Interactive Classification and Mapping of Multi-Dimensional Remotely Sensed Data Using n-Dimensional Probability Density Functions (nPDF)

Haluk Cetin and *Donald W. Levandowski*
Department of Earth and Atmospheric Sciences, Purdue University, West Lafayette, IN 47907

ABSTRACT: The *n*-dimensional probability density functions (nPDF) is an algorithm for displaying, analyzing, and classifying data. The technique is developed from what are often called frequency perspective plots, but overcomes the inherent limitation of earlier approaches. The interactive classification procedure using the nPDF algorithm has led to new approaches for the classification of multi-dimensional data. The nPDF plots provide a clear perspective of the distribution of remotely sensed data, as well as the training fields selected in supervised classification schemes. After mapping the nPDF of multi-dimensional data and training field distributions, the nPDF space can be divided according to the distribution of training field data. This nPDF division is then used as a look-up table to classify the data. For unsupervised classification, the nPDF plots may provide a valuable representation of data distribution that can be used directly to select the number of classes and locations of class means for initial clustering of the data. Complimentary to the use of nPDF for both supervised and unsupervised classification, the routine may also be used for data transformation and reduction. Besides the speed and low memory requirements, this transformation can be user directed to enhance particular features of interest.

INTRODUCTION

UNSUPERVISED AND SUPERVISED CLASSIFICATION STRATEGIES are commonly used for the classification of digital remotely sensed data. Previous studies of supervised classification techniques (e.g., maximum-likelihood, minimum distance, etc.) have demonstrated the difficulty of selecting training fields for classifying digital remotely sensed data. A conceptually similar problem is encountered during unsupervised classification in the selection of the number, standard deviations, and location of means (Swain and Davis, 1978; Wharton and Turner, 1981; Jensen, 1986; Chuvieco and Congalton, 1988).

Many techniques rely on purely statistical approaches to describe data and training field distribution. However, a graphical method, in conjunction with statistical techniques, has the advantage of providing a conceptually simple view of highly complex data distributions.

Several studies have been carried out on graphical methods in order to display remotely sensed data statistics using two-dimensional (2D) or pseudo three-dimensional (3D) plots. Several methods have been proposed to use two-dimensional displays of intersample distances through the use of distance measures. These are well explained by Fukunaga (1972) and are limited to two classes for each two-dimensional plot. Eyton (1983) described an instructional package for the use of frequency perspective plots which are also limited to a maximum of two bands per plot. Coggeshall and Hoffer (1973), Anuta (1977), and Swain and Davis (1978) showed the use of two-dimensional displays. Esbensen and Geladi (1989) described the use of principal components analysis for multivariate image analysis using frequency plots. In order to store more complex data distribution, Mori and Gotoh (1989) developed a method for the analysis of SPOT HRV (*Haute Resolution Visible*) data statistics in three-dimensional histogram space; however, their method is limited to three channels. Jensen (1979) presented a graphic method of analyzing training class statistics by viewing parallelepipeds, but this method is also limited to three channels or dimensions. Hodgson and Plews (1989) introduced a method to display cluster means in up to six-dimensional space using different sizes of numbers depending on the distance to the graphics plane and using red, green, and blue color combinations. This method, however, is not only inherently complex and difficult to interpret, but is limited to the plotting of mean locations of training fields.

This paper deals not only with the mapping of multi-dimensional digital remotely sensed data in order to overcome the inherent limitations of the earlier approaches, but also with the interactive classification of the data in terms of both supervised and unsupervised classification strategies by using the *n*-dimensional Probability Density Functions (nPDF) algorithm (Cetin, 1990). Finally, we demonstrate the usefulness of the nPDF algorithm for data transformation and reduction.

METHODOLOGY

In two-dimensional feature space, the position of any point is uniquely determined by the intersection of two scalars having different origins (Figure 1a). The intersection of the two arcs created by the scalars gives the location of the point in the feature space. The axes of the graph in Figure 1a represent two spectral bands: Band 1 (x_1) and Band 2 (x_2). For this discussion, we assume all original spectral bands are orthonormal. The range "R" is 255 (i.e., gray scale) for the 8-bit TM data. The feature vector is defined by

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

The magnitude of the scalars can be calculated by using the Euclidean distances

$$D_1 = (x_1^2 + x_2^2)^{1/2} \text{ and} \tag{1a}$$

$$D_2 = [(R - x_1)^2 + x_2^2]^{1/2}. \tag{1b}$$

A generalized distribution of highly correlated digital remotely sensed data in three-dimensional feature space is shown

PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING,
Vol. 57, No. 12, December 1991, pp. 1579–1587.

0099-1112/91/5712–1579$03.00/0

in Figure 1b. When a third dimension is added, we define the feature vector by

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

and the distances to the two corners of a cube (Figure 1b) will be

$$D_1 = (x_1^2 + x_2^2 + x_3^2)^{1/2} \tag{2a}$$

$$D_2 = [x_1^2 + x_2^2 + (R - x_3)^2]^{1/2} \tag{2b}$$

Although Graphical representation of four or higher dimensional data is impossible, it is still possible to calculate the distances vectorially. For the multi-dimensional case, the feature vector is defined by

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$$

where n is the dimension of the data. When a hyperdimensional cube is used, the vector magnitudes (the distances to the two corners) for n-dimensional data will be

$$D_1 = \left(\sum_{j=1}^{n} x_j^2 \right)^{1/2} \tag{3a}$$

$$D_2 = \left(\sum_{j=1}^{n} (x_{jj}^2 * a_j + (R - x_j)^2 * b_j) \right)^{1/2} \tag{3b}$$

$$\text{if} \begin{cases} j = 1,2,4,5,7,8,10,11,13,14, \dots & a = 1, b = 0 \\ j = 3,6,9,12,15, \dots & a = 0, b = 1 \end{cases}$$

where j is the band (dimension) number. The formula for the distance to the corners of a hyperdimensional cube can be generalized as

$$D_i = \left(\sum_{j=1}^{n} (x_j^2 * a_j + (R - x_j)^2 * b_j) \right)^{1/2}. \tag{4}$$

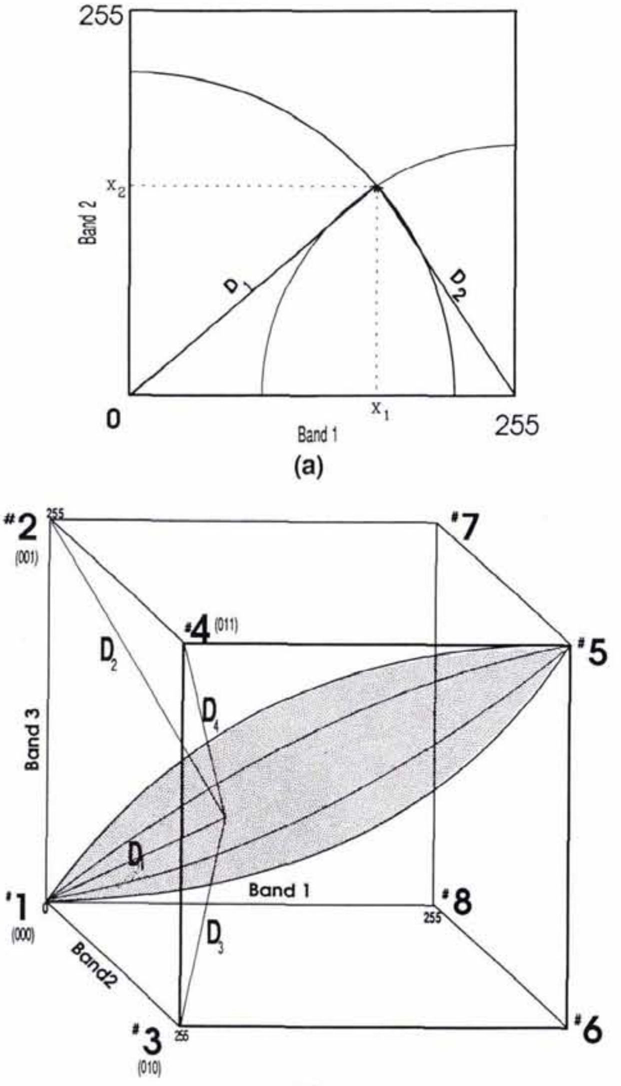


FIG. 1. (a) Two-dimensional feature space. (b) Three-dimensional feature space.

There are eight possible corners of a three-dimensional cube as is shown in Figure 1b. Four of the corners can be selected as principal corners (1 through 4), while the remaining corners (5 through 8) are the complimentary to the four principal corners. Thus, the complimentary pairs are corners 1 and 5, 2 and 6, 3 and 7, as well as 4 and 8. For the hyperdimensional cube model, a and b values needed to calculate the distances to the principal corners using the Equation 4 are as follows (j is the band number):

D_1: For all j values $\quad a = 1, b = 0$

$$D_2\text{: if} \begin{cases} j = 1,2,4,5,7,8,10,11,13,14,\dots & a = 1, b = 0 \\ j = 3,6,9,12,15,\dots & a = 0, b = 1 \end{cases}$$

$$D_3\text{: if} \begin{cases} j = 1,3,4,6,7,9,10,12,13,15,16,\dots & a = 1, b = 0 \\ j = 2,5,8,11,14,17,\dots & a = 0, b = 1 \end{cases}$$

$$D_4\text{: if} \begin{cases} j = 1,4,7,10,13,16,\dots & a = 1, b = 0 \\ j = 2,3,5,6,8,9,11,12,14,15,\dots & a = 0, b = 1 \end{cases}$$

The a and b values can also be determined in the following way: Each corner of a cube (3D) has three coordinate values. The corner #1 is selected as the origin of the cube; therefore, it has (0,0,0) for x,y,z coordinates, respectively. Similarly, coordinates for corner #2 are (0,0,255), for #3 are (0,255,0), and for #4 are (0,255,255). For the hyperdimensional cube, depending on the dimension or number of bands, a binary code can be used to determine the a and b values. If we use 0 for 0 coordinate and 1 for 255, the a and b values can be determined easily. The following example has 15 dimensions (bands 1 through 15 are from left to right):

corner #1: 000000000000000
corner #2: 001001001001001
corner #3: 010010010010010
corner #4: 011011011011011

For the Equation 4, if the binary number is 0 for the band, a is 1 and b is 0. Similarly, if the binary number is 1 for the band, a is 0 and b is 1.

The nPDF components are then calculated using the already determined distances, as well as a number of scaling factors: i.e.,

$$\text{nPDF}_i = S * D_i / (2^{\text{BIT}} * \text{NB}^{1/2}) \tag{5}$$

where

nPDF_i = component i of nPDF,
i = corner number,
S = desired scale factor for the nPDF axes (256, 512, etc.),
D_i = calculated distance for component i (the distance to corner i, calculated with Equation 4),
BIT = number of bits of input data (8 bits for TM, etc.), and
NB = number of bands used.

For this study, the reference points chosen were the two corners of a hyperdimensional cube, whose size was set within the 0 to 255 range, to calculate each nPDF component of the data sets described below. One corner was selected at one end of the maximum data distribution (corner #1 in Figure 1b), while the other corner is perpendicular to the maximum data distribution (corner #4 in Figure 1b). Depending on the distribution of the classes of interest in nPDF space, the user can select corners by which the separation of the classes is maximum.

In order to calculate the frequency values of the nPDF components, two of the nPDF components must be used at a time (i.e., nPDF_1 and nPDF_4 for this study). When the four corners are used (two at a time), there are six possible plots of the nPDF (1-2, 1-3, 1-4, 2-3, 2-4, and 3-4).

The frequency calculation for corners #1 and #4 can be performed by using the following routine:

```
            ....
First initialize Freq(i,j) = 0
Calculate Freq(i,j) values:
      For row=1 to nrow
            For column = 1 to ncol
                  ....(nPDF_i calculations)
                  Freq(i,j) = Freq(i,j) + 1
            Continue
      Continue
            ....
```

where *n*row = number of rows of the data, and *n*col = number of columns of the data.

The nPDF scale for this study was selected as 256. Therefore, a 256 by 256 array is needed for the frequency (Freq) calculations. The memory required for the array is summarized in Table 1.

The nPDF algorithm was first tested (Figure 2a) on six-band synthetic data (Cetin, 1990) using a scale of 256 and corners #1 and #4. Although the synthetic data have 16 separable classes, each band individually can be analyzed to discriminate between only two broad classes. Therefore, no one, two, or three bands can uniquely separate all the 16 classes. When the traditional two-dimensional frequency plots are used, 15 two-dimensional graphs are needed to display the data distriubtion, and a maximum of four classes could be separated per plot. On the other hand, the nPDF mapping function, using corners #1 and #4, provides a single perspective plot that successfully maps the data into 16 discrete classes (Figure 2b).

The algorithm was then applied to actual Landsat Thematic Mapper (TM) data consisting of seven bands and covering an agricultural area in Tippecanoe County, Indiana (Plate 1). Due

TABLE 1. THE REQUIRED VIRTUAL MEMORY (RAM) FOR DIFFERENT SCALES.

Scale	Frequency<65,536	Frequency<4.29×10^9
	(Integer*2 mode)	(Integer*4 mode)
256×256	132 Kilobyte	263 Kilobyte
512×512	525 Kilobyte	1.05 Megabyte
1024×1024	2.1 Megabyte	4.20 Megabyte

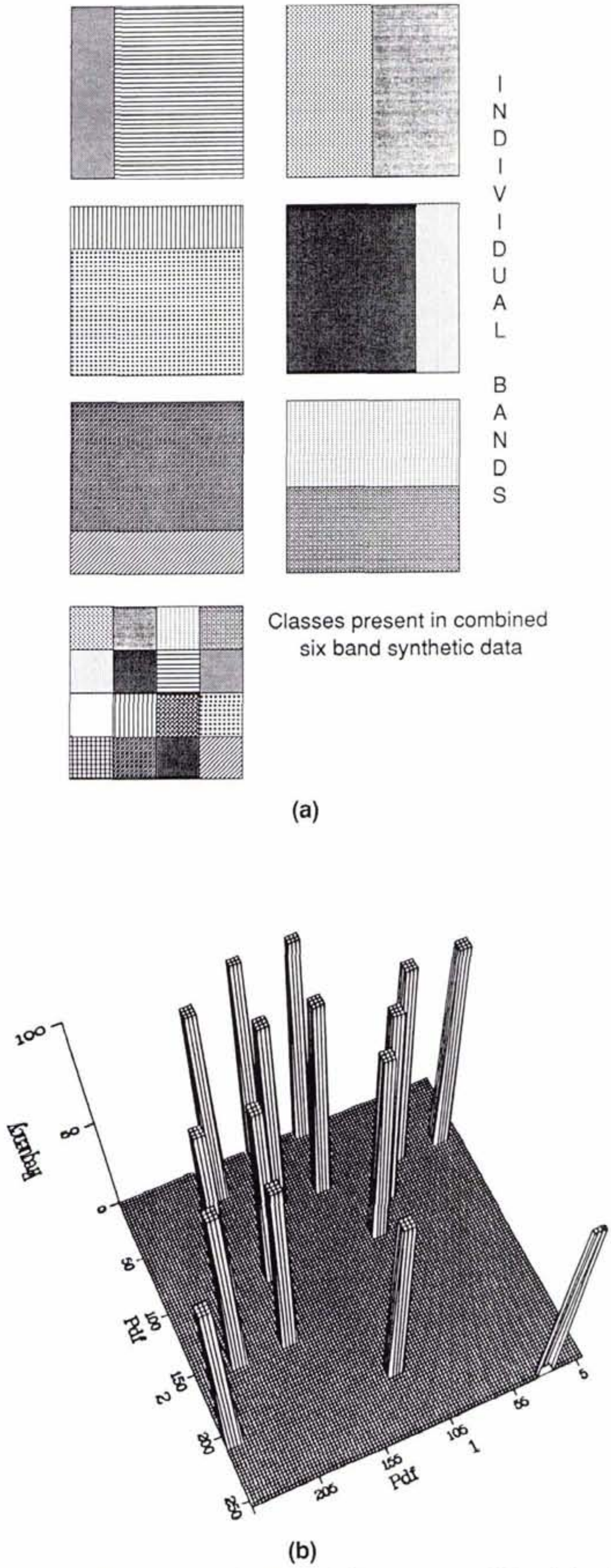


FIG. 2. (a) Six-band synthetic data. (b) The nPDF plot of the six-band synthetic data.

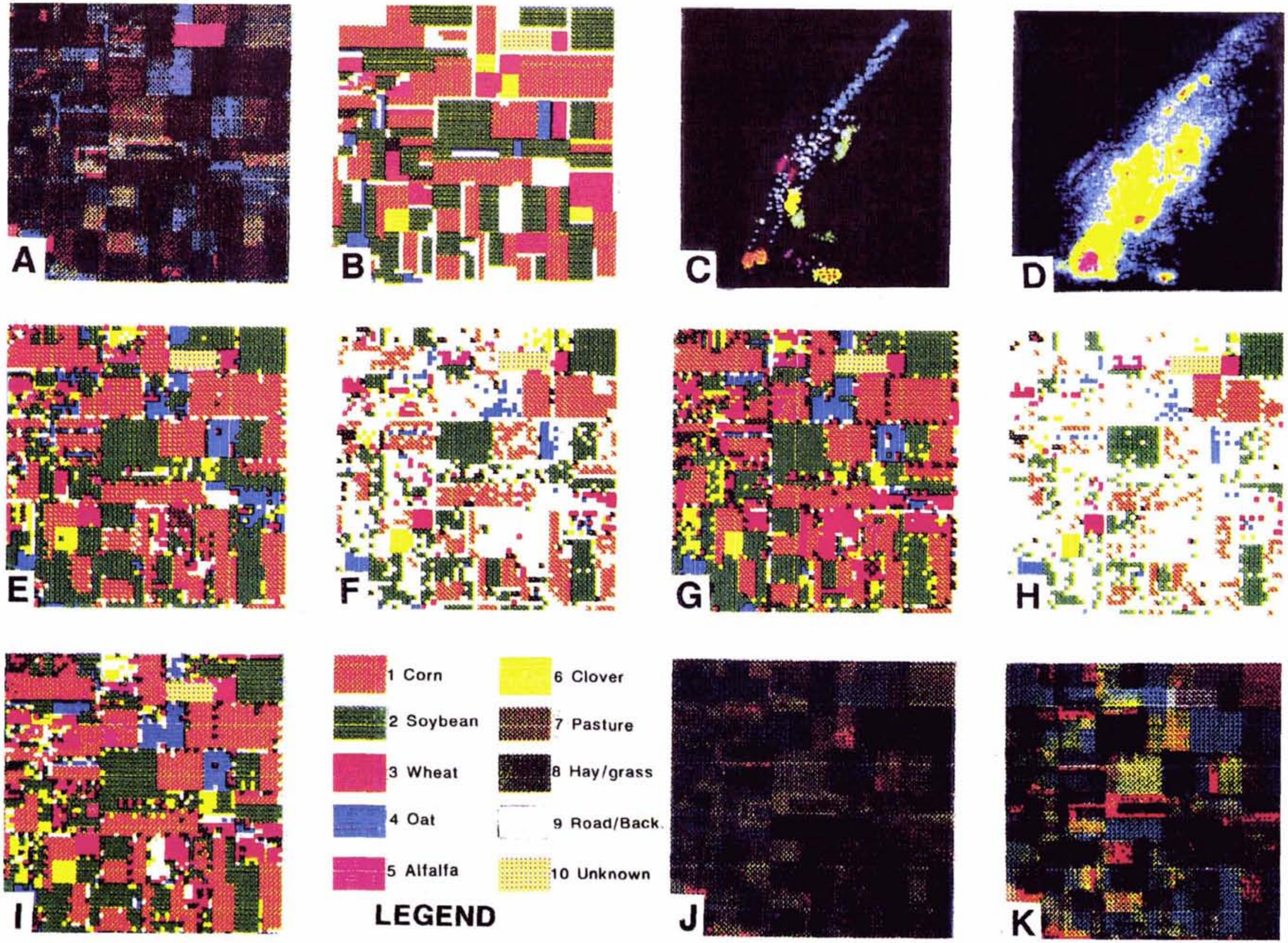


PLATE 1. (A) False color composite image (Red=4, Green=3, and Blue=2) of the TM scene. (B) The ground information of the area. (C) The NPDF plot of the training field distribution in image (raster) form. (D) The NPDF plot of the TM data distribution in image form. (E) The results of supervised classification using NPDF algorithm. (F) Maximum-likelihood classification results. (G) Minimum distance classification results. (H) Mahalanobis distance classification results. (I) The results of unsupervised classification (clustering) using NPDF algorithm. (J) Principal component (PC) analysis image of the TM data (Red= PC1, Green= PC2, and Blue= PC3). (K) The NPDF transformation image.

to the high correlation between the TM bands, the distances between the clusters are very small; therefore, very accurate calculations must be made. Consequently, after the range of output is determined, it is usually advantageous to stretch the range of the nPDF axes used. Although this can be done automatically, noise present in the data will tend to limit the results. Thus, the distribution should be examined first, and then the range to stretch the nPDF distribution can be selected by the user.

The nPDF of the remotely sensed data and training field distributions can be displayed using pseudo three-dimensional plots, in a two-dimensional contour map (Figure 3), in ASCII character mode (Figure 4) or in image (raster) form (Plate 1).

INTERACTIVE CLASSIFICATION OF REMOTELY SENSED DATA USING nPDF

The nPDF algorithm can be used not only to display multidimensional data and training field distributions, but also to classify the data using either supervised or unsupervised classifications. Two classification routines (supervised and unsupervised) were written to map and classify the remotely sensed data using the nPDF algorithm.

UNSUPERVISED CLASSIFICATION

The nPDF algorithm was first used to map the TM data and then, by using the "valley-peak seeking" approach, the number and location of means were selected for unsupervised classification (clustering). Thus, localized zones of high frequency (peaks in contour plots) were chosen as initial class means. The nPDF plot of the TM data is shown in color coded form in Plate 1D and in ASCII code in Figure 4a. Two different approaches can be used for the nPDF clustering:

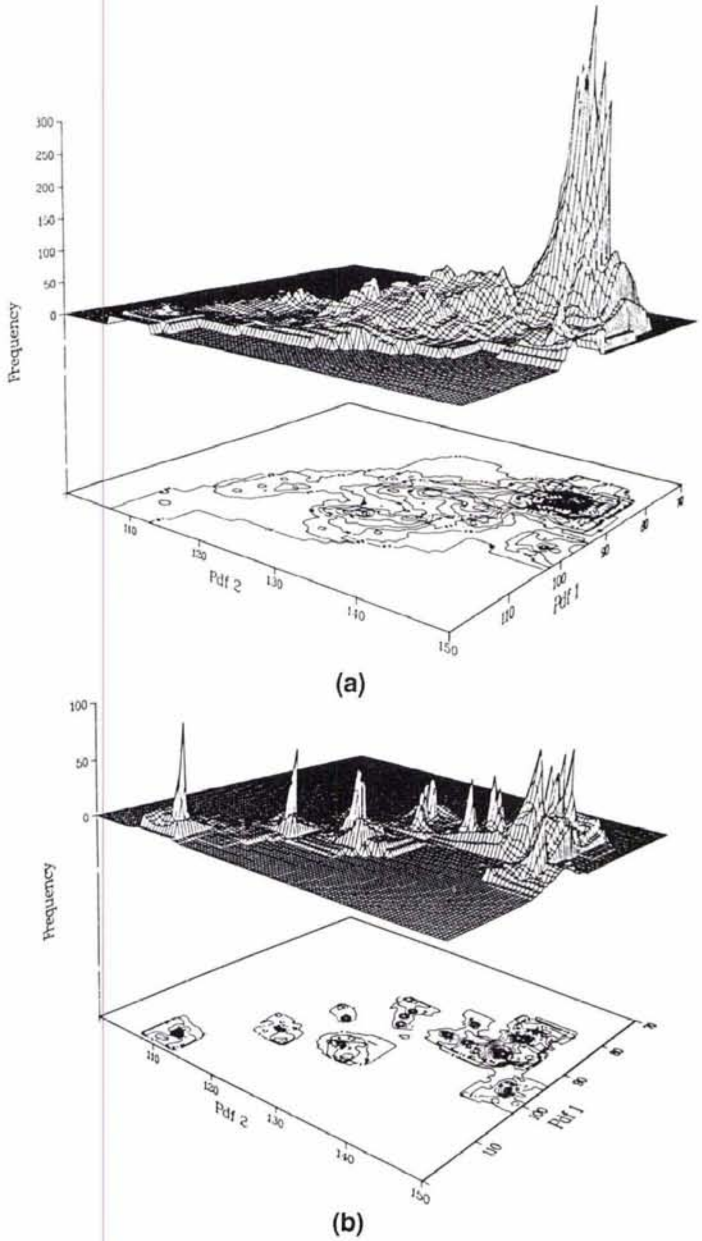
(1) Mean locations are selected depending on a maxima or mean of a cluster from the nPDF plot. After selecting the mean (or maxima) locations, the nPDF coordinates of the means are entered into the nPDF program as an input so that the data can be classified into the desired number of classes (depending on the number of means) by using a reverse calculation and a minimum distance seeking approach: i.e.,

$$d_{jk} = \|\text{nPDF}_j - \mathbf{M}_k\|$$

$$d_{jk} = \left(\sum_{i=1}^{2} (\text{nPDF}_i - m_{ki})^2\right)^{1/2}$$

where $\mathbf{M}_k$ is the mean vector obtained from nPDF plot for class k, nPDF_j is the calculated nPDF values for pixel j of the input data, and d_{jk} is the distance between them. The class with the smallest distance to the nPDF coordinates (values) is assigned to the pixel for the classification. After examining the results visually, some of the classes can be merged into distinct classes (in terms of spatial relationship).

(2) The clusters are outlined by simply drawing boundaries between the clusters on the nPDF plot of the data. Each of the windows created by the boundaries is considered as the class distribution on the plot. These boundaries are digitized and a raster form of image data with the same scale selected for the nPDF plot (256 by 256 for this study) is obtained. Because the location of every point (nPDF coordinates) with a class value is known, these data are used as a look-up table to classify the TM data. The 256 by 256 data are read into two-dimensional array, Class (i,j) array, for example. The value of the array, i being column (nPDF_1) and j being row (nPDF_4) of the raster image data, is the class number. The classification starts from the first pixel of the TM data and continues until the last pixel is classified. The following example serves to make the procedure clear:



(a)

(b)

FIG. 3. (a) The nPDF plot of the TM data distribution. (b) The nPDF plot of the training distribution.

The digital values for a given pixel are 10, 20, 30, 40, 50, 60, and 70 for the bands 1, 2, 3, 4, 5, 6, and 7, respectively. The range (R) is 255, scale (S) = 256, and TM data used are 8 bit. The results of the nPDF calculations are (for corners #1 and #4)

$$\begin{aligned}
D_1 &= (10^2 + 20^2 + 30^2 + 40^2 + 50^2 + 60^2 + 70^2)^{1/2} \\
&= 118.322 \\
D_4 &= (10^2 + (255-20)^2 + (255-30)^2 + 40^2 + (255-50)^2 \\
&\quad + (255-60)^2 + 70^2)^{1/2} \\
&= 438.748 \\
\text{nPDF}_1 &= 256 * 118.322/(256 * 7^{1/2}) = 44.72 = 45 \\
\text{nPDF}_4 &= 256 * 438.748/(256 * 7^{1/2}) = 165.83 = 166
\end{aligned}$$

If the class at nPDF location (45,166) was earlier classified as 3 (class 3), then the pixel is classified into that class.

Because the second approach is computationaly intensive and relies on boundaries which are sometimes difficult to identify,

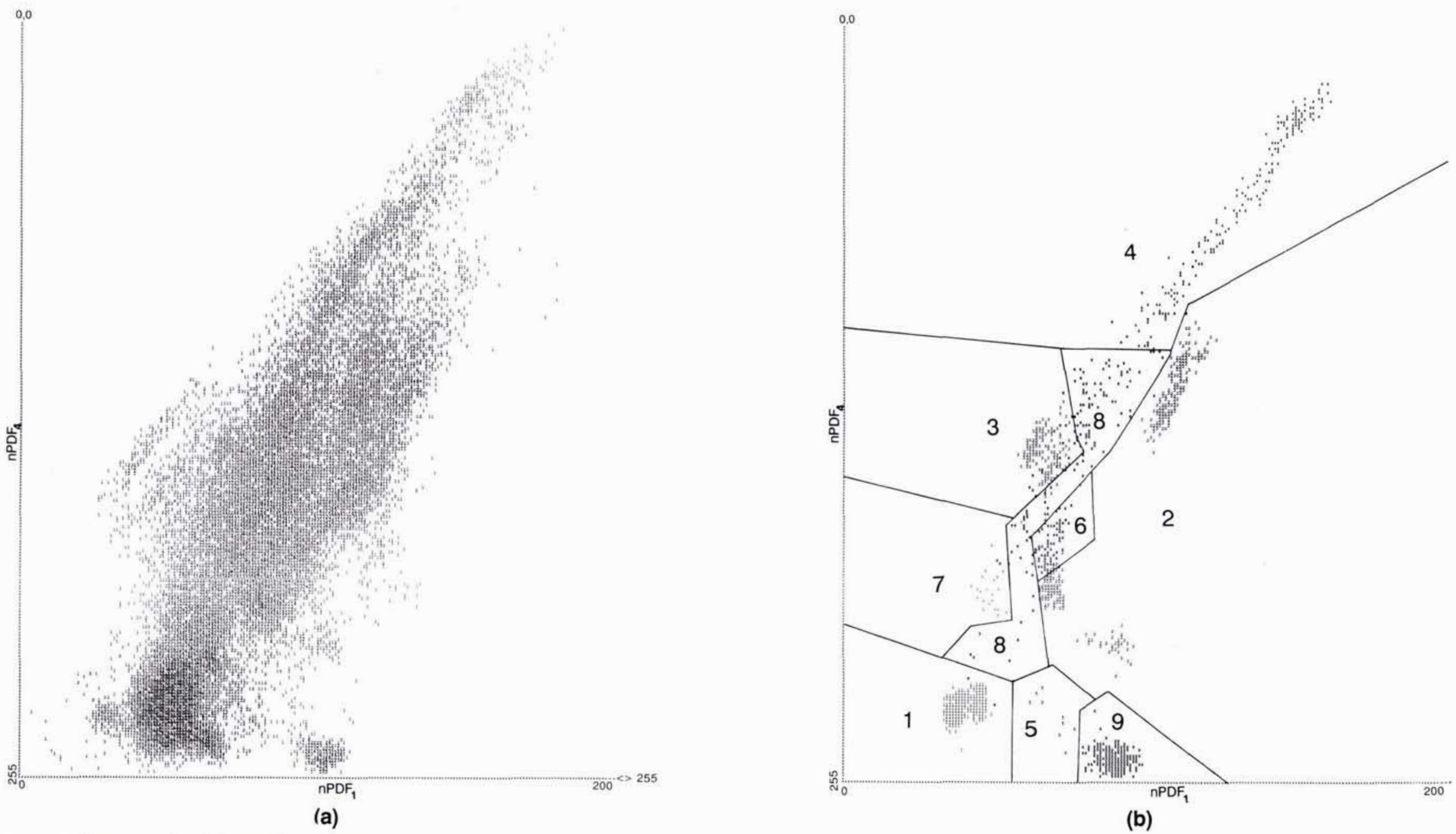


FIG. 4. (a) The nPDF plot of the TM data distribution in ASCII form; Frequencies are A = 10-15, B = 16-20, C = 21-25, D = 26-30, E = 31-35, F = 36-40, and G > 41. (b) The nPDF plot of the training field distribution in ASCII form; 1: corn, 2: soybean, 3: wheat, 4: oat, 5: alfalfa, 6: clover, 7: pasture/hay/grass, 8: road/background, 9: unknown field.

the first approach, which relies on a minimum distance analysis, was used for the clustering in this study. The nPDF clustering results are shown in Plate 1.

Supervised Classification

The algorithm was then used for supervised classification of the data. The nPDF of the TM data and training field distributions were mapped in image form (Plate 1 C and D) and in ASCII form (Figures 4a and 4b) with a scale of 256 by 256. For displaying and digitizing the data, ERDAS^R image processing software was used. The following procedure was used to classify the data:

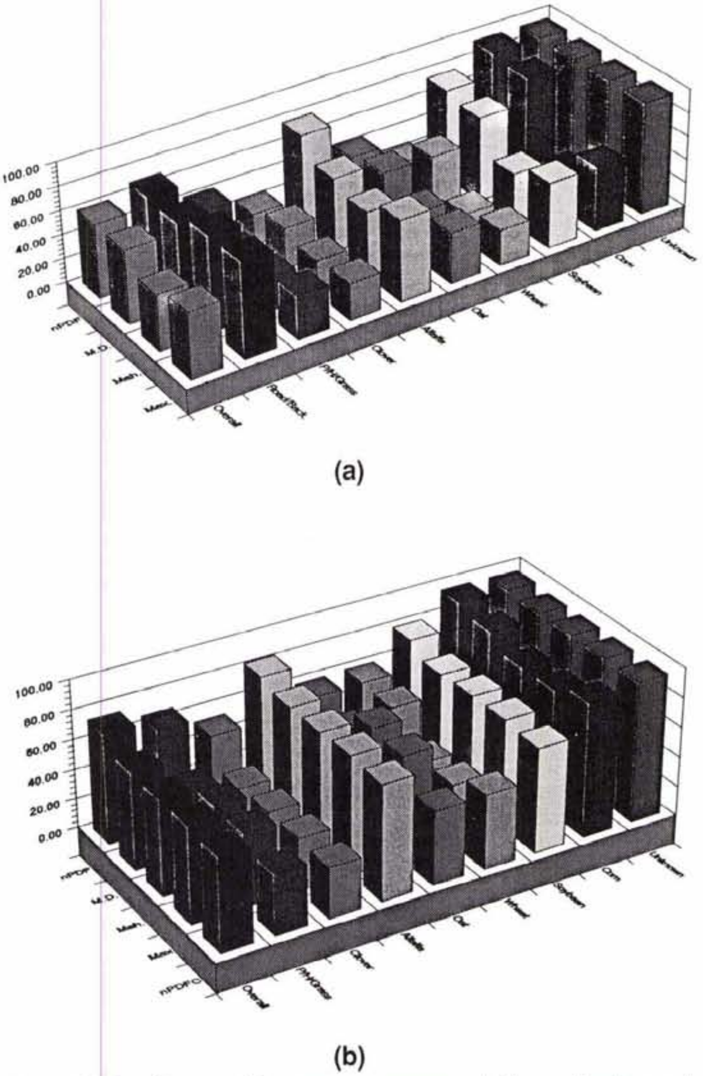
First, training fields were selected and their coordinates were entered into the nPDF program to map the training field distribution in nPDF space (Figure 4b). The training fields selected were three training blocks from soybeán fields; two training blocks each from corn, wheat, oat, hay/grass/pasture fields, and road/background; and one training block each from clover, alfalfa, and unknown field. As is seen from Figure 4b, the classes, excluding the road/background class, have distinct clusters. The boundaries between the classes were drawn midway between two closest neighboring classes by using screen digitizing on the monitor display of the nPDF training data map (Figure 4b). Then, the digitized areas were converted to a raster form GIS file with 256 rows and 256 columns, and each pixel was assigned a class determined by the digitized boundaries between the data classes. The 256 by 256 GIS file was read into a data array of the same dimension, as discussed in the previous section under part two of unsupervised techniques. The 256 by 256 data array was then used as a look-up table for the classification (every coordinate, or pixel, in the GIS file has a class value). The classification starts from the first pixel of the input data (TM, etc.) and the nPDF$_i$ values are calculated for the pixel. The nPDF values, or coordinates, are then used to look up the appropriate class number from the data array. The classification continues until the last pixel has been assigned a class.

Other classification algorithms, such as maximum-likelihood, minimum distance to means, and Mahalanobis distance, were used to compare the results obtained from the nPDF classification (Plate 1). Due to the large standard deviation of the "road/background" class (class #8 in Figure 4b), the results obtained from the maximum-likelihood and Mahalanobis distance classifications are not as good as the results obtained from the nPDF and minimum distance classifications. The "unknown" field was classified well by all the classification techniques used. The performance of the nPDF classification depends on how one draws the boundaries on the nPDF map (image) of training field distribution. Due to the influence of the "road/background" class, some of the fields belonging to the other classes (especially pixels belonging to the wheat and pasture/hay/grass classes) were classified into the "road/background" class by the nPDF procedure, as well as the other classification routines. The accuracy measures of the classifications are shown in Table 2 and Figure 5a.

Because the road/background class has a large standard deviation, this class was excluded from a new set of classifications. Table 3 and Figure 5b show the accuracy measures of the second set of classifications. For almost all of the classification schemes, the accuracy was improved. The nPDF clustering and minimum distance classification (without road/background class) results are almost identical because of the similarity in the method of classifying the data (both use a minimum distance approach); however, the nPDF approach is both faster and more flexible for user interaction. The run time of the classifications, obtained for the TM data with seven bands and 169 by 169 pixels, is shown in Table 4 (tested on Northgate-486 PC).

TABLE 2. ACCURACY MEASURES (WITH ROAD/BACKGROUND CLASS) OF THE nPDF, MINIMUM DISTANCE (M.D.), MAHALANOBIS (MAH.), AND MAXIMUM-LIKELIHOOD (MAX.) CLASSIFICATIONS (VALUES IN PERCENT).

Class	nPDF	M.D.	Mah.	Max.
Corn	94.70	91.55	29.72	54.36
Soybean	75.71	71.32	42.68	54.73
Wheat	26.57	51.95	22.93	29.80
Oat	45.87	51.87	38.62	42.25
Alfalfa	78.46	61.54	53.85	67.69
Clover	25.69	34.41	27.95	30.69
P/H/Grass	42.96	26.06	28.17	35.92
Road/Back	69.52	68.13	78.66	79.21
Unknown	94.70	94.70	94.95	94.95
Overall	61.58	61.28	46.39	54.40



(a)

(b)

FIG. 5. (a) Bar diagram of the accuracy measures (with road/background class) of the nPDF, Minimum Distance (M.D.), Mahalanobis (Mah.), and Maximum-Likelihood (Max). classifications. (b) Bar diagram of the accuracy measures (without road/background class) of the nPDF, nPDF Clustering (nPDFC), Minimum Distance (M.D.), Mahalanobis (Mah.), and Maximum-Likelihood (Max.) classifications (values in percent).

DATA REDUCTION USING nPDF TRANSFORMATION

Complimentary to the use of nPDF for both supervised and unsupervised classification, the nPDF approach may also be used

TABLE 3. ACCURACY MEASURES (WITHOUT ROAD/BACKGROUND CLASS) OF THE nPDF, nPDF CLUSTERING (nPDFC), MINIMUM DISTANCE (M.D.), MAHALANOBIS (MAH.), AND MAXIMUM-LIKELIHOOD (MAX.) CLASSIFICATIONS (VALUES IN PERCENT).

Class	nPDF	M.D.	Mah.	Max	nPDFC
Corn	96.03	91.55	82.34	82.79	90.63
Soybean	79.72	71.45	72.63	73.00	70.75
Wheat	59.70	58.55	37.60	39.68	50.34
Oat	59.12	52.25	71.00	69.25	50.00
Alfalfa	83.08	78.46	78.46	80.00	78.46
Clover	52.67	34.41	35.70	35.70	34.89
P/H/Grass	66.90	42.96	50.70	50.00	41.55
Unknown	95.71	94.70	94.95	94.95	94.70
Overall	74.12	65.54	65.42	65.67	63.92

TABLE 4. THE RUN TIME OF THE CLASSIFICATIONS OBTAINED FROM NORTHGATE-486 PC FOR THE TM DATA WITH SEVEN BANDS AND 169 BY 169 PIXELS

Maximum-Likelihood Classification:	130 seconds,
Mahalanobis Classification:	126 seconds,
Minimum Distance to Means Classification:	67 seconds,
nPDF Clustering:	52 seconds,
nPDF Classification:	24 seconds.

for data reduction and transformation. There are two ways to approach data reduction using the nPDF method.

(1) The first method is to select three reference points that one point is one end of the direction of the maximum distribution of the data, and the other two reference points are at the ends of the axes which are perpendicular to both the maximum distribution axis and to each other. The bands of the output data are created by using the distance formula (Equation 4)

$$D_i = \left(\sum_{j=1}^{n} (x_j^2 * a_j + (R - x_j)^2 * b_j) \right)^{1/2}$$

where, for each band (j), $a = 1$ and $b = 0$ or a = 0 and $b = 1$ and by using the nPDF formula (Equation 5). An example of a choice of a and b values for six-band input data is given in Table 5. For each of the output transformed bands, we desire a particular class of interest to have higher values than the other classes. With reference to Equation 4, it can be seen that, if the original DN values for the class of interest are higher relative to the other classes, a should be set to 1 and b to 0, and thus the original DN values are squared. This will tend to raise D_i for the class of interest. However, where the class of interest has comparatively low original DN values, we take the square of the difference between the range, R, and the DN values, which will again raise D_i for that class. This is achieved through setting a to 0 and b to 1 in Equation 4. Thus, when the summation is performed, D_i will tend to be highest for the class of interest. The following example, from a Pioche, Nevada, TM scene uses representative pixels from three classes that are difficult to separate in that scene: hydrothermally altered areas, vegetation, and light-toned soils (see Table 5). By comparing the DN values for each band, for each class, appropriate a and b values may be chosen to separate a particular class in each output band. Table 5 shows that light-toned soil class tends to have the highest DN values in all bands. Thus, for band 1, we select a to equal 1 and b to equal 0 for all bands. D_1 is thus the square root of the sum of the squares of the DN values and the resulting nPDF_1 value is 112 for the light-toned soil, compared to 66 for vegetation and 80 for the hydrothermally altered areas. In band 2, vegetation is enhanced. Table 5 shows that vegetation has higher DN values than hydrothermally altered areas in bands 2, 4, and 6 (2.08 to 2.36 μm), and thus a value of 1 is chosen for a and 0 for b. For these three TM bands, the original DN values are therefore used in the calculation. For bands 1, 3, and 5, which have lower DN values for vegetation than the hydrothermally altered area, by selecting 0 for a and 1 for b, the DN values are subtracted from 255 (R = range = 255), thus giving vegetation higher value than the altered area. The calculated nPDF_2 of 138 is therefore higher than hydrothermally altered area (120) and light-toned soil (111).

Similarly, a and b values are chosen for band 3 so that the output data will have high values for hydrothermally altered areas. Because the altered areas have high DN values in bands 1, 3, and 5, b values of 0 and a values of 1 are used for those bands. In bands 1, 3, and 5, where the hydrothermally altered area has lower DN values than vegetation, b values are set to 1 and a values are set to 0. Therefore, band 3 will have higher values for the hydrothermally altered area than those of the vegetation or soil.

When an RGB color combination is used for the bands 1, 2, and 3 of the output data, respectively, the light-toned soil class will tend to be red, the vegetation green, and the hydrothermally altered areas blue. The mixed pixels will have colors that are combinations of these three bands.

(2) A second alternative approach for data reduction is to obtain four bands of output data by using the four corners of the hyperdimensional cube. For the band 1 of the output data, the corner #1 is used (Figure 1b). Similarly, the corner #2, #3, and #4 are used for band 2, band 3, and band 4 of the output data, respectively. The calculations are done as described before and the nPDF_i values are used as DN (digital number) for the corresponding band.

These calculations are significantly more efficient than the principal components (PC) analysis in terms of CPU time. For the TM data with 169 by 169 pixels and seven bands, the run time for the nPDF reduction was 22 seconds, whereas PC analysis calculations took 55 seconds. When the dimension of the data increases, the CPU time increases dramatically for the PC analysis compared to the nPDF method. Furthermore, the nPDF does not require as much memory as is needed by PC calculations. This is especially significant, because memory requirements become a problem with the PC analysis when higher dimensional data are used.

The stretching of the PC data is limited to the transformation matrix and thus it is essentially a "blind approach" driven by the data distribution. In contrast, the nPDF transformation axes may be chosen so as to take advantage of the distributions of selected pixels. A PC rotation will tend to be dominated by the major spectral differences in the scene (i.e., snow, cloud, and bare ground). However, these distributions can be identified on the nPDF plots, and a transformation and enhancement can be selected that will tend to separate the cover types of interest. Even where snow or cloud is not a problem, such as for the Tippecanoe County scene (see Plate 1), the nPDF analysis can be used to produce an image that enhances classes better than that of a PC analysis.

DISCUSSION AND CONCLUSIONS

Applications of the nPDF algorithm show that it can be used not only to display multi-dimensional data and training field distributions, but also to classify the data. Additionally, the algorithm can be used for data reduction, which can further aid in identifying classes and showing the relative intraclass distribution.

One of the limitations of the commonly used classifiers is that training fields representing the entire data must be selected in

TABLE 5. THE DN, a, AND b VALUES OF THE TM SCENE AND DISTANCE-nPDF CALCULATIONS FOR THE LIGHT-TONED SOIL, VEGETATION, AND HYDROTHERMALLY ALTERED AREA CLASSES

	DN values for TM bands (a and b values)						D_i			nPDF_i		
Class	1	2	3	4	5	6	D_1	D_2	D_3	nPDF_1	nPDF_2	nPDF_3
Light-toned soil	126 (10)	66 (10)	88 (10)	97 (10)	172 (10)	92 (10)	274.7	272.1	375.1	112	111	153
Vegetation	96 (01)	54 (10)	44 (01)	78 (10)	71 (01)	35 (10)	162.5	338.0	369.4	66	138	151
Hydrothermally altered area	102 (10)	50 (01)	62 (10)	67 (01)	126 (10)	33 (01)	195.5	292.9	396.2	80	20	162

order to classify the data. The nPDF algorthim does not have this limitation.

In order to get reasonable results, very accurate calculations, Integer*4 mode (long integer) for distance calculations, and double precision for square root calculation, should be made and appropriate reference systems should be selected (hyperdimensional cube, etc.). The classification results depend on how one selects the maximum data distribution of a class of interest from the image created using the nPDF training field distributions. Therefore, the classification performance depends on the user, not on the computer. The reduction method using the nPDF method is faster and does not require as much memory as is required by the PC analysis.

SOFTWARE IMPLEMENTATION

The nPDF software was written in FORTRAN-77 code on an IBM 3090 main frame using the DISSPLA graphics package and in "C" language on an IBM-PC based image processing system (ERDAS). It is interactive software that allows one to use different reference coordinates and scales to map and classify the multi-dimensional data. For information on the availability of the software, please contact the first author.

ACKNOWLEDGMENTS

We wish to thank Mr. Timothy Warner and Dr. Arvind Chaturvedi for their helpful suggestions. The Laboratory for Applications of Remote Sensing (LARS) at Purdue University kindly provided the Landsat TM data and ground information. This study was supported by Indiana Mining and Mineral Resources Research Institute (IMMRRI), Grant No. G1184118 and G1194118.

REFERENCES

Anuta, P. E., 1977. Computer-Assisted Analysis Techniques for Remote Sensing Data Interpretation. *Geophysics,* 42(3): 468–481.

Cetin, H., 1990. nPDF-An Algorithm for Mapping n-Dimensional Probability Density Functions for Remotely Sensed Data. *Proceedings of the 10th Annual International Geoscience & Remote Sensing Symposium, IGARSS'90,* I: 353–356.

Coggeshall, M. E., and R. M. Hoffer, 1973. *Basic Forest and Cover Mapping Using Digitized Remote Sensor Data and ADP Techniques.* LARS Information Note 030573, Purdue University, 131 p.

Chuvieco, E., and R. G. Congalton, 1988. Using Cluster Analysis to Improve the Selection of Training Statistics in Classifying Remotely Sensed Data. *Photogrammetric Engineering & Remote Sensing,* 54(9): 1275–1281.

Esbensen, K., and P. Geladi, 1989. Strategy of Multivariate Image Analysis (MIA). *Chemometrics and Intelligent Laboratory Systems,* 7: 67–86.

Eyton, J. R., 1983. A Hybrid Image Classification Instructional Package. *Photogrammetric Engineering & Remote Sensing,* 49(8): 1175–1181.

Fukunaga, K., 1972. *Introduction to Statistical Pattern Recognition.* Academic Press, New York.

Hodgson, M. E., and R. W. Plews, 1989. N-Dimensional Display of Cluster Means in Feature Space. *Photogrammetric Engineering & Remote Sensing,* 55(5): 613–619.

Jensen, J. R., 1979. Computer Graphic Feature Analysis and Selection. *Photogrammetric Engineering & Remote Sensing,* 45(11): 1507–1512.

———, 1986. *Introductory Digital Image Processing.* Prentice-Hall, New Jersey.

Mori, M., and K. Gotoh, 1989. Three-Dimensional Analysis of SPOT HRV Data. *Proceedings of the 12th Canadian Symposium on Remote Sensing, 9th Annual International Geoscience & Remote Sensing Symposium, IGARSS'89,* 2: 467–470.

Swain, P. H., and S. M. Davis, 1978. *Remote Sensing, The Quantitative Approach.* McGraw Hill Inc., New York.

Wharton, S. W., and B. J. Turner, 1981. ICAP: An Interactive Cluster Analysis Procedure for Analyzing Remotely Sensed Data. *Remote Sensing of Environment,* 11: 279–293.

(Received 9 May 1990; revised and accepted 27 February 1991)