulation (LIST) Application of RIDIT Analysis	5. Report Date March 1980 6. Performing Organization Code
7. Author(s) T. B. Dennis Lockheed Engineering and Management Services Company, Inc.	8. Performing Organization Report No. LEMSCO-14390 10. Work Unit No.
9. Performing Organization Name and Address Lockheed Engineering and Management Services Company, Inc. 1830 NASA Road 1 Houston, Texas 77058	11. Contract or Grant No. NAS 9-15800 13. Type of Report and Period Covered Technical Report
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Lyndon B. Johnson Space Center Houston, Texas 77058 Technical Monitor: J. D. Erickson	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

In Label Identification from Statistical Tabulation (LIST), analysts are asked to respond to a set of key questions about each pixel which is a candidate for labeling. The categorical responses are assigned ordinal values and combined with spectral features for use in a discriminant labeling process. In this paper, a mathematical method for assigning values to categorical data, called RIDIT's, is applied to the LIST data and studied as a replacement for the arbitrary assignment of values which was used previously.

17. Key Words (Suggested by Author(s)) RIDIT Categorical data Discriminant analysis	18. Distribution Statement

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 17	22. Price*

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

SR-LO-00430
JSC-16345

TECHNICAL REPORT

LABEL IDENTIFICATION FROM STATISTICAL TABULATION (LIST) APPLICATION OF RIDIT ANALYSIS

Job Order 73-302

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

PREPARED BY

T. B. Dennis

APPROVED BY

T. C. Minter

T. C. Minter, Supervisor
Techniques Development Section

J. E. Wainwright

J. E. Wainwright, Manager
Development and Evaluation Department

LOCKHEED ENGINEERING AND MANAGEMENT SERVICES COMPANY, INC.
Under Contract NASA 9-15800

For

Earth Observations Division

Space and Life Sciences Directorate

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
LYNDON B. JOHNSON SPACE CENTER
HOUSTON, TEXAS

March 1980

LEMSCO-14390

CONTENTS

Section Page

1. INTRODUCTION.. 1

2. THE RIDIT TRANSFORMATION.. 2

3. APPLICATION OF DISCRIMINANTS...................................... 4

4. RESULTS.. 5

5. CONCLUSIONS AND RECOMMENDATIONS.................................. 9

6. REFERENCES... 13

TABLES

Table Page

1 LANDSAT DATA SEGMENTS, COUNTIES, AND ACQUISITIONS WHICH COMPRISE THE TRANSITION YEAR LIST DATA.................. 6

2 A COMPARISON OF TEST AND TRAINING ACCURACIES OBTAINED WITH CONVENTIONAL AND TRANSFORMED DATA................ 7

3 ACCURACY OF USING THE MAJOR EIGENVECTOR FOR DISCRIMINATION ON INDEPENDENTLY TRANSFORMED DATA SETS.......... 8

4 ACCURACY OF USING THE MAJOR EIGENVECTOR FOR DISCRIMINATION ON UNIVERSALLY TRANSFORMED DATA SETS............ 10

5 LISTING OF MAJOR EIGENVECTORS GENERATED FOR FIVE SEGMENTS.. 11

FIGURES

Figure Page

1 Example of distributions of classes (0 = small grains, 1 = nonsmall grains) projected onto the major eigenvector..... 12

1. INTRODUCTION

Label Identification from Statistical Tabulation (LIST) is a semiautomated labeling technology developed during the Large Area Crop Inventory Experiment (LACIE) conducted at the National Aeronautics and Space Administration (NASA) Lyndon B. Johnson Space Center (JSC) in Houston, Texas. A complete description of this procedure can be found in reference 1. Basically, this procedure transforms analyst responses to a set of questions about a picture element (pixel) and the raw spectral data (four 4-by-1 vectors for each pixel) into a set of features and then applies a linear discriminant to these features to obtain a label of small grains or nonsmall grains. To develop the features used in LIST, the analyst responses to the pixel-level questions were assigned numerical values in such a way that smaller, assigned values were more characteristic of small grains. The automated features, derived from the spectral data, were also designed to be correlated with the probability of the presence of small grains. In this transformed feature space, the theoretically ideal discriminant boundary would be obtained with a discriminant vector, certain components of which were all nonnegative. However, in the LIST application, the discriminant vector did not have this desired property.

A literature search yielded a suggestion that the arbitrary assignment of values to each response be replaced by an assignment based on the distribution of all observed responses. (See Levine et al., ref. 2.) The transformation, called a RIDIT transformation, will be discussed in section 2. The name RIDIT was chosen by Bross (ref. 3) for this transformation. The first three letters are an acronym for "relative to an identified distribution," while the last two letters were chosen to form an analogy with "Probits" and "Logits." In addition to the RIDIT transformation, a discriminant is introduced in reference 2 to be used in conjunction with the transformation. This discriminant and others will be discussed in section 3. In section 4, the results of several discriminant analyses are shown, and conclusion and recommendations are given in section 5.

2. THE RIDIT TRANSFORMATION

The use of RIDIT's was introduced by Bross (ref. 3) and has appeared in many applications in the field of epidemiology (refs. 4 to 6). In reference 2, Levine et al. propose the use of RIDIT's in analyzing questionnaires. The use of RIDIT's applies when the difference in responses to a questionnaire is related to two classes within the data. The responses to each question must be ordered monotonically so that an increase in the response is related to the probability of membership in one of the two classes. The RIDIT transformation then replaces the arbitrary assignment of a value to a given response with an assignment of values based on the distribution of all responses to the question. To formulate the transformation, let X denote a response to the K^{th} question of the questionnaire. Let P_k $(Z < X)$ denote the probability that a response to question K is less than X and likewise let P_k $(Z > X)$ denote the probability that a response to question K is greater than X. The transformed value of the response X to question K is then given by the formula

$$T_k(X) = P_k(Z < X) - P_k(Z > X)$$

$$= F_k(X) - 1 + \lim_{E \to 0} F_k(X - E^2)$$

where F_k is the cumulative distribution for the responses to question K. In practice, empirical estimates of these distributions are used to obtain the RIDIT transformation.

Data which are transformed by this method have the following properties.

a. The expected value of the transformed responses to each question, $E[T_k(X)]$, is zero, and hence, if all responses to a given question are identical, the transformed value is zero and cannot contribute to a discriminant score. Another consequence of this property is the fact that zero lies between the means of the two transformed classes. To prove this assertion for the discrete case, suppose X takes on the n discrete values

$Z_1 < Z_2 < \cdots < Z_n$ with frequency $W_1, W_2, \cdots, W_n$, respectively. Then $T_k(X)$ takes on the values

$$\sum_{p=1}^{j-1} Wp - \sum_{p=j+1}^{n} Wp$$

with frequency W_j for $j = 1, 2, \cdots, n$. Thus, the expected value of T_kX is

$$\sum_{j=1}^{n} W_j\left(\sum_{p=1}^{j-1} W_p - \sum_{p=j+1}^{n} W_p\right)$$

$$= \sum_{j=1}^{n}\sum_{p=1}^{j-1} W_jW_p - \sum_{j=1}^{n}\sum_{p=j+1}^{n} W_jW_p$$

$$= \sum_{j=2}^{n}\sum_{p=1}^{j-1} W_jW_p - \sum_{p=2}^{n}\sum_{j=1}^{p-1} W_jW_p = 0$$

b. If two original sets of data with arbitrary value assignments are equivalent in the sense that there is an order-preserving function from one onto the other, then the transformed values of the two sets of responses will be identical. For example, if a data set contained responses of 0, 1, 2 for a particular question, and the value 1 was reassigned as 3, and the value 2 was reassigned as 7, then, after applying the RIDIT transformation, the two data sets (one coded 0, 1, 2 and the other coded 0, 3, 7) would yield the same RIDIT coding.

c. If all original responses, X, in a given interval $A < X < B$ are reassigned the single response $\frac{A + B}{2}$ (or any number between A and B), then the transformed values of responses outside the interval $A < X < B$ are not changed.

d. The RIDIT transformation is itself order-preserving.

Other properties of this transformation are given by Brockett and Levine (ref. 4).

Two methods are available for the application of a RIDIT transformation to the LIST data set depending on how the distributions, $F_k(X)$, are derived. One method is to pool a collection of training segments to obtain a transformation to be applied universally. The other method is to treat each segment independently and derive a separate RIDIT transformation to be applied to that segment. Tests were conducted to determine the discriminability after applying each of these methods, and comparisons were made to results obtained with the original data. These results are presented in section 4.

3. APPLICATION OF DISCRIMINANTS

In reference 2, Levine et al. describe a method for computing a discriminant to be used on the transformed data. Briefly, the method is the following:

a. Let A denote the matrix of transformed data; i.e., A_{ij} is the RIDIT value of the response to question j for the i^{th} respondent (pixel i).

b. Let Z_0 be the vector with the same dimension as the number of questions and with each component equal to the number 1.

c. Define a sequence of vectors, Z_j, recursively by the formula

$$Z_{j+1} = \frac{A^T A Z_j}{||A^T A Z_j||}$$

where $||.||$ is the usual euclidean norm. It is well known that the sequence Z_j will converge to the eigenvector corresponding to the largest eigenvalue of $A^T A$ when the multiplicity of that eigenvalue is 1. Since the RIDIT transformation has the property that each column has a mean of zero, $A^T A$ is merely the covariance matrix of the transformed data. Thus, the discriminant plane determined by this method is orthogonal to the direction of greatest variance in the data.

When no information is available concerning the underlying classes in the data and with the knowledge that the variance in the data is related to class membership, this discriminant seems reasonable. In applications involving the LIST data, however, ground truth observations were available. Thus, conventional discriminations were trained on the transformed data, and their performance was compared to the performance of the above converging routine and to the direct use of the major eigenvector. With the data sets available in this study, no example of the covariance matrix having a principal eigenvalue with multiplicity greater than 1 has been found.

4. RESULTS

To determine the applicability of RIDIT's to LIST data, two tests were conducted. The first test was to compare test and training discriminant accuracies obtainable with the normal data and with data transformed by the two RIDIT transformations discussed earlier. The discriminant used in this test was the quadratic discriminant (weighted by priors) which is available in the Statistical Analysis System (SAS, ref. 7). The transition year LIST data used in the test are comprised of 24 blind sites from North Dakota, South Dakota, and Minnesota. (See table 1.) In each discriminant run, the first 16 segments were used for training and the remaining 8 for testing. The resulting test and training accuracies are presented in table 2.

The second test was conducted to compare the accuracy obtained using the SAS discriminant on the two sets of transformed data obtained in the first test with the accuracy obtained using Levine's converging routine and the accuracy of using the major eigenvector directly. First, the data obtained by applying a separate RIDIT transformation to each segment were considered. For each segment, a separate covariance matrix was obtained from the transformed data. The major eigenvectors were extracted from SAS routines, and Levine's algorithm was applied to approximate the major eigenvector. The data were then projected onto each of these vectors and the resulting one-dimensional data sets were submitted to the SAS discriminant procedure to determine an optimal decision boundary for each segment. The accuracies obtained for each segment are presented in table 3, along with the means and standard deviations of the accuracies of each method.

Next, the separately transformed segments were grouped into training and test sets using the first 16 segments from table 1 as a training set and the remaining 8 as a test set. A covariance matrix was obtained from the training data. The major eigenvector and Levine's approximation were computed as before, and the training and test data were projected onto these vectors. The SAS discriminant procedure was applied to the projected training set to obtain the appropriate decision boundaries. The decision boundaries were then

TABLE 1.- LANDSAT DATA SEGMENTS, COUNTIES, AND ACQUISITIONS WHICH COMPRISE THE TRANSITION YEAR LIST DATA

Segment	County and State	Acquisitions
	Training set	
1380	Rockwood, Minnesota	78115, 78169, 78204, 78222
1394	Burke, North Dakota	78120, 78174, 78228, 78264
1457	Ward, North Dakota	78174, 78228, 78246, 78264
1461	Pierce, North Dakota	78137, 78190, 78217, 78236
1472	Barnes, North Dakota	78117, 78135, 78216, 78243
1473	Cass, North Dakota	78116, 78197, 78207, 78251
1518	Roseau, Minnesota	78135, 78188, 78224, 78243
1566	Grant, Minnesota	78133, 78169, 78196, 78232
1584	Pembina, North Dakota	78117, 78198, 78216, 78243
1602	Montrail, North Dakota	78174, 78211, 78228, 78264
1612	McHenry, North Dakota	78137, 78155, 78199, 78236
1619	Grandforks, North Dakota	78135, 78207, 78243, 78252
1636	Stutsman, North Dakota	· 136, 78154, 78217, 78243
1650	Hettinger, North Dakota	78156, 78209, 78218, 78246
1658	Dickey, North Dakota	78117, 78135, 78207, 78252
1668	Perkins, South Dakota	78153, 78174, 78228, 78264
	Test set	
1676	Brule, South Dakota	78135, 78207, 78224, 78234
1755	Jerauld, South Dakota	78117, 78153, 78197, 78225
1909	Kidder, North Dakota	78136, 78154, 78208, 78217
1918	Grant, North Dakota	78137, 78209, 78236, 78263
1656	Mormon, North Dakota	78137, 78155, 78209, 78263
1825	Norman, Minnesota	78133, 78169, 78196, 78232
1842	Yellow Medicine, Minnesota	78133, 78205, 78223, 78241
1784	Minnehaha, South Dakota	78134, 78169, 78196, 78223

TABLE 2.- A COMPARISON OF TEST AND TRAINING ACCURACIES OBTAINED WITH CONVENTIONAL AND TRANSFORMED DATA

Type of data set used	Probability of correct labeling	
	Training samples	Test samples
Normal LIST transition year data	72.32	66.64
LIST transition year data with the entire training set used to develop the RIDIT transformation	77.44	67.85
LIST transition year data with the RIDIT transformation derived and applied independently on each segment	75.75	77.10

TABLE 3.- ACCURACY OF USING THE MAJOR EIGENVECTOR FOR DISCRIMINATION ON INDEPENDENTLY TRANSFORMED DATA SETS

[Computed independently on each segment]

Segment	Probability of correct labeling (PCL)	
	Major eigenvector	Levine's method
1380	91.34	91.34
1394	66.46	65.85
1457	72.28	72.28
1461	75.00	75.00
1472	73.37	71.20
1473	64.14	64.14
1518	89.66	89.66
1566	80.62	80.62
1584	70.37	69.84
1602	68.28	68.82
1612	88.52	88.52
1619	74.03	74.03
1636	58.99	54.68
1650	86.67	86.67
1658	69.67	70.49
1668	90.38	90.38
1676	91.11	91.11
1755	87.10	89.68
1909	81.94	81.94
1918	81.32	81.32
1656	94.74	94.74
1825	79.46	79.46
1842	84.11	84.11
1784	78.20	78.20
Mean PCL	79.07	78.92
Standard Deviation	9.83	10.40

applied to the projected training and test sets to obtain training and test accuracies. These accuracies are presented in table 4. Also presented in table 4 are the accuracies obtained by repeating this above procedure starting with the data for which one RIDIT transformation was computed from the 16 training segments and applied to each segment. These results will be discussed in more detail in the next section.

5. CONCLUSIONS AND RECOMMENDATIONS

The identical test and training accuracies in table 4 occurred because of poor separation of the data along the direction of the discriminant vector. The separation was so poor, in fact, that the best discriminant decision boundary was placed so that all training samples (and test samples) would be classified into the predominant class, nonsmall grains.

This phenomena also occurred in the application to individual segments presented in table 3. There was also a great deal of variance noted in the individual eigenvectors generated for each segment, and the small-grains class was not projected in a way that would lead to a unified decision rule. Table 5 lists the major eigenvectors generated for five segments, and figure 1 gives two extreme examples of the distribution of the data along the major eigenvectors. Because of this poor separation, the use of the major eigenvector as a discriminant vector is not recommended.

The RIDIT transformation computed independently for each segment does, however, appear to hold promise for use as a data normalization technique. Note in table 2 that the test accuracy is significantly improved using the data where a separate RIDIT transformation was applied to each segment. There are examples in the LIST data of segments in which one of the LIST responses is constant across the segment. When no RIDIT transformation is performed, each of these responses is then weighted in comparison to the entire data set, and some must contribute to erroneous discriminant scores. The same principle holds when the response is compared to the entire set of responses to obtain a unified RIDIT transformation. However, if a separate RIDIT transformation is

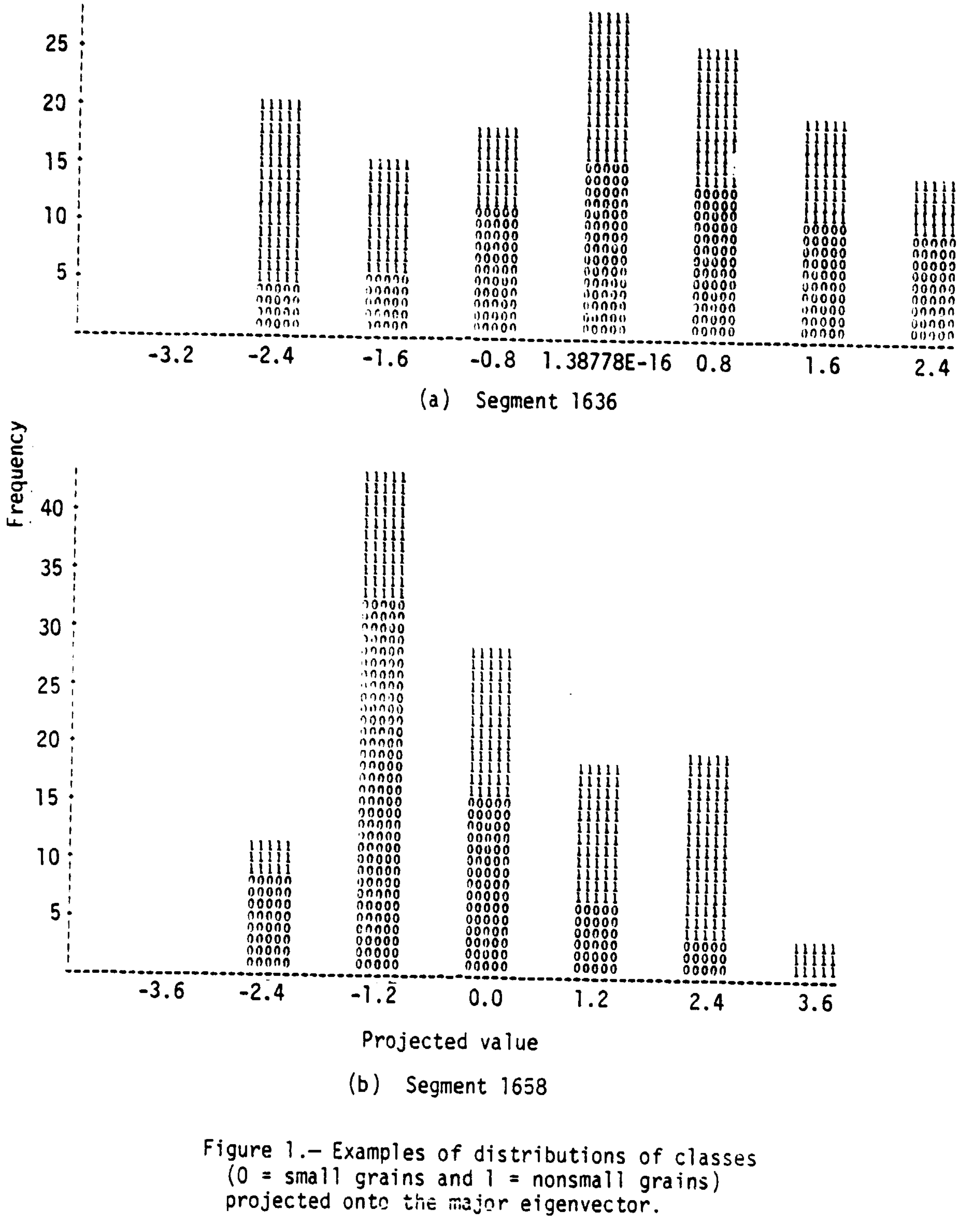
TABLE 4.- ACCURACY OF USING THE MAJOR EIGENVECTOR FOR DISCRIMINATION ON UNIVERSALLY TRANSFORMED DATA SETS

[Computed from pooled data sets]

Method of obtaining RIDIT transformation	Probability of correct labeling			
	Major eigenvector		Levine's method	
	Training data	Test data	Training data	Test data
Each segment transformed independently	64.46	83.31	64.46	83.31
One RIDIT transformation computed from training segments	64.46	83.31	64.46	83.31

TABLE 5.- LISTING OF MAJOR EIGENVECTORS
GENERATED FOR FIVE SEGMENTS

Major eigenvectors by segment				
1612	1619	1636	1650	1658
-0.02211	-0.0607	-0.15535	-0.0497	0.236357
-.03654	-.1967	-.30945	-.10001	.196131
-.14804	-.20283	.071697	-.0322	.135674
-.05088	-.17429	.087242	.025475	.040506
-.27146	.019571	-.29896	-.24965	.263358
-.27039	-.24168	-.3547	-.24342	.292772
-.26766	-.27172	.106559	-.25154	.191614
-.25388	-.24932	-.08914	-.31063	.217026
.181374	.218742	.091025	.176134	.097652
.315556	.30019	.341235	.374018	-.34074
.178593	.230127	.071158	.192524	.125144
.322089	.303404	.351265	.361622	-.32461
.160846	-.03046	.264795	.100246	.19588
.160148	.208172	.088044	.099044	-.09507
.151644	.199183	.011309	.080798	.133114
.144349	.163862	.074912	-.0635	.148612
.073459	.107764	-.14005	-.01276	.217098
.092598	.02344	.067957	.147505	.217098
.192346	.16167	.093618	.008633	.173433
.070835	.192986	.111979	.062737	-.03686
.254892	.174262	.294942	.174891	-.26336
.265028	-.03243	.33591	.249916	-.27712
.267663	.243377	.194668	.235427	-.17985
.253922	.262672	.113267	.251545	-.21703
	.249733		.312498	



(a) Segment 1636

(b) Segment 1658

Figure 1.— Examples of distributions of classes (0 = small grains and 1 = nonsmall grains) projected onto the major eigenvector.

computed for that segment, then the transformed value of the channel in question is zero for each pixel in that segment. Thus, that particular channel does not contribute to the discriminant score for the pixels in that segment but may indeed be applicable to pixels in other segments. The use of the RIDIT transformation in conjunction with conventional discrimination techniques seems to be justified under these circumstances.

6. REFERENCES

1. Dennis, T. B.; and Pore, M. D.: A Labeling Technology for Landsat Imagery. LEC-14357, December 1979.

2. Levine, A.; Roizen, P.; Roze, R.; and Christensen, H.: A Mathematical Method for Analyzing Questionnaires. Bulletin of World Health Organization, vol. 47, 1972, pp. 87-97.

3. Bross, I. D. J.: How to Use RIDIT Analysis. Biometrics, vol. 14, pp. 18-38.

4. Brockett, P. L.; and Levine, A.: On a Characterization of RIDIT's. The Annals of Statistics, vol. 5, no. 6, 1977, pp. 1245-1248.

5. Lynch, G. W.: A Decision Theoretic Approach to RIDIT's. Communications in Statistics, Theor. Meth., A7(6), 1978, pp. 607-614.

6. Selvin, S.: A Further Note on the Interpretation of RIDIT Analysis. American Journal of Epidemiology, vol. 105, no. 1, 1977, pp. 16-20.

7. Barr, A. J.; Goodnight, J. H.; Sall, J. P.; and Helwig, J. T.: A User's Guide to SAS 76. SAS Institute Inc., 1976.