AgRISTARS

SR-T9-00402
NAS9-14689
NASA CR-
160427

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

November, 1979

Supporting Research

FINAL REPORT

DEVELOPMENT AND EVALUATION OF CLUSTERING PROCEDURES

By: L. F. Guseman, Jr.

DEPARTMENT OF MATHEMATICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS

OMIT

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4 Title and Subtitle
FINAL REPORT: Development and Evaluation of Clustering Procedures

5. Report Date
November 1979

6. Performing Organization Code

7. Author(s)
L. F. Guseman, Jr.

8. Performing Organization Report No

9 Performing Organization Name and Address
Department of Mathematics
Texas A&M University
College Station, Texas 77843

10. Work Unit No.

11. Contract or Grant No.
NAS-9-14689-9S

13 Type of Report and Period Covered
FINAL (11/1/78-10/31/79)

2 Sponsoring Agency Name and Address
Earth Observations Division
NASA/Johnson Space Center
Houston, Texas 77058

14. Sponsoring Agency Code

5 Supplementary Notes

6. Abstract

Work carried out under the above contract was concerned with:

Development of Spatial Clustering Algorithms

Development of Multipass Color Displays

Investigation of Histogramming Techniques

Evaluation of Data Reduction Systems

7 Key Words (Suggested by Author(s))	18. Distribution Statement

9 Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
		10	

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

FINAL REPORT

DEVELOPMENT AND EVALUATION OF

CLUSTERING PROCEDURES

Contract NAS-9-14689-9S

November 1, 1978 - October 31, 1979

Prepared for:

Earth Observations Division
NASA/Johnson Space Center
Houston, Texas 77058

by

L. F. Guseman, Jr.
Principal Investigator
Department of Mathematics
Texas A&M University
College Station, Texas 77843

ACKNOWLEDGMENTS

The investigations discussed herein were carried out for the Earth Observations Division, NASA/Johnson Space Center, Houston, Texas, under Contract NAS-9-14689 to the Texas A&M Research Foundation, College Station, Texas, 77843, during the period November 1, 1978 to October 31, 1979. The work was coordinated through several discussions with key personnel from NASA/EOD, especially Dr. Mickey Trichel. The various tasks were pursued in conjunction with Professor Jack Bryant of Texas A&M University, and Professor Patrick L. Odell, University of Texas at Dallas. Graduate Assistants David Egle and Gary Breaux, and student assistants Keith Albright and David Taylor provided programming support for the tasks.

L. F. Guseman, Jr.
Principal Investigator

DEVELOPMENT AND SELECTION OF CLUSTERING PROCEDURES

1. INTRODUCTION

A practical application of remote sensing which is of considerable interest is the use of satellite-acquired (LANDSAT) multispectral scanner (MSS) data to conduct an inventory of some crop of economic interest such as wheat over a large geographical area. Any such inventory requires the development of accurate and efficient algorithms for analyzing the structure of the data. The use of multi-images (several registered passes over the same area during the growing season) increases the dimension of the measurement space. As a result, characterization of the data structure is a formidable task for an unaided analyst.

Cluster analysis has been used extensively as a scientific tool to generate hypotheses about structure of data sets. Sometimes one can reduce a large data set to a relatively small data set by the appropriate grouping of elements using cluster analysis. In some cases, the algorithm which effects the grouping becomes the basis for actual classification. In other cases, the cluster analysis produces groupings of the data which in turn serve as training sets for a classification procedures (e.g. Gaussian maximum likelihood). Additional uses of cluster analysis arise in conjunction with dimensionality reduction techniques which are used to generate displays for purposes of further interactive analysis of the data structure.

Investigations carried out under this contract centered around the refinement of previously-developed spatial clustering algorithms and their

application to the generation of color displays of multipass data. Additional investigations into incorporating histogramming techniques into these algorithms were also carried out. A preliminary (somewhat theoretical) approach to analyzing large data reduction systems was also formulated. Specifically, investigations were carried out in the following areas:

Development of Spatial Clustering Algorithms

Development of Multipass Color Displays

Histogramming Techniques

Evaluation of Data Reduction Systems.

The results of each of these investigations is discussed in turn in the sequel.

2. DEVELOPMENT OF SPATIAL CLUSTERING ALGORITHMS

In a LACIE type application, the input data for a clustering algorithm is a multi-image; namely, a set of registered images, taken at different times, of the same subject. In addition to having multidimensional data (multispectral measurements) we also have "multi-picutres" of the subject. The availability of this spatial aspect of LACIE data and attempts to preserve its spatial integrity are the basis for the clustering algorithms developed throughout this contract period.

In previous contract periods initial versions of spatial clustering/classification algorithms were formulated and implemented. The results of these investigations appear in:

Jack Bryant, AMOEBA: A Spatial Clustering Algorithm for Multi-temporal LANDSAT Data, Contract NAS-9-14689-6S, Report #15, Department of Mathematics, Texas A&M University, August, 1977.

Jack Bryant, Clustering Boundary Pixels in LANDSAT Data, Contract NAS-9-14689-8S, Report #16, Department of Mathematics, Texas A&M University, July, 1978.

Jack Bryant, Applications of Clustering In MUlti-image Data Analysis, Contract NAS-9-14689-8S, Report #18, Department of Mathematics, Texas A&M University, August, 1978.

J. Bryant, On the Clustering of Multidimensional Pictorial Data, J. Pattern Recognition 11, (1979), 115-126.

J. D. Bryant, On the Clustering of Multidimensional Pictorial Data, Proceedings of Technical Sessions, Vol. II, The LACIE Symposium, NASA/Johnson Space Center, Houston, Texas, 77058, July, 1979, pp. 647-656.

The above reports form the basis for investigations carried out during this contract period in the area of spatial clustering/classification algorithm development.

Work was continued on refinements of the AMOEBA algorithm. The details for this work is contained in the expository paper:

> Jack Bryant, The Easy Remote Sensing Problem, Contract NAS-9-14689-9S, ROUGH DRAFT, Report #20, Department of Mathematics, Texas A&M University, August, 1979.

This report presents a model for multi-imagery which is applicable to cloud and haze-free LANDSAT MSS multitemporal data taken from flat areas of the earth dominated by agricultural activity. A detailed discussion is presented which explores the properties of a multitemporal LANDSAT data set and how they can be realistically incorporated into an underlying model for multi-imagery. An unsupervised classification problem is considered in the setting of the model and a computer program for solving the problem is discussed. The methodology unifies the spatial, temporal and measurement-space aspects of the problem.

3. DEVELOPMENT OF MULTIPASS COLOR DISPLAYS

Investigations under this contract have concentrated on the continued development of techniques for combining registered multitemporal LANDSAT data into a single 3-color display. An algorithm has been formulated which attempts to make use of some of the spatial aspects of the data. Investigations into the development of the algorithm can be divided generally into the following three areas:

1. Prototype selection
2. Dimensionality reduction
3. Color assignment

The prototype selection portion of the algorithm is a consequence of the approach used to perform the actual dimensionality reduction. Details of the dimensionality reduction are presented in the attached report:

> Jack Bryant and L. F. Guseman, Jr., Distance Preserving Linear Feature Selection, Contract NAS-9-14689-8S, Report #19, Department of Mathematics, Texas A&M University, July, 1979.

Initially, prototype selection was accomplished using a representative from each of the distinct classes in the multi-image as determined by the AMOEBA algorithm. Another approach to prototype selection which appears to be satisfactory (using Robert's gradient of distance 2) is discussed in the above report.

The color assignment portion of the algorithm is still under development. Only recently were the ranges for the output of the dimensionality reduction step determined so as to be in the domain of the inverse Faugheras transformation. This is currently being implemented for preliminary testing.

Additional investigations in this area were concerned with attempts to speed up the computational method for determining the feature selection matrix. The Davidon-Fletcher-Powell unconstrained minimization algorithm appears to be the most accurate and reliable.

4. HISTOGRAMMING TECHNIQUES

Chapter 5 (Applications of Histogramming) of Report #18 cited above contains a survey of several image analysis systems in which multi-dimensional histogramming is used for clustering, data compression and interactive analysis. During this contract period, FORTRAN programs were developed at Texas A&M University implementing some of these techniques. A new technique was developed to allow multi-temporal multi-spectral histogramming to be performed on a small computer. Programs using this new technique in conjunction with distance preserving feature selection were installed on a minicomputer at the Data Analysis Laboratory, EROS Data Center during August, 1979. A report on this work is being written by J. Bryant, TAMU, and S. Jenson, EROS.

The TAMU developed software for histogramming, and the modification of the dimensionality reduction and previously developed clustering programs, is operational. Documentation is at the "well commented listing" stage.

5. EVALUATION OF DATA REDUCTION SYSTEMS

Data reduction systems which utilize multitemporal MSS data to produce proportion estimates of several crop classes are large and complicated. Large numbers of vector-valued observations are used, in conjunction with algorithms based on various models, to produce these estimates. Testing the validity of these models and determining the subsequent effect on the accuracy of the porportion estimates cannot (in many instances) be carried out. In addition, when the software system is (conceptually) the best it may be that proporties of the original data set in fact impose the accuracy limitations.

A theoretical approach to determining the limiting accuracy of the data set is set forth in the report:

> Patrick L. Odell, The Cramer-Rao Lower Bound as a Criteria for Evaluating A Large Complicated Data Reduction System Such as LACIE, Contract NAS-9-14689-9S, Report #21, Department of Mathematics, Texas A&M University, September, 1979.

The problem of how such an approach can be implemented and tested on real data is a problem for further investigation.

DISTANCE PRESERVING LINEAR FEATURE SELECTION

Jack Bryant and L. F. Guseman, Jr.
Department of Mathematics
Texas A&M University
College Station, Texas 77843

Report #19

Prepared For
Earth Observations Division
NASA/Johnson Space Center
Houston, Texas
Contract NAS-9-14689-9S
July, 1979

D/2

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle
Distance Preserving Linear Feature Selection

5. Report Date
July, 1979

6. Performing Organization Code

7. Author(s)
Jack Bryant and L. F. Guseman, Jr.

8. Performing Organization Report No.
19

9. Performing Organization Name and Address
Department of Mathematics
Texas A&M University
College Station, Texas 77843

10. Work Unit No

11. Contract or Grant No
NAS-9-14689-9S

12. Sponsoring Agency Name and Address
Earth Observations Division
NASA/Johnson Space Center
Houston, Texas 77058

13. Type of Report and Period Covered
Unscheduled Technical

14. Sponsoring Agency Code

15. Supplementary Notes
Principal Investigator: L. F. Guseman, Jr.

16. Abstract

A new method for linear feature selection is described which has as its underlying theme the preservation of actual distances between training data points in the lower dimensional space. Comparison with existing methodology places the method closer to the principle components or Karhunen-Loève approach than to methods based on an approach through statistical pattern recognition. A computer program implementing the technique is described. An example application to 12 dimenisonal LANDSAT data is given.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Linear feature selection Principle components Clustering LANDSAT data	

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

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

DISTANCE PRESERVING LINEAR FEATURE SELECTION

Jack Bryant and L. F. Guseman, Jr.
Department of Mathematics
Texas A&M University
College Station, Texas 77840

Abstract

A new method for linear feature selection is described which has as its underlying theme the preservation of actual distances between training data points in the lower dimensional space. Comparison with existing methodology places the method closer to the principle components or Karhunen-Loève approach than to methods based on an approach through statistical pattern recognition. A computer program implementing the technique is described. An example application to 12 dimensional LANDSAT data is given.

Linear feature selection Principle components Clustering LANDSAT data

INTRODUCTION

The setting for this paper is the following: We are given M measurement vectors $x_1,\ldots,x_M$. Each vector is an n-tuple of real numbers $x_i = (x_{i1},\ldots,x_{in})^T$. Following Andrews,[1] we call the complete data set pattern space; it is naturally embedded in n-dimensional Euclidean space E^n. Both the dimensionality n and the number M of data points are assumed to be large. We also have at hand a set of p distinct prototypes $y_1,\ldots,y_p$ which are assumed to contain representatives of the classes present in pattern space. They may be specified externally (i.e., in a supervised sense), or may be the result of an unsupervised clustering technique.[2,3,4] Since the actual class structure present in the real world (from which our data comes) is the subject of investigation (i.e., is unknown), we make no assumption that the prototypes exactly match real classes. We just assume we have a representative sample of most of the real classes, or, in Haralick's terminology[5], we know "where the action is." In particular, the number of classes is unknown, and there are too few prototypes to make parametric estimates on class statistics.

In this setting, we investigate the structure of the data. For example, we may use a descendant of the famous ISODATA clustering algorithm.[6] Or we may wish to display the data (not just the prototypes) in two or three dimensions. We hope that ISODATA produces a nice clustering, or that obvious clusters appear in two or three dimensional scatter plot displays. Of course, ISODATA is expensive to run on the original high dimensional large data set. The main problem is the dimensionality, which also hampers human understanding of the geometric structure of the raw data. These problems naturally lead to feature selection. Heuristically, the idea is that the

"intrinsic dimensionality" of pattern space is much less than the dimension of the Euclidean space in which it is embedded. This can be expected for at least two reasons:

(i) Two different measurements from the real world are almost never independent.

(ii) The structure we are looking for actually exists.

Daly[7] has noticed a relationship between clustering (i.e., finding prototypes) and factor analysis (here, feature selection). The relationship goes something like this: if the prototypes can be found which represent the actual classes present in the data, then a basis in pattern space consisting of $k < n$ vectors can sometimes be found for representing all the data. Daly calls this relationship "duality." Qualitatively, the relationship is similar to the concept of duality which "has been used extensively in Mathematics, Engineering and Statistics."[7, p. 79] We digress briefly to discuss an instance of Daly's duality at its extreme.

STATISTICAL PATTERN RECOGNITION

In statistical pattern recognition, certain assumptions are made about the structure of each class and how the classes are related to each other. An important special case is the following; the pattern space classes each have known multivariate normal distributions with known a priori probabilities, and it is desired to classify each data point using a Bayes' optimal classifier.[8] Pattern space is embedded in E^n, and there are M measurements. Consequently, for c classes, on the order of $c \cdot M \cdot n^2$ floating point multiplications and additions are required to classify the data. The dependence on n is thus dominant. The motivation for linear feature selection is the following well-known

Theorem. Let B be a linear transformation from E^n to E^k. If {x} is normally distributed in E^n with mean μ and covariance Σ, then {Bx} is normally distributed in E^k with mean $B\mu$ and covariance $B\Sigma B^T$.

Since it is easy to formulate an expression for the probability that a point is misclassified (under the assumption of normality), we can form a function depending only on B and the known statistics of the classes which should be minimized. The problem becomes to find k less than n (hopefully much less) and B so that the probability of misclassification (PMC_B) in E^k is acceptably close to the probability of misclassification (PMC) is pattern space. Unfortunately, the expressions for PMC_B and PMC are analytically and numerically intractable, and so a number of computable measures have been proposed for telling when class separation is lost after a linear dimensionality reduction. These are surveyed in Kanal,[9] with certain recent advances being noted by Decell and Guseman.[10] If we can assume that these measures do indeed have the hoped for relationship to PMC_B, then the duality of Dale assumes the form shown in Fig. 1.

Fig. 1 illustrates what we think Daly means by duality. Roughly speaking, classification in E^k is the inverse to B when B is restricted to classes in E^n. Pattern space, containing classes, is transformed to feature space in which class identification is performed. By the techniques of statistical pattern recognition the linear transformation B is determined which preserves class separation in the lower dimensional space.

What makes the last two paragraphs a digression is the lengthy list of patently false assumptions (false, at least, in our setting). In general, classes in the data are not usually distributed multivariate normal. Furthermore, we have no idea of the number of classes, much less what the parameter

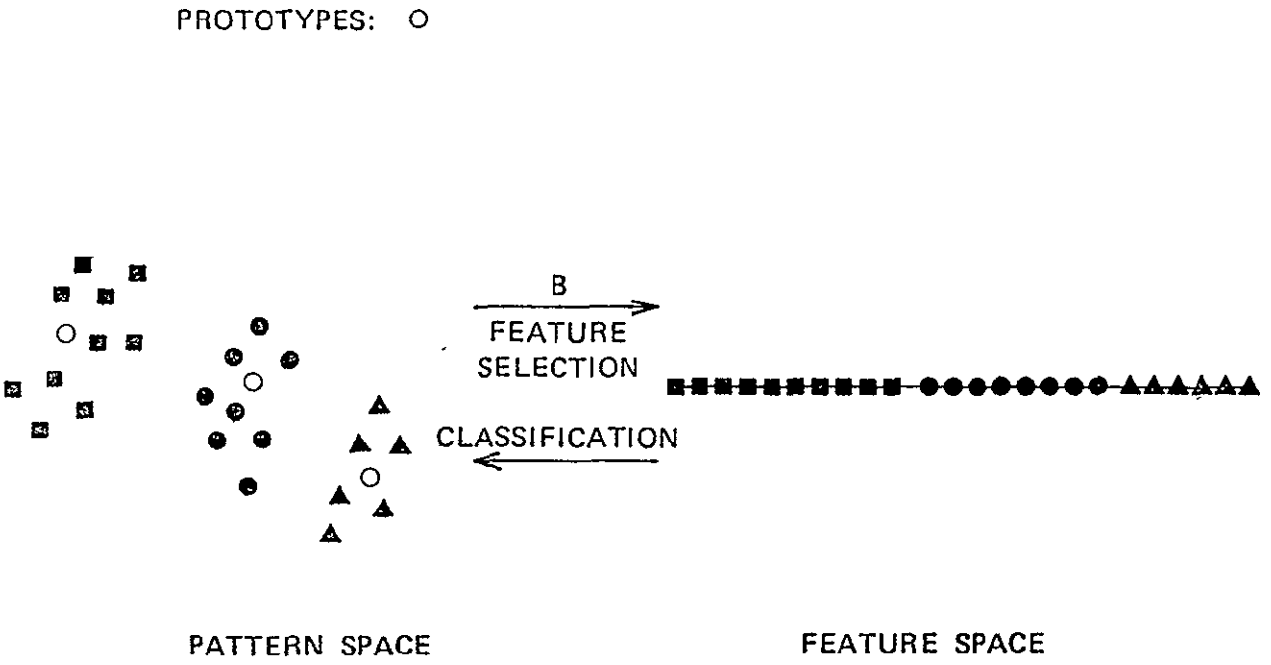

Fig. 1. Duality Between Feature Selection and Clustering

values for each class are. In addition, the separation measure in E^k which led to B is only qualitatively related to PMC_B. Daly, of course, does not have statistical pattern recognition in mind at all. His technique for feature selection is the principle components map (truncation in the Karhunen-Loève expansion), trained by the prototypes.

Class statistics in the parametric setting are often estimated by sampling the data. The samples must be labelled, and comprise what is usually called training data. There are serious questions on how multivariate statistics can be estimated from small training sets (see, for example, Kempthorne[11, pp. 11-28]). Thus many labelled samples must be furnished for this approach to statistical pattern recognition. The acquisition of enough labelled samples is often prohibitively costly.

A new approach to the problem of minimizing the increase in PMC following a dimensionality reduction has recently been formulated by Odell[12]. He assumes the data is continuous, dependent on a single parameter t, $0 \leq t \leq T$ and that the data is from one of a finite number m of known stochastic processes. Furthermore, the requirement for training data, in the absence of actual knowledge of the exact form of the m stochastic processes, remains.

PRINCIPLE COMPONENTS

In common with parametric methods, the principle components method requires samples in pattern space, but not labelled samples. The prototypes

should be selected to represent the "spread" of the data.[3, p. 313 et seqq.] One then forms the pooled covariance matrix Σ of the prototypes after the overall mean vector has been subtracted from each prototype. The principle components map from E^n to E^k based on these prototypes is the k×n matrix whose rows are orthonormal eigenvectors corresponding to the k largest eigenvalues of Σ. (Since Σ is non negative semi definite, the eigenvalues are real and non negative.) The principle components map is motivated by the following simply stated problem: Given p prototypes $y_1,\ldots,y_p$, find an orthonormal basis $B = \{e_1,\ldots,e_n\}$ such that, for every k, the number

$$g_k = \sum_{i=1}^{m} \left|\left| y_i - \sum_{j=1}^{k} \langle y_i | e_j \rangle e_j \right|\right|^2$$

is smallest, where $\langle u|v\rangle$ denotes the inner product in E^n. That is, the principle components feature selection technique can be thought of as arising from the minimization of a certain objective function.

There have been many recent books and papers[13-20] on using the principle components map in pattern recognition problems. Several of these papers use the principle components map in intimate association with cluster analysis, a use for which it is well suited. In many applications, however, the principle components map tends to destroy separation between data points. For example, consider the points (-1,0), (1,0), (0,α) and (0,-α) in E^2 where $0 < \alpha < 1$. The principle components map from E^2 to E^1 is simply B = (1,0). Thus both points (0,α) and (0,-α) are mapped into (0,0), and are therefore not separable in E^1. In this paper, our primary objective is to attempt to model the separation the prototypes enjoy in E^n after a suitable linear transformation to E^k. A secondary objective is to be able to apply the transformation to all of pattern space.

Actually, the principle idea is not new. Sammon[21,22] and Shepard[23] look at the problem from this same point of view (keeping the prototypes separated), but their methodology does not lead to a linear map. No method is given for transforming points other than the prototypes. With a linear map (such as is furnished by the principle components approach), dimensionality reduction of all data can be performed using $M \cdot n \cdot k$ multiplications and additions. In the next section we describe our objective and show how it can be achieved numerically. An outline of the steps for implementing the technique in a computer program is given along with an illustrative example of the technique applied to LANDSAT data.

THE OBJECTIVE FUNCION

Starting with the p distinct prototypes $\{y_1,\ldots,y_p\}$, $y_i \in E^n$, our objective is to preserve, as closely as possible, the Euclidean distances between linearly transformed prototypes in E^k. That is, if B is a k×n matrix, and $w_i = By_i$, then we seek to make $||w_i-w_j||$ as close as possible to $||y_i-y_j||$ for all $1 \leq i < j \leq p$. This idea leads to an objective function:

$$f(B) = \sum_{i\leq i<j\leq n} [||By_i-By_j||-||y_i-y_j||]^2 \quad .$$

The process of determining B is thus an optimization problem, which can be solved using standard programs such as Davidon-Fletcher-Powell.[24,25] For such programs the gradient of f is required. To this end, let $m = p(p-1)/2$ and let $\{z_i : i = 1,\ldots,m\}$ be the set of differences $y_\mu - y_\nu$ with $1 \leq \mu < \nu < p$. No z_i is zero since the prototypes are distinct. Let $b_\mu = (b_{\mu 1},\ldots,b_{\mu n})^T \in E^n$ denote the μ-th row of B. Let $\langle x|y\rangle = \sum_1^n x_i y_i$ denote the inner product in E^n.

Then straightforward calculation yields

$$\frac{\partial f}{\partial b_{\mu_0\nu_0}} = -2 \sum_{i=1}^{m} (||z_i|| - ||Bz_i||) \frac{\sum_{\nu=1}^{n} b_{\mu_0\nu} z_{i\nu} z_{1\nu_0}}{||Bz_i||}$$

$$= -2 \langle b_{\mu_0} \mid \sum_{i=1}^{m} \left(\frac{||z_i|| z_{i\nu_0}}{||Bz_i||} - z_{i\nu_0} \right) z_i \rangle$$

$$= -2 \langle b_{\mu_0} \mid \sum_{i=1}^{m} z_{i\nu_0} \frac{||z_i||}{||Bz_i||} z_i \rangle + 2 \langle b_{\mu_0} \mid \sum_{i=1}^{m} z_{i\nu_0} z_i \rangle .$$

Let $S = (s_{\mu\nu})_{n\times n}$ denote the scatter matrix $s_{\mu\nu} = \sum_{i=1}^{m} z_{i\nu} z_{i\mu}$, and let $T(B) = (t_{\mu\nu})_{n\times n}$ be defined by

$$t_{\mu\nu} = \sum_{i=1}^{m} \frac{||z_i||}{||Bz_i||} z_{i\nu} z_{i\mu} .$$

With this notation,

$$\frac{1}{2}\nabla f(B) = BS - BT(B) .$$

Remarks. 1. A proof that minima exist is given in the Appendix. However, even in simple cases f can have a poor local minimum. For example, consider again the four points $\pm(1,0)$ and $\pm(0,\alpha)$ in E^2 where $0 < \alpha < 1$. (Recall these yield the principle components map $(1,0)$.) The objective function $f(b_1,b_2)$ is easily computed and points at which $\nabla f = 0$ can be found using calculus. The symmetry of the problem leads to two distinct classes of minima. In class 1, points $\pm(1,0)$ are mapped outside $(0,\alpha)$ on an interval. These are

global minima. The other minima map points $\pm(0,\alpha)$ outside points $\pm(1,0)$. If α is small, these are very poor local minima. This is illustrated in Fig. 2.

2. The minimum is never unique if $n > 1$. Indeed, if A is $k \times k$ and orthogonal, then $f(AB) = f(B)$ for every B. As a result, there are fewer than $n \times k$ degrees of freedom in the choice of elements b_{ij} in so far as varying b_{ij} changes the value of f.

3. The minimization problem is to find B with $\nabla f(B) = 0$. This can also be written as a (nonlinear) fixed point problem. If $F(B) = B\,T(B)S^{-1}$, then a fixed point of F is a stationary point for f.

4. The nonlinear map F is homogeneous of degree zero. The problems of finding fixed points of such mappings are interesting and seem to have other applications, for example in iterative image processing. One consequence of this is that when B is a fixed point and $a > 0$ then $F(aB) = F(B) = B$; that is, one iteration starting at aB solves the fixed point problem. It would seem possible to use the special nature of this map to simplify the computational problems involved in finding fixed points.

A PROGRAM FOR FINDING B

The technique for finding B with f(B) minimum (and so for linearly modelling pattern space in E^k while preserving distances) has been coded in FORTRAN and tested on several unrelated data sets. We suppose the prototypes are given and $k<n$. (This may involve extensive preprocessing.)

Step 1. Form the set of difference vectors and balance the difference set by subtracting the overall mean vector of this set from each difference

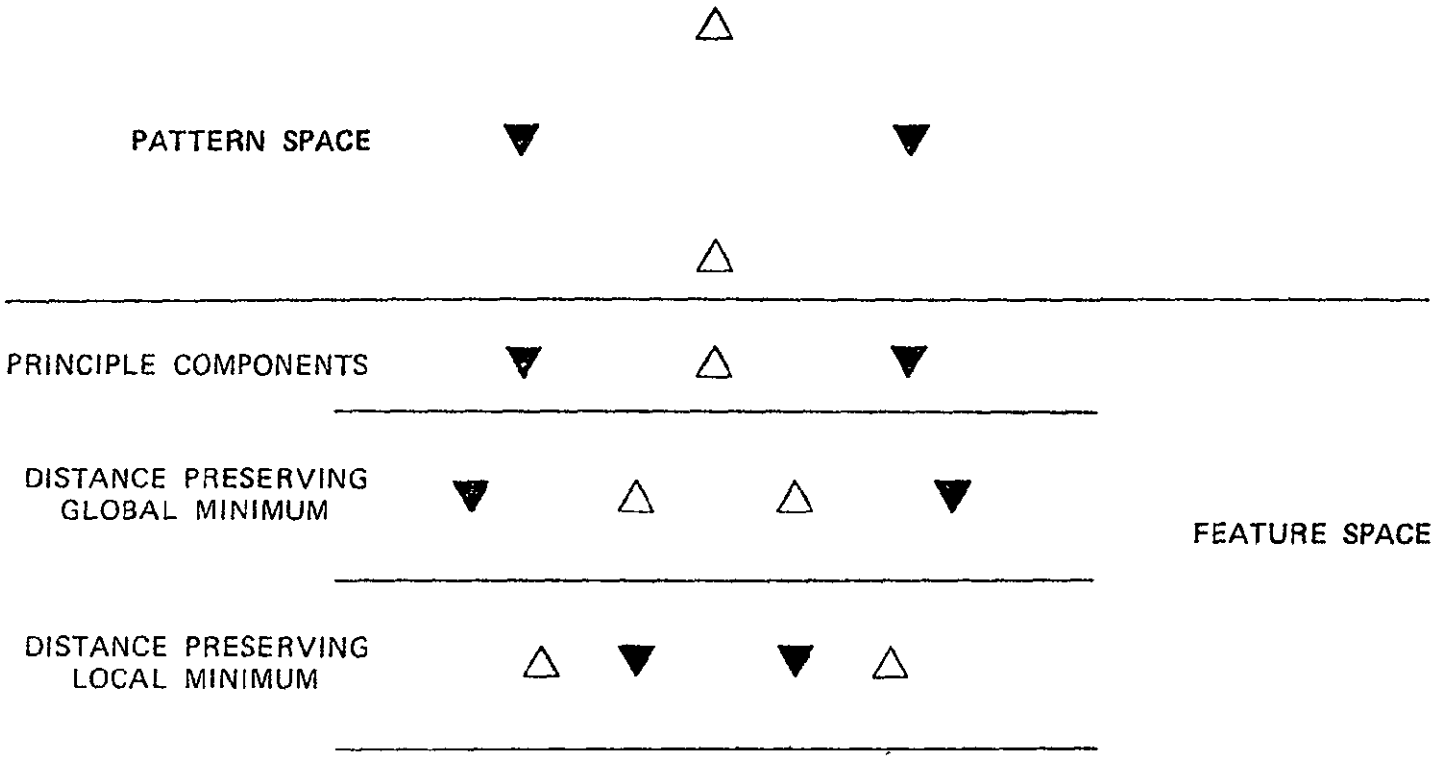

Fig. 2. Illustrating Distance Preservation, $E^2 \rightarrow E^1$.

vector. This translation, of course, does not change any distance.

Step 2. Form the scatter matrix of the balanced prototype differences. Except for the subtracted mean, this is the matrix S above.

Step 3. Find the eigenvalues λ_i and associated eigenvectors $e_i \in E^n$, $i = 1,\ldots,n$, of S ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$; note $\lambda_n \geq 0$. Let $B_0 = (e_1,\ldots,e_k)^T$ be the k×n matrix whose μ-th row is e_μ^T. Note that B_0 is the k×n principle components map based on the balanced prototype differences. This B_0 is used to start the optimization method or the fixed point iteration.

At this point, either of two steps may be followed:

Step 4A. Using the nonlinear map F, iterate: that is, let $B_i = F(B_{i-1})$, $i = 1,\ldots$.

Step 4B. Using the Davidon-Fletcher-Powell method, solve the problem $\nabla f(B) = 0$ starting at B_0. This optimization program has been found to be entirely reliable in this application, although two data dependent tolerances need to be set to prevent the program from making wasteful searches.

Step 5. Once satisfactory convergence is obtained, examine the error (value of the objective function) to decide whether to try a larger or possibly smaller value of k.

RESULTS

The program outlined here has been tried on several data sets. One of the most spectacular successes has been on multispectral multitemporal LANDSAT data. Since both M and n are large, any reduction in n will greatly improve classification efficiency. Kauth and Thomas,[26] and Wheeler et al.[27]

have independently found that each temporal acquisition of LANDSAT data is about two (rather than four) dimensional when restricted to problems of agricultural interest. This is also predictable based on the spectral correlation between bands 4 - 5 and 6 - 7. The two dimensionality of one pass LANDSAT data thus results from a combination of the two causes of low intrinsic dimensionality mentioned in the introduction. Knowing this, we are led to believe three pass data is at most 6 dimensional.

In the example reported here, the data consists of three registered passes of LANDSAT data consisting of about 23,000 pixels. A fast technique for selecting prototypes (from classes of agricultural interest) is the following: For each pixel inside the scene, compute Robert's gradient of distance 2 (see Haralick[5]) and consider the pixels in the lower five percent. (These pixels are unusually pure, that is, not mixtures.) In each pattern space coordinate find two pure pixels, one with greatest and one with least measurement value. This procedure furnishes 24 pixels which we use as prototypes (they were distinct). This gives 276 pairs. Using the technique discussed above, the mean error of best distance-preserving map from 12 to k dimensions is shown. Also shown is the mean error associated with the principle components map. Since our iterative technique uses the principle components map to start, it is easy to compare how the two preserve between-prototype distances. These results are shown in Table 1.

Looking at Table 1, it is difficult to resist a belief that this particular set of prototypes is four dimensional, and that even three dimensions are enough to preserve most of the between-prototype structure. The distance in E^{12} between points which differ by 1 in each coordinate is, of course,

Table 3. Mean Square of Difference Between $||Bz_i||$ and $||z_i||$

Reduced Dimension	Principle Component Map	Distance Preserving Map
1	1053.4	309.7
2	844.0	20.6
3	564.7	8.1
4	435.0	0.6
5	263.3	0.2
6	112.8	0.2

$\sqrt{12} = 3.46$. An RMS difference of $\sqrt{8.1} = 2.85$ corresponds to errors less than this in E^{12}, and a mean square difference of 0.6 indicates essentially perfect modelling of the prototypes. We doubt that this conclusion could have been obtained using the principle components map approach. That is, our approach models the data in lower dimensional space and provides a clear measure of how accurately the model preserves inter-prototype distances.

We have performed several other tests, for instance on Iris data and on marketing data (for hand-held calculators). Rather than report these here, we would suggest the interested reader take them as "things to do" in the sense of Hartigan.[3] Take your favorite data set in which the principle components approach has been effective after some study and manipulation, and try this approach. We think you will find obvious clusters of calculators in two dimensions, and that Iris data is essentially one dimensional.

SUMMARY

A new approach to linear feature selection is presented. The technique is operationally like the principle components approach, but differs in that the objective function is derived from an attempt to preserve the separation between prototypes in pattern space. The linear transformation is derived either by standard optimization programs or by fixed point iteration of a certain nonlinear mapping. The technique has been applied to several data sets. Particularly exciting is the application to three pass LANDSAT imagery. From the value of the objective function when the reduced dimensionality is three, we can infer that a very accurate representation of the data can be obtained in a single color image.

REFERENCES

1. H. C. Andrews, Introduction to Mathematical Techniques in Pattern Recognition. Wiley-Interscience, New York (1972).
2. M. R. Anderberg, Cluster Analysis for Applications. Probability and Mathematical Statistics, A Series of Monographs and Textbooks, number 16. Academic Press, New York (1973).
3. J. A. Hartigan, Clustering Algorithms. Wiley series in Probability and Mathematical Statistics. John Wiley and Sons, New York (1975).
4. B. S. Duran and P. L. Odell, Cluster Analysis: A Survey. Lecture Notes in Economics and Mathematical Systems, Managing Editors M. Beckmann and H. P. Künzi, Volume 100, Springer-Verlag, Berlin-Heidelberg-New York (1974).
5. R. M. Haralick and I. Dinstein, A spatial clustering procedure for multi-image data, IEEE Trans: on Circuits and Systems, CAS-22, 440-450.(1975).
6. G. H. Ball and D. J. Hall, ISODATA, a novel method of data analysis and pattern classification, Technical Report, Stanford Research Inst., Menlo Park, California. 72 pp. (1965).
7. J. A. Daly, Some dual problems in pattern recognition, Pattern Recognition 3, 73-84 (1971).
8. T. W. Anderson, An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, New York (1958).
9. L. Kanal, Patterns in pattern recognition, IEEE Trans. Pattern Recognition, IT-20, 697-722 (1974).

10. H. P. Decell, Jr. and L. F. Guseman, Jr., Linear feature selection with applications, Pattern Recognition 11, 55-63 (1979).

11. O. Kempthorne, Some aspects of statistics, sampling and randomization, in Contributions to Survey Sampling and Applied Statistics, Papers in Honor of H. O. Hartley, H. A. David, ed., Academic Press, New York, New York (1978).

12. P. L. Odell, A model for dimension reduction in pattern recognition using continuous data, Pattern Recognition 11, 51-54 (1979).

13. Y. T. Chen and K. S. Fu, Selection and ordering of feature observations in a pattern recognition system, Inf. Control 12, 394-414 (1968).

14. H. H. Harmon, Modern Factor Analysis, Univ. of Chicago Press, Chicago (1960).

15. M. Shimura and T. Imai, Nonsupervised classification using the principle component, Pattern Recognition 5, 353-363 (1973).

16. B. R. Kowalski and C. F. Bender, An orthogonal feature selection method, Pattern Recognition 8, 1-4 (1976).

17. S. S. Yau and S. C. Chang, A direct method for cluster analysis, Pattern Recognition 7, 215-224 (1975).

18. S. Wold, Pattern recognition by means of disjoint principle components models, Pattern Recognition 8, 127-139 (1976).

19. G. E. Lowitz, Stability and dimensionality of Karhunen-Loève multispectral image expansions, Pattern Recognition 10, 359-363 (1978).

20. C. M. Hay and R. W. Thomas, Development of techniques for producing static strata maps and development of photointerpretive methods based on multitemporal LANDSAT data, Remote Sensing Research Programs, Space Sciences Laboratory, Series 19, Issue 1, University of California, Berkeley (1977).

21. J. W. Sammon, Jr., A. H. Proctor and D. F. Roberts, An interactive-graphic subsystem for pattern analysis, Pattern Recognition 3, 37-52 (1971).

22. J. W. Sammon, Jr., A non-linear mapping for data structure analysis, IEEE Trans. Computers C-18, 401-409 (1969).

23. R. N. Shepard, The analysis of proximities: Multidimensional scaling with an unknown distance function. I, Psychometrika 27, 219-246 (1962).

24. R. Fletcher and M. J. D. Powell, A rapidly convergent descent method for minimization, Comput. J. 6, 163-168 (1963).

25. J. E. Dennis, Jr. and Jorge J. Moré, Quasi-Newton methods, motivation and theory, SIAM Review 19, 46-89 (1977).

26. R. J. Kauth and G. S. Thomas, The tasselled cap--a graphic description of the spectral-temporal development of agricultural crops as seen by LANDSAT, LARS Symposium: Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana (1976).

27. S. G. Wheeler, P. N. Misra and Q. A. Holmes, Linear dimensionality of LANDSAT agricultural data, LARS Symposium: Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana (1976).

28. P. R. Halmos, Finite Dimensional Vector Spaces, Undergraduate texts in mathematics, second edition, Van Nostrand, Princeton, N.J. (1958).

29. Martin Schecter, Principles of Functional Analysis, Academic Press, New York, N.Y. (1971).

APPENDIX

We prove here that the function f has a minimum.

Theorem. Let $Z = \{z_1,\ldots,z_\ell\}$ be a finite set of vectors in E^n, let T_{nk} be the set of linear transformations from E^n to E^k, and define $f(B) = \sum_{i=1}^{\ell}[||Bz_i||-||z_i||]^2$, $B \in T_{nk}$. Then there exists B_0 such that for all B, $f(B) \geq f(B_0)$.

Proof: Let H be the m-dimensional subspace of E^n spanned by Z. Let P denote the projection of E^n onto H along $H^\perp$ (see [28], pp. 146, et seqq.). When restricted to H, P is one to one and, for all $x \in E^n$, $||Px|| \leq ||x||$. Since H is m-dimensional, H is linearly isometric to E^m; let $U : H \to E^m$ denote one such isometry. The map $A = UP : E^n \to E^m$ is one to one when restricted to Z. Let $W = AZ$ and $w = Az$; clearly $||w|| = ||UPz|| = ||Pz|| \leq ||z||$, with equality if $z \in H$. In particular, $||w_i|| = ||z_i||$ if $w_i = Az_i$, $z_i \in Z$, and W spans E^m. Let $f_m : T_{mk} \to R^+$ be defined as in the statement of the theorem (with m replacing n). Let A^+ be any linear transformation from E^m to E^n such that $A^+ w_i = z_i$, $i = 1,\ldots,\ell$. Then

$$f(B) = \sum_{i=1}^{\ell}[||Bz_i|| - ||z_i||]^2 = \sum_{i=1}^{\ell}[||BA^+w_i|| - ||w_i||]^2 = f_m(B'),$$

where $B' = BA^+ \in T_{mk}$. Similarly, if $B' \in T_{mk}$, then $B = B'A$ has $f(B) = f_m(B')$. Thus minima of f correspond exactly to minima of f_m, and it suffices to show f_m has a minimum.

For $B \in T_{mk}$, define $||B|| = \sup\{||Bx|| : ||x|| = 1\}$. Under this norm, closed bounded sets in T_{mk} are compact. Clearly, the function f is continuous. Thus it is sufficient to show the closed set $K = \{B : f_m(B) \leq f_m(0) = C\}$ is bounded in T_{mk}. Suppose not. Then there is a sequence $\{B_j\}$ in K with $||B_j|| \to \infty$ and $f_m(B_j) \leq C$. By definition of f_m, this implies $\{||B_j w_i||\}$ is bounded for each fixed i. Since W contains a basis and each B_j is linear, $\{||B_j w||\}$ is bounded for each $w \in E^m$. By the uniform boundedness theorem[29], the sequence of norms $\{||B_j||\}$ is uniformly bounded, contrary to $||B_j|| \to \infty$. Thus K is compact and f_m has a minimum.

Remark: Note $f(B_0) = 0$ if $m \leq k$.

The Easy Remote Sensing Problem

Jack Bryant
Department of Mathematics
Texas A&M University
College Station, Texas 77843

Report #20

Prepared For

Earth Observations Division
NASA/Johnson Space Center
Houston, Texas
Contract NAS-9-14689-9S

D-2

1 Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4 Title and Subtitle	5. Report Date
The Easy Remote Sensing Problem	August, 1979
	6. Performing Organization Code
7 Author(s)	8. Performing Organization Report No.
Jack Bryant	20
9 Performing Organization Name and Address	10 Work Unit No
Department of Mathematics Texas A&M University College Station, Texas 77801	11. Contract or Grant No. NAS-9-14689-9S
12 Sponsoring Agency Name and Address	13 Type of Report and Period Covered
Earth Observations Division NASA/Johnson Space Center Houston, Texas 77058	Unscheduled Technical
	14. Sponsoring Agency Code

15 Supplementary Notes

Principal Investigator: L. F. Guseman, Jr.

16 Abstract

A model for multi-imagery applicable to LANDSAT multi-spectral multi-temporal data taken from agricultural areas with flat terrain is proposed. An unsupervised classification problem is defined in the setting of the model. Methods to solve this problem are discussed. A computer program incorporating the methods which solves the problem is described. The methodology unifies the spatial, temporal and measurement-space aspects of the data analysis problem. The paper includes a lengthy expository introduction motivating the approach.

17 Key Words (Suggested by Author(s))	18. Distribution Statement
Automatic cluster analysis, Spatial clustering technique, Field detection, Non-parametric clustering, Multitemporal multispectral data, Remote sensing, Large area crop inventory, Agricultural scene modelling	

19 Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
		50	

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

ROUGH DRAFT

Background

Early in the space age, it was recognized that the view of Earth from space had many potential applications. Oil companies wanted to seek promising geological formations, foresters were worried about the environmental consequences of clear-cutting, commodities brokers recognized profits in advanced yield information. . . and government: the military, in the remote sensing business since the days of balloon flight, was quick to acquire and use high resolution systems. In the United States, a systematic program, beginning with the Tiros and Nimbus series of satellites, to chart cloud formation, movement, and dissipation in an attempt to improve weather prediction, was undertaken. There is no doubt this attempt has proved successful: the "big picture" from space of areas on Earth not otherwise visible has been the key.

However, neither the high resolution military systems nor the low resolution weather satellite data furnished the imagery required for prospecting, crop monitoring, land resource management and so on. This void was filled by the Earth Resources Technology Satellite program (1972), later named LANDSAT, and scheduled to be replaced by Landsat-D with higher performance[1] (and much more data) by late 1981. These programs have been and will probably remain under NASA management in a developmental state until 1985. Even in the developmental state, the data has been a useful tool for commercial ventures.[2] The Corn Blight Watch[3] of 1970 and the Large Area Crop Inventory Experiment (LACIE)[4] showed that remotely sensed data could be effectively used in agricultural applications. Both of these

required the collection and management of huge amounts of multi-image data; LACIE, in addition, required large quantities of ancillary information. One principal finding from LACIE was that the data management problem could be solved in a timely fashion.

The Corn Blight Watch and the LACIE both concern agriculture. Agricultural applications of remote sensing are unique for several reasons. The most obvious is that in many parts of the world crops are grown in large fields. Within a field under active cultivation, remotely sensed measurements are relatively much more homogeneous spectrally than within other areas, giving these areas smooth textural appearance. Also, in many parts of the world, the underlying terrain is essentially flat. This suggests that a field in one position in a multi-image which is spectrally like a field in another is probably in the same class. Contrariwise, spectrally different fields are probably in different classes, because the reflectance properties of the ground cover and not artifacts from the terrain (such as shadows or vastly different soil types) account for the spectral separation.

There are, of course, at least four undefined terms in the preceeding paragraph: field, spectrally like, spectrally different and class. It might be argued that these terms have their usual definitions. This means, ultimately, that human intervention is required to decide whether this blob is a field, those are alike, etc. It is the purpose of this paper to suggest that these terms can be defined implicitly; that is, in this limited setting (flat terrain, large fields), the relationship between these terms can be defined logically. What is surprising is that these

essentially philosophical issues lead directly to an automated multi-image data analysis procedure.

The Data Analysis Problem

Before going on, it is instructive to examine some of what an analyst might go through to understand multi-image data. The analyst is equipped with several temporal acquisitions of Landsat data which have been registered to one another. The imagery is available in two forms: as false color transparencies approximating color infra-red film and as the actual Landsat measurements accessible by computer. With only this, the analyst interested exclusively in agriculture can give little more than the interpretation of the phenological concepts of class, field, like and unlike. Of course, cities, mountain-tops, lakes and so on can be recognized at a glance; but the issue here is much more complex: it is desired to

(a) label the fields with crop type;

(b) estimate the condition of each field;

(c) estimate the total area by crop type; and,

(d) predict the yield at harvest.

It has been determined[5] that these questions cannot be resolved without ancillary information, including but not limited to

(1) crop calendar information for the area including this and recent years;

(2) recent history of crop proportions;

(3) regional crop rotation, cultivation and irrigation practices;

(4) a history of meterological events which might affect crop development.

The analysis and interpretation of imagery is creative: it is in no rigid way determined by given circumstances. The best and for the foreseable future only way to understand remotely sensed agricultural information is to use a human analyst so that this creative process is nourished, not stifled. It is in this spirit that I approach these problems: what part of the analyst's job is tedious, boring, almost inhumane? Can any of it be replaced by automated processing? Is there anything about the data which is independent of ancillary information that can be extracted automatically?

One operation an analyst will find nearly impossible is to label and specify the properties of an isolated image element without knowing how the spatial neighbors go. In a single image, it is easy for the analyst to locate points interior to fields for interpretation; it is much harder to find interior points for multi-imagery. The computer, however, has no such problem. A properly instructed computer can find areas which are relatively homogeneous and prepare image products which identify such areas for the analyst. Thus, even with multi-imagery, analysts can first concentrate on whole regions rather than on isolated picture-elements, using their creativity where it is most suited rather than trying to correlate the spectral-temporal behavior of a spatially isolated element. (It is not surprising that the so-called "Procedure 1"[6] met with such analyst resistance. This statistically-inspired methodology <u>forced</u> an analyst to deal with individual multi-image elements with their location specified a priori.)

Another difficult task for an analyst is to identify when another area in the picture has similar spectral-temporal behavior as one already processed. This might at first seem trivial; after all, the computer "knows" the data, and even a fairly simple interactive system will allow the analyst to interrogate the data values particular pixels enjoy. But problems arise: large images contain a lot of data, usually more than can be economically kept in the computer's random access memory. Thus these questions may require considerable real time overhead to answer, so that the analyst may attempt to match the false multi-image colors rather than raw data. The film products, however, have poor relation to data values because of severe nonlinearity of the video-film system (including saturation) and (perhaps more to the point) because of subjective local[7] and large scale[8] reaction to color. Thus actually identical colors (with different surrounding color) appear different and actually well separated data values appear to produce the same apparent color on the film products. In addition, it is difficult to keep in mind just where everything is, so without a fancy interactive system much tedious counting may be involved.

However, one's first thought about this problem is actually correct: we <u>can</u> automatically identify which fields are like spectrally. The secret is that the scene itself tells us when sample multi-image elements are alike: namely, once spatially connected spectrally homogeneous fields have been found, then samples from a fixed field should belong to the same real class. Given any assignment rule based partly or entirely on spectral-temporal behavior, the assignment of elements to classes should preserve the

spatially derived fact that samples from the same field are actually in the same class. If it does this and still distinguishes different classes well, then spatially separated fields which are assigned to the same class by the classifier can be believed to be in the same real class.

The last paragraph contains the germ of the idea behind the new clustering and classification program AMOEBA.[9] In that paper, there is a brief discussion of a model for Landsat agricultural data. Here, we go deeper into the underlying assumptions. We first note in detail the properties Landsat data has which must be considered before automatic processing can be seriously considered.

The Data

Landsat data for a given area on Earth is acquired about once every 18 days. Because of overlap, it sometimes happens that an area will be seen on successive days. Four spectral bands are sensed: two in the visible and two in near infra-red. The dynamic range of real world reflectance is such that the data collected must be clipped on board the satellite. There are six sensors for each band, which are switched to the lens system by a rotating mirror. This leads to calibration problems. Because the sensors are not identical, they do not respond in the same way to incoming photons, and so this is adjusted by those who preprocess the data. Residual nonuniformities appear as stripping down scan lines in an unregistered film product. Another problem is that every six lines the registration between the satellite and the ground jumps by exactly one image element. This amounts to about ten meters displacement per scan

line with a nearly 60 meter jump at the end of six scan lines. A straight narrow high contrast feature such as a road with vegetation on each side will show this distortion even on the film products. Let us call this displacement jitter.

The periodic scan line distortions degrade the resolution in a single image. In multi-temporal imagery, registration has been performed, and even the periodicity of the distortion is lost. Scan line stripping jumps around, the formerly predictable saw-tooth roads now appear wormy and so on. Since the underlying crop identification-yield prediction problem cannot be solved using single pass data, these errors must be faced and handled. One result of the jitter is that registration cannot in general be very good: certainly if peak displacements are judged, then no better than ± 2 image element registration can be hoped for. Thus, while the nominal resolution of single pass data may be 80 meters, the effective resolution of multi-imagery is probably no better than 120 meters, if it is even possible to speak of resolution in the presence of these highly nonlinear distortions.

At first, a ± 2 peak displacement might seem too pessimistic. In order to obtain a quantitive feel for how bad the registration might be, assume that of the eight nearest neighbors a point has, the probability that the registration places the resolution cell being viewed within 25 meters of the image element to which it is registered is about 1/3. One might arbitrarily assign probabilities of 1/8 to each of the four nearest neighbors and then 1/24 to each of the next four nearest. Since each temporal

acquisition is spatially independent of each other, the jitter is also independent. Thus, even if each pass can be registered with this much precision, the absolute theoretical limit in this example, with three or more pass data the probability that a point is, in all passes, registered within one image element of each other pass decreases rapidly with the number of passes. Even for three pass data, this probability is 0.60 (under the assumptions of misregistration probability mentioned above); with four passes, the probability that a pixel is, in all four passes, registered no worse than one image element each pass to each other pass is 0.22. This is in spite of the fact that the RMS registration error in this simple model is $(\frac{1}{24}(8\cdot 0^2+3\cdot 4\cdot 1^2+4\cdot(\sqrt{2})^2))^{1/2} \simeq 0.91$, about that found in real data. The conclusion that registration errors of more than one image element are common seems rather firm.

One immediate consequence of the registration problem is that a single sample of multi-image data obtained by image element is not a set of averaged measurements from an area on the ground. Not at all. Even if the scene consisted of exactly two classes, with no problems of resolution, the samples would not be samples in the statistical sense. Imagine a crowd of animals (say lions and sheep). By a sample (of one), we mean the selection of one animal from the crowd. In the situation here, however, we select a head from one, a body from a neighbor, ... Even if a selection happened to extract all parts from the class lions, there might be the head of an adult male attached to the body of a female. Such a creature would not be a lion at all. Similarly, samples from Landsat multi-temporal data are not samples in the statistical sense.

Since the point I want to make seems to be hard to grasp, let me give an example in a less fanciful setting. In a real wheat field, there will usually be parts which do better than other parts. There is no need, here, to suggest reasons why: it is given. Suppose we are looking at averaged greenness over one-acre areas within the field. The wheat was planted and harvested on the same day throughout the field. In spite of this, the rate of development over one of the acres is quite different from a neighboring one. If we assume four temporal acquisitions, the conditions which prevail on the ground might be as follows:

	1	2	3	4
Sample 1	0 green	.4 green	.6 green	0 green
Sample 2	0 green	.6 green	.8 green	0 green

These samples are close spatially in this example; suppose, because of registration errors, the measurements from sample 1 and sample 2 got interchanged in acquisition 2:

	1	2	3	4
Measurement 1	0	.6	.6	0
Measurement 2	0	.4	.8	0

An analyst, observing this situation, would have two comments. Measurement 1 seems unlike wheat because of the plateau in development. But measurement 2 is also strange: wheat does not usually develop so rapidly.

These problems, let us review, are not a consequence of poor resolution, inadequate sampling density or measurement errors. These problems are due to registration errors. The statistical implications are hard to grasp at

once. The main implication is that off block-diagonal terms in the covariance matrix are garbage.

An even nastier problem accompanies image elements which are near the physical boundary between two classes. Their population can be modelled easily in the absence of registration errors, although current statistically based classification methodology ignores this fact. Registration errors, even very slight ones, defeat completely the model. The best we can do is simplistic but seems to work well. We will return to the mixture problem later. Also, registration errors must be considered when searching for boundaries between fields. This, too, will be addressed later.

One other artifact deserves mention: during the calibration process (over which the analyst has no control), the data is mapped from one digital image to another. The output image histogram will almost always have strange properties. To illustrate, suppose the input histogram in a single band were shaped like an inverted "w". The input data might range from values 10 to 50 (see Fig. 1). Suppose in calibration the data is scaled linearly from 8 to 64. Now the output histogram has missing values. On the other hand, were the data scaled linearly from 16 to 45, the output histogram would have narrow false peaks. Both of these conditions are found in real data, somewhat limiting the usefulness of mode-seeking multi-dimensional clustering techniques.[10] Also, parallelepiped classification techniques[11] are adversely affected by this distortion which causes a measurement to jump around wildly from scan line to scan line.

One way to overcome this measurement distortion is to purposely limit the range of incoming values; for example, all data could be divided by 3.

Another is to take combinations of measurements. One-third of the sum of the first three channels of Landsat data, and one-half of the sum of the two infra-red channels minus the second visible, have been found experimentally to have about the same variability from field center to field center, and to correlate well with the behavior of growing plants. The second combination amounts to a template match filter for chlorophyll; this will be large when ground measurements are green. The first is large when ground reflectance is high, that is, when the scene is bright. Most workers[12,13] have found that brightness-greenness space is more appropriate for agricultural pattern recognition problems than the higher dimensional untransformed space, although recently, Hlavka et al.[14] find otherwise. In any case, the simple $(C_1+C_2+C_3)/3$, $(-C_2+C_3+C_4)/2$ transformation does overcome the digital range distortion almost completely, and weights brightness and greenness so that the variability of each is roughly the same.

There are other sources of distortion or noise in the data. Clouds and cloud shadows, for example, are clearly not of agricultural interest. and appear at random to get in the way. Perhaps more severe is haze, which causes even successive day views of the same area to appear drastically different. The affect of the sun angle must also be considered by the analyst. These sources of distortion are much easier to overcome than the highly nonlinear ones mentioned before, and will be ignored here. In fact, it is said that haze and sun angle correlation can now be performed automatically.[15]

Before attempting to define the "Easy Remote Sensing Problem," we need to make more precise some of the terms we have been using, and introduce

some new ones. Most of the definitions given here are narrower than the standard ones[16] because we do have in mind Landsat data. At the same time, we continue our discussion of special properties Landsat data has.

A digital multi-image is a finite rxcxd array I ; r is the number of rows, c the number of columns and d is the number of measurements associated with each pixel (formerly picture-element). There is an approximately known correspondance between the pixels in the image and coordinates in the real world, but as noted before, the uncertainty might easily exceed 100 meters. The measurements themselves are averages, in the standard optical sense, of spectrally filtered real world reflectances. The spatial modulation transfer function of the system before digital operations begin can be assumed given. From what we estimate about registration, the main limit to resolution is not optical; in fact, the data seems to be slightly under sampled from the optical point of view. The data is sampled in a particular direction; to fix ideas, we shall think of the rows as being scan lines. When the data is registered, this only approximately remains true. One reason for locating the scan line direction is that data down scan lines is sometimes subjected to bandwidth limited processing which may distort the system modulation transfer function; this does not seem to be a major issue with Landsat data.

A partition of an image is a collection of pairwise disjoint subsets of the image whose union is the image. A clustering of an image is a partition such that members of the same subset are alike and members of distinct subsets are different. A cluster map is a map C of indices 0 - ℓ the same size as the image. The pixel at (n,m) is said to be in

cluster $C(n,m)$. We do not necessarily assume each cluster is non-void. Index 0 is reserved for a special cluster of pixels distinguished by the fact that their temporal-spectral-spatial behavior cannot be understood. A clustering technique is a systematic method for constructing cluster maps. A clustering technique is supervised is human intervention is required during at least part of the method. Supervision can be more or less interactive. One example of supervision might be to terminate iterations in an iterative procedure such as ISODATA[17]; another might be to tell the clustering technique that it should merge clusters i and j , or try to split cluster k . An unsupervised clustering technique is said to be automatic if no user-supplied parameters are required.

A label is a tag with meaning in the real world applied to spatial associations of pixels. Examples might be "mountain ridge" or "wheat field." The general problem of agricultural scene understanding is to cluster the data and label the clusters. When more is asked of the accuracy or detail of the labels, the problem becomes more difficult. Examples of labels in a relatively easy agricultural labelling problem might be "non-agricultural," "small grains" and "non-small grains." The labels "wheat," "barley," "corn," "soybeans," "sunflowers" and "other" correspond to a much more difficult problem. One property the agricultural labelling problem has which distinguishes it from others is that the objects of interest are fields on the ground, and variability within a field is much less than variability between distinct fields. Let us attempt to relate this to the data independently of the labelling problem.

The Model

We first require some topological concepts. Two pixels are neighbors if their spatial coordinates differ by exactly one. (A pixel inside the image has exactly four neighbors.) A path in the image is a sequence $p_1,\ldots,p_i$ of pixels such that p_{j-1} is a neighbor of p_j for $j = 2,\ldots,i$. A set P of pixels is said to be connected if for each p,q in P there is a path $p_1,\ldots,p_i$ in P with $p = p_1$ and $p_i = q$. A component K of a subset S of the image is a maximal connected subset. Each component K is simply the set of all pixels which can be joined to an arbitrary pixel in K with a path lying entirely in S. Each subset S of the image is the union of uniquely determined pairwise disjoint components.

Axiom 1 Real classes exist.

In the above, there is a discussion of within-field variability. Let R denote the set of pixels in the image which have the property that each measurement is taken from the same class; R is well-defined, although unknown. In an agricultural scene, a point on a spatial boundary between two real classes will not itself be in a real class. This serves to limit our model to agricultural-like images.

Axiom 2 If p and q are neighbors and are in R then p and q are in the same real class.

Let P be the set of pixels in R such that each neighbor of a point in P is in R. The elements of P are called pure pixels. They have the property that they are very like their neighbors, for each neighbor of a

pure pixel is in the same real class. A patch is a component of P. Obviously, each pair of points in a patch is in the same real class.

The following assumption formalizes the "large fields" condition. A square field of about 250 x 250 meters (about 14 pixels in area) will probably correspond to a patch of one to four elements. If all fields in a given real class are smaller than 200 x 200 meters (about 10 acres), then this real class will probably not be found in the set of patches.

Axiom 3 Each real class is represented in at least one patch.

The next axiom formalizes the "flat terrain" assumption.

Axiom 4 The real class associated with a patch does not depend on the location of the patch within the image.

This implies that if an area containing a patch is copied in the image to another location then the new patch and old patch will belong to the same real class. This property is rather like large scale translation invariance of the concept of a real class.

Locating Pure Pixels

The solution of practical problem of finding the set R requires bringing together the spectral, temporal and spatial aspects of the data. We continue to assume the scene is dominated by agricultural activity. The first observation is:

Note 1. A boundary between fields in real data almost always accompanies a significant change in spectral values for at least one acquisition. That is, adjacent fields may be close in all but one measurement and still be

identified as being distinct. This simple remark has a practical consequence with startling power: namely, boundaries are best recognized by multiple univariant gradient thresholding. This is helpful because:

Note 2. If boundaries between fields can be recognized automatically, then the patches of the model can be extracted from the data without supervision.

The artificial intelligence community has been active in the boundary detection problem for many years. One paper[18] is of particular interest because of the thorough parametric comparison of three gradients contained therein. In order to describe these, denote by $||u||_1$ the norm $\sum_{i=1}^{d} |u_i|$ of a vector $u = (u_1,\ldots,u_d)$. Pixels are examples of d-dimensional vectors. Let $P(i,j)$ denote the pixel with row index i, column index j. Robert's gradient[19] is defined by $R(i,j) = ||P(i,j) - P(i+1,j+1)||_1 + ||P(i+1,j) - P(i,j+1)||_1$. The extended Robert's gradient is

$$E(i,j) = ||P(i-1,j) - P(i+1,j)||_1 + ||P(i,j-1) - P(i,j+1)||_1 .$$

The maximum gradient is

$$M(i,j) = \max\{||P(i,j)-P(k,\ell)||_1 : k=i \text{ and } \ell=j+1 \text{ or } k=i+1 \text{ and } |\ell-j| \leq 1\}.$$

(See Fig. 2) Although the extended Robert's gradient does not use the actual value of the measurement at the point, it did experimentally lead to gradient thresholded imagery much more credible. The data in question was from one temporal acquisition. The observation in Note 1 suggests a vector "gradient," defined on pairs of neighboring pixels rather than at a single pixel. If u is a vector, define the vector $A(u)$ by $A(u) = (|u_1|, \ldots, |u_d|)$. If u

and v are vectors, write $u > v$ provided for some i, $u_i > v_i$. Let H and V be vector tolerances (the derivation of which will be described presently). Neighboring pixels p and q in the same row will be marked boundary pixels provided $A(p-q) > H$; neighbors p and q in the same column will be marked boundaries if $A(p-q) > V$. The tolerances H and V can be determined by sampling the data and accumulating the average A_h and A_v vector distance from a pixel to the right and (respectively) above neighbor. A constant k can then be determined experimentally so that $H = k\,A_h$ and $V = k\,A_v$ are appropriate vector tolerances.

Sometimes it happens in the nearest neighbor registration process that a line of pixels looks exactly like a line of neighbors in one acquisition. This pattern can interfere with the boundary pattern, sometimes leading to one or two pixel missed boundaries. Accordingly, once the thresholded boundary pixels have been marked, these cracks which otherwise join larger areas toplogically should be filled. The logic is very simple: if a pixel is not marked 'boundary' and the above and below or left and right neighbors are, then mark the pixel boundary.

The evolution of this method of boundary detection has a mildly interesting history which will be recorded here. We initially used Robert's extended gradient to mark single pixels as boundaries. This was improved by changing $||u||_1$ to $||u||_2^2 = u_1^2 + \ldots + u_d^2$ in the definition above. The next improvement of note was to threshold $||u - v||_2^2$ for neighboring pixels u and v and mark both (not merely one) as boundary pixels when the threshold was exceeded. Then, motivated by Note 1, we replaced $||u||_2^2$ by $||u||_\infty = \max\{|u_1|,\ldots,|u_d|\}$, so that $||u-v||_\infty$ was thresholded.

Finally, vector thresholds were introduced. If should be remarked that, while the general topological nature of the patches remaining after boundary detection did not change much, the agreement with ground truth improved with each change. Along the way we tried and discarded other gradients as well as sophisticated boundary following techniques.[20] Roughly equal to Robert's extended gradient was the more rational maximum distance gradient $M^*(i,j) = \text{Max}\{||P(i,j)-P(k,\ell)||_1 : |i-k|+|j-\ell| = 1\}$. Boundary following probably failed because of registration errors. A boundary recognized in a single image was lost in the multi-imagery, while false boundary segments were found. Attempts to reconstruct multi-image boundaries from logical combinations of multiple single image boundaries were generally disappointing. Although much of this work was unsuccessful, it did have one positive benefit. It convinced us that thin boundaries can not be found, so we had to be satisfied with thick ones. Pixels for which the hypothesis that, in all passes, each neighbor is in the same real class, can be doubted should be marked as possible boundaries, and what remained unmarked would serve as the set of pure pixels.

In Fig. 3, the average brightness of a four pass Landsat image is shown. (This one dimensional image superficially resembles each of the four images.) Fig. 4 shows the pure pixel structure obtained by the strategy outlined in this section. Maximum patch size was limited to 50 by the software which grew the patches.

Clustering

So far, we have coolly evaded what is actually at the heart of this data analysis problem. True, we have formalized assumptions about what form reality takes in the agricultural scene understanding problem. We have proposed a method by which the patches of the model can be found. We have eschewed the application of the methods of mathematical statistics[21] since we recognize that our pixels are not samples in the sense of this methodology. We have, by using the term "real class," intended to suggest a label, but not necessarily "the" label. That, of course, is the problem.

The trouble with our real classes is there is no way to tell when distinct patches correspond to different real class. The hint from Axiom 4 is not enough: this axiom only tells when two must be the same, not when they might, and the condition is far too severe. In order to motivate the idea which leads to a method for finding distinct real classes, let us return to LACIE.

In that experiment, the initial emphasis was on a two class problem: wheat and other (i.e. non-wheat). Wheat production was a sensitive political issue in those days, and the program attracted political support. It was finally noticed in LACIE that wheat and barley were difficult to distinuish on the ground, and therefore more so from over a thousand kilometers. At the start of LACIE, no one would have argued that wheat and barley are the same label. At the end, they were routinely regarded as indistinguishable sublabels of the class small grains. The separation between wheat and barley in an area in which both are grown is much less than the separation between a wheat field planted in early April and one planted in May. The point of all this is there may well be several classes of small grains

detectable from the imagery acquired on a few given dates, but it is too much to hope one could be labelled barley.

Since this point seems to be the source of some confusion, it can be repeated: by "real class" I mean a natural real class. Wheat and barley are obviously different labels, but these labels apply to microscopic differences on the ground, and not at all to remotely sensed measurements. Our real classes must be observable in the measurements. Put another way, in clustering we are looking for natural structure: we are asking that the date itself suggest the clustering.[22,23]

It is an open question at present how this relates to the subjective but all-important expectations of the analyst. An analyst will actually see certain patterns in the data which are expected to be clustered alike and certain others which should be different. Let us call the disagreements between the analyst and the clustering mistakes. It is common to recognize two distinct cases:

Type I mistake: Clusters different analyst-labelled associations in the same cluster.

Type II mistake: Clusters same analyst-labelled associations in different clusters.

Obviously, it is difficult for an analyst to correct a type I mistake, for the clustering has identified areas the analyst recognizes as being distinct, and therefore the analyst must rework these areas almost as if clustering had not been performed. Type II mistakes are usually a consequence of too many clusters, at least from the analyst's point of view. These mistakes increase the complexity of the cluster map, making it difficult for the analyst to comprehend. Yet type II mistakes are easier to correct.

Return for a moment to the abstract formulation in the model: a partition which assigns each patch to the same "cluster" fits the axioms. If this partition <u>is</u> a clustering (so that only one real class is present in the data), then it is perfect. But usually it won't be. A scene containing only one real class will not challenge an analyst, and so no automatic clustering technique will be needed. In fact, if it is desirable to avoid type II mistakes, then a clustering should tend to produce somewhat more than the "correct" number of clusters.

If data is clustered into N clusters, each cluster containing n_i of $n = \Sigma\, n_i$ points, then the probability that a pair of points selected at random will be in the same cluster is $\frac{1}{n^2} \sum_{i=1}^{N} n_i^2$. For example, the clustering shown in Table 1 has this probability .047. Selecting the points in image data from different spatial locations will slightly improve

the chance of finding points from different clusters. (A Monte Carlo experiment in which 129,726 pairs were selected at random from different patches resulted in 5,720 pairs in the same cluster, with an estimated probability of .044.) Therefore, most of the time a random pair will be from different clusters. If we have a partition of the data in which each pure pixel is assigned to a member of the partition, then we can test how well the partition is fitting the model by testing whether pixels from the same patch are clustered alike and whether randomly selected pairs are usually clustered differently.

Actually we can do better than this, but to do so requires another assumption about the data. Several studies[24,25] have shown that the first linear combination of the principle components map[26] preserves over half of the "structure" of the imagery. For a single image, experiments[27] have shown that the first combination is approximately the brightness. It would seem to follow that the average brightness of a patch taken over the number of acquisitions is a significant one dimensional atribute. Given a one dimensional measurement, the patches can be ordered by the value of this measurement, and samples spread out in this ordering, but not too spread out, can be selected. We get a way of testing type I mistakes.

In Table 2, the inter-patch Euclidean distances are shown. The cluster centers from Table 1 have been ordered and numbered by increasing brightness. Looking at this table, one gets the impression that about one-fourth of the way on each side of a cluster are other clusters which are different, but not too different, from the one. The average inter-cluster distance in this data set is about 25. The strategy of selecting samples spread out by

average brightness actually does work better than random selection of pixels from different patches, but not as much better as could be hoped. We will see shortly it is possible to more reliably select samples from different real classes.

Spectral Metric Considerations

So far we have not discussed what is central to any clustering technique: it is the Classifier. For multi-dimensional data, one might consider a Classifier based on a distance function preceeded or followed by spatial processing. Let us, for simplicity, ignore temporarily the spatial aspects of the Classifier, and concentrate on the underlying metric. That is, we consider a per-pixel classifier. We are given M cluster centers $C_1,\ldots,C_M$ and hope to classify a single pixel p into one of the M clusters. We might assign p to the nearest neighbor using a Minkowski metric, or we might use a Maximum Likelihood classifier. It needs to be noted that there are many objections to per-pixel classification based on a metric (e.g. Maximum Likelihood classification) although it is currently the technique most often used. We wonder if there is something about the data which tells us something about the metric.

Imagine a scene with a real class present in two fields. Suppose the spectral measurements throughout each field are constant. We would hope the Classifier assigns them to the same cluster. The underlying metric should therefore usually assign individual pixels to the same nearest neighbor cluster center. In real data, a pixel is a little more than an acre, and the measurements are averages over this much of the real world.

Imagine another scene in which the East half of the first field is interchanged with the East half of the second. As part of our "flat Earth" assumption, the East-West measurements could be different but not too different. A pixel on the boundary between East and West in either field will be measured as a convex combination of the supposedly constant East and West samples. It is obvious that this sample is in the same real class. Therefore, the metric classifier should assign these measurements to the same cluster. To not do so would be an error.

The point here is fundamental, and so, at a risk of repeating myself, I would like to describe in more earthly terms the point. Take a farmer with 80 acres under cultivation; he or she plants the "back 40" with wheat in the week of April 20 and prepares the rest for immediate planting. Early in May, the rest is planted with wheat. The conditions in this region are such that these two areas result in slightly different development. However, suppose they are not all that different: suppose another field in a spatially separated area managed like the farmer's field planted in May were placed in the same real class as the "back 40" class. It seems completely inarguable that a Landsat view of the boundary between farmer A's "back 40" and the rest of the farm would be placed in different real classes. They are both planted using techniques which, were the areas on the ground spatially separated, would have been recognized as being in the same real class. Therefore, the boundary between these two slight variations in spectral-temporal behavior is in the same real class.

There is a geometric way to describe this situation. Unless spatial considerations override the underlying metric classification, a pixel spectrally on a straight line between samples from the same cluster should be assigned to that cluster. Mean vectors of patches, of course, are less bothered by spatial problems: the nearest neighbor to patch mean classification should have this property. A set in Euclidean space is called convex is for each pair of points in the set, the line segment between the points lies in the set. Denote the underlying distance between a measurement-space vector p and a cluster center C by $d(p,C)$. We do not suppose d is a metric or even that d is defined everywhere. For each cluster j, let $K_j = \{p : d(p,C_j) \leq d(p,C_i) \text{ for } i \neq j\}$. These M sets are called the decision regions; the decision boundary between classes i and j is the set $H_{ij} = \{p : d(p,C_i) = d(p,C_j)\}$. We would hope the decision regions are convex subsets of measurement space (in order to incorporate what we noticed in the preceeding paragraph). However, the concept of an underlying metric really applies to methodology, not to the structure of the data. Instead of a methodological assumption, we refer directly to the model, and then discuss techniques which are consistent with the axiom.

Axiom 5. Let P, Q and R be patches with spectral mean vectors p,q and r, respectively. Suppose that $q = \alpha p + (1-\alpha)r$, $0 < \alpha < 1$, and that P and R are in the same real class. Then Q, P and R are all in the same real class.

This axiom is a statement in the same general direction as Axiom 4. That is, we are further limiting the term "in the same real class."

At first glance, it appears that this axiom forces an underlying distance function (if any) to be weighted Euclidean distance. Not quite; consider Fig. 5. Three cluster centers are shown. The decision boundaries are hyperplanes, which of course implies linear discriminant functions, or, equivalently, weighted Euclidean distance between any two classes. The weights in this example depend on the cluster pair which compete for the point. It is hard to see how such a distance function could arrise in an unsupervised clustering program, or even why it would be considered in such connection. Because of the example, we cannot prove (without making more assumptions) that the distance function must be essentially Euclidean, or even that the distance function need be derived from a metric.

This discussion does suggest the following: if we assume the weights in a weighted Euclidean distance have been incorporated in the data, then ordinary Euclidean distance seems to play a fundamental role in our data analysis problem. It is the only natural distance function compatible with Axiom 5 for a Classifier which incorporates such a function at some level. In particular, the commonly applied per-pixel classifiers using Maximum Likelihood or "city block" distance fail to satisfy Axiom 5. This may account for what was noticed by Bauer et al.[11] (p. C-38): they are comparing 5 classification methods, and find

> . . . Through the use of Euclidean distance measure, the CPU time required by the minimum distance classifier was much less than for any of the maximum likelihood classifiers. Analyst involvement was also minimal. . . . The overall cost was therefore lowest of all. . . . Suprisingly, the accuracy was relatively high, equalling . . . [or] being slightly higher than the maximum likelihood per point classifier . . . .

The per-pixel classification accuracy of nearest Euclidean neighbor was best of all, even exceeding the spatial classification program ECHO.

The fundamental role Euclidean distance commands suggests we return to the sticky point we abandoned above. Instead of relying on some ad hoc one dimensional attribute such as average brightness, suppose we look for a linear combination of measurement vectors which preserves, in one dimension, the Euclidean distance between points in measurement space. Of course, this is impossible, so we do as well as we can. We select distinct prototypes $p_1,\ldots,p_k$ which represent the data (in the same sense as a selection of prototypes used to define a principle components map). We form the $I = k(k-1)/2$ distinct differences $z_i = p_\mu - p_\nu$, and define

$$f(a) = f(a_1,\ldots,a_n) = \sum_{i=1}^{I} (||z_i|| - |(z_i|a)|)^2 ,$$

where $||z_i||$ is the Euclidean norm and

$$(z_i|a) = \sum_{j=1}^{n} z_{ij}\, a_j$$

is the inner product in n-space. The function f measures how well the linear map from E^n to E^1 preserves the separation between prototype pairs. Methods for finding minima of this function are discussed in a more general setting by Bryant and Guseman,[28] with applications being noticed by Bryant and Jenson.[29] Prototype selection is also discussed there.

In Table 3, inter-patch Euclidean distances are shown after ordering based on the transformation which attempts to model the original separation. That is, instead of average brightness, we order the patch centers by this most significant (in the Euclidean distance sense) feature. It is clear from Table 3 that this ordering results in a better match between ordinal separation and Euclidean separation. There are exceptions, as there must be (since the data is not at all one dimensional), but they are less drastic than before. By selecting pixels spread out in this order (but not too spread out), we hope to obtain type II error training without having to worry our weary analyst.

More on Clustering

Suppose we are given two clusterings of data. By observation of how samples are classified in the two clusterings, we can judge qualitatively how they differ without having any labels. Two quantitative methods to compare clustering are by the use of a contingency table of cross classifications,[30] and Rand's[31] idea of using pairs to estimate similarity between clusterings. Green and Rao[32] earlier used a substantially equivalent approach. Dubes and Jain[33] study several clustering algorithms using this technique. The technique requires counting the number of times the two clusterings agree on whether a pair is in the same cluster or in different clusters; the estimate of similarity is this number divided by the total number of pairs.

As is discussed in Anderberg,[34] the contingency table provides more information (as might be expected). (In fact, it is this method by which one transfers ground truth labels to a cluster map. One observes which label is associated with the most elements in a given cluster and transfers that label to the entire cluster. This works best when there are very few labels.) Unfortunately, the contingency table is an $N \times M$ matrix, and is harder to incorporate in a simple comparison than Rand's single number.

Three drawbacks to Rand's method are seen in image data: first, there are a lot of pairs, so we have to sample. But second, it is hard to adequately sample the alike-alike case. Recall we earlier found the probability of two independent choices selecting a pair from the same cluster to be

$$p_s = \frac{1}{n^2} \sum_{i=1}^{N} n_i^2 \quad .$$

We can further compute the number of distinct pairs both from the same class

$$P_s = \sum_{i=1}^{N} n_i(n_i-1)/2 = (n^2 p_s - n)/2$$

and the number of distinct pairs from different classes

$$P_d = \frac{1}{2} \sum_{i=1}^{n} n_i(n-n_i) = \frac{1}{2}(n^2 - \sum_{i=1}^{n} n_i^2) = n^2(1-p_s)/2$$

Since, in a practical remote sensing problem, p_s is usually small (say < 0.1), it will be necessary to either consider all pairs, or else sample the data many times, to obtain estimates on the alike-alike case. There are, of course, $n(n-1)/2$ distinct pairs in all.

The third drawback is possibly the most difficult. The first two are mere computational problems. The third is an interpretative problem. The contribution to Rand's measure of similarity by the alike-alike case can never be very much, and yet here is where most of the information about similarity conveyed by the contingency table is found. That is, since for both clusterings, even unrelated random ones, $P_s \simeq P_d(p_s/(1-p_s))$ we expect different-different matches to be found for most selections of pairs. If we refer to the example already discussed, for that clustering $n = 22\ 932$, so that (approximately) $P_s = 12.3 \times 10^6$ and $p_d = 250.6 \times 10^6$. It is 20 times more likely that a random pair will be from class different than alike. In fact, the discrete partition which assigns every point in the image to its own "cluster" will show a Rand measure of similarity of $P_d/(n(n-1)/2) \simeq 0.953$. Yet the original clustering was a clustering, and this partition, with 22,932 clusters, is nonsense.

Note that the difficulty we perceive here acts in opposition to an earlier one. There, we didn't know what to do about the indiscrete clustering: the partition of the data which assigned everything to the same cluster seemed to fit the axioms. Rand's measure fails at the other extreme: comparison of a clustering to the discrete partition will not be significant. Somehow, the Rand measure fails to weight the two "mistake" cases equally.

One way around this might be to normalize the measure to weight the two difference cases equally. Let A and B be clusterings, and let P_{sa} be the number of distinct pairs classed in the same cluster by A , and let P_{da} be the number of distinct pairs classed in different clusters by

A . Similarily define P_{sb} and P_{db} . Denote by N_s the number of alike-alike matches and N_d the number of different-different matches. The new similarity measure proposed is

$$S = \frac{N_s}{P_{sa}+P_{sb}} + \frac{N_d}{P_{da}+P_{db}} \quad .$$

Note that, like Rand's measure, it is difficult to compute for a large data set, and values are between 0 and 1. If the clusterings are the same (i.e. if A = B) then $S = 1$, whereas if A and B never treat a pair alike, then $S = 0$. When we compare the above example clustering to the discrete one, we obtain $N_d = P_{da} = 250.6 \times 10^6$ and $P_{db} = n(n-1)/2 = 262.9 \times 10^6$, so that (as $N_s = 0$) $S = 0.49$. Comparison to the indiscrete clustering gives $N_s = P_{sa} = 12.3 \times 10^6$, $P_{sb} = 262.9 \times 10^6$ and $N_d = 0$; thus $S = 0.045$, very close to Rand's measure.

Unfortunately, the computational drawbacks to this measure preclude its use for data sets of this size. Also, both clusterings must cluster all of the data in order that the measures be defined. As suggested before, we are able to select some pairs known to be from the same real class and some from probably different real classes. The underlying nearest neighbor assignment strategy will assign these measurement vectors to cluster centers. They (our samples) are all pure pixels, and threfore should be less affected by registration errors than pixels nearer spatial boundaries, yet registration errors will persist even for pure pixels. A nearest Euclidean neighbor assignment will be less bothered than many other measures. Consider again an example like one we had before:

	Ground Measurement	Sample
Pixel 1	(4,6)	(6,6)
Pixel 2	(6,8)	(4,8)

Because of registration errors, the first measurement is swapped. Now the mean of this class of ground measurements is (5,7). Note that this is also the mean of the samples, and that the Euclidean distance from a ground measurement to mean is the same as the distance from sample measurement to mean. It would seem that this situation is also common in real data measurements. That is, patch sample means should be closer to ground measurement means than samples are. Since it is the behavior of ground measurements that an analyst will understand best, we take as clusters actual patch means.

All pixels in a patch are in the same real class, and the patch mean better models the patch spectral-temporal behavior than any sample. Thus, once a number N of cluster centers $c_1,\ldots,c_N$ have been selected from the set of all patch means, the entire patch plus its border will be assigned to the nearest cluster center to patch mean. Pixels not bordering patches can initially be classified on a per-point basis by nearest neighbor. (Spatially motivated reclassification of these per-point classification will be discussed later.) It is this classification, the per-point classification, that we would like to adjust the clusters for.

Return for a moment to the problem of comparison of clusterings. Suppose again we have two clusterings A and B ; about A , however, we only have sample information. We have a certain set A_s of pairs known to be in the same cluster and another set A_d , containing the same

number P of pairs as A_s, known to be in different clusters. Each of these sets is much smaller than the set of all possible pairs. It is therefore computationally feasible to estimate the similarity between A and B (which depends, of course, on the samples). Let B_s be the pairs in $A_s \cup A_d$ which are clustered alike and B_d the pairs in $A_s \cup A_d$ which are clustered differently. Let $|V|$ denote the number of elements in a finite set V. The quantity

$$q = \frac{|A_s \cap B_s|}{|A_s|+|B_s|} + \frac{|A_d \cap B_d|}{|A_d|+|B_d|}$$

is an estimate on the similarity between the two clusterings, and $p = 1-q$ is an estimate on the difference. Both are between 0 and 1. Whether q is close to the actual similarity between clustering A and B depends on the sampling.

Now, finally, return to the model. We have an actual but unknown clustering A of patches into real classes. The quantity $p = 1-q$ is an estimate for half the probability that a pair is in the same real class but is clustered differently plus the probability that a pair is in different real classes but is clustered in the same cluster. We call this quantity the pair probability of misclustering (PPMC). We obtain the set A_s by picking samples from the same patch which are spread out spatially. We obtain A_d by ordering patch samples as suggested before and picking pairs separated, but not too separated, in this order. By this strategy, we obtain an estimate for how close an arbitrary clustering based on per-point classification restricted to pure pixels is to the actual one.

In a typical small image (such as our running example), perhaps 250 patch means will be candidates for being cluster centers. On the other hand, data of this complexity typically contains 25 real classes. Given the objective of finding cluster centers so as to make the PPMC smallest, there is the problem of selecting which clusterings to evaluate. Even given 25 real classes, the number of 25 cluster center subsets of 250 candidate centers is about 1.65×10^{34}. The method we use is based on a simple analysis of the source of errors: clusters which split same pairs or attract different pairs contribute errors. We now make this more precise.

From each patch containing at least 5 pixels, select 5 test pixels spread out spatially in the patch. Call each set of 5 a test set. Order the test sets by transforming the first member of each set to one dimension using the first principle component or the distance preserving technique discussed above. Each test set contributes 10 distinct pairs known to be in the same real class. Pixels spread out in the sense of this order can be considered to be from distinct real classes: for each test set, select the first test pixel in the test set 1/4 of the way in this order on either side. If this leads to an invalid index, select one 1/2 of the way. Each test set now furnishes 10 distinct pairs for testing how well a clustering can discriminate between different real classes.

We now evaluate the initial candidate cluster centers: take the set of all patch means. Classify each test pixel by nearest cluster center, and eliminate each cluster center to which nothing is assigned. Next eliminate each center to which only one test pixel is assigned, immediately re-classifying each test pixel which was assigned to an eliminated center.

Continue in this fashion, eliminating and reclassifying, until 100 cluster centers remain. These are the starting cluster centers. Estimate the square of the diameter D of the set of starting clusters, and set E = 1+D/5; E is the initial elimination protection threshold.

Each of the clusters which split a pair from the same test set is credited with an error, as, similarily, is a cluster which joins a pair from different sets. (For a time, we will use "cluster" instead of "cluster center.") Each test set furnishes ten tests of the clustering for each type of error. The total number of errors is proportional to an estimate for the PPMC, and the cluster which contributes the most errors is marked for elimination.

Whether the cluster is actually eliminated now depends on what happens to the pixels which we assigned to the cluster. Because, once eliminated, a cluster has no chance of re-entering the contest, we are careful to try to retain clusters which seem essential for some test pixel. If one of the test pixels, when reassigned, is more than E from the new cluster, do not eliminate the cluster, replace E by $E \cdot 20/19 + d$ (d is the dimensionality), locate the cluster with next most errors, and, in the same way, attempt to eliminate this cluster. Any time all reassignments pass the elimination protection test, do eliminate the cluster, reassign the test pixels to their new clusters and replace E by $E - d$. In the course of reducing the number of clusters from 100 to 5, the elimination protection logic is invoked typically 25 times. (The adaptive threshold is of interest independently of its use in the cluster formation process.)

We continue in this fashion, counting errors and eliminating erroneous clusters, until two clusters survive. Each clustering will have associated with it a certain PPMC: the estimate of how well nearest neighbor per point clustering performs on pure pixels. It is actually rather disappointing: about .20 to .30 for a typical scene. In any case, we select the clustering with the lowest PPMC as our initial set of cluster centers.

It needs to be emphasized that these cluster centers are not the final ones, and patches are not to be classified on a per point basis. The size of the smallest PPMC certainly suggests that per point classification leaves much to be desired. However, no improvement can be expected by changing the distance function: we have already noticed that Euclidean distance is the appropriate choice. To improve this clustering we return to the spatial nature of agricultural scenery.

Using Spatial Information in Clustering

The Classifier which has evolved so far has a spatial preprocessor in which patches are identified. Pixels within or bordering a patch could be classified in the same class, namely by assignment to the nearest cluster center from this patch mean. Many pixels, nonetheless, remain unclassified. Some of these pixels may be on sharp spatial boundaries between patches. Most, however, will be "contaminated" in various ways. For example, a one acre pond in a pasture, or an isolated stand of trees, will show much different spectral-temporal behavior than the surrounding region. Others will be on contaminated boundaries: a fence-line will frequently be bordered by perhaps 5 meters of uncultivated land. Narrow roads and their surrounding will similarly not look like a mixture between the spectrally homogeneous

areas they bound. Even the wider (paved) county roads (perhaps 25 meters wide field-to-field) are too narrow to recognize as a class. Although a human observer can infer the existence of these roads and locate them while viewing multi-imagery, this inference involves extensive cerebral processing which no computer program has approached. In addition, registration errors discussed before cause all these artifacts to occupy much more space in multi-temporal data, and complicate the problem of assigning these pixels to clusters.

The general problem of per-pixel classification thus seems hopeless. The temptation is to classify patches and quit, without apology. My personal experience with this approach has been interesting: it seems no one (except me) likes a clustering of the data in which up to one-third of the points are not assigned to any cluster. This is in spite of the fact that these pixels are not modelled. Even sharp boundaries, that is, boundaries between fields which are not contaminated by roads or wide fence-lines, are not modelled because of registration errors. Remember the problem: we must use multi-temporal data to resolve the real classes of interest. But also remember registration errors are inescapable in multi-temporal data. It seems clear that any methodology or theory in which registration errors are not taken into account is of questionable value in the agricultural scene understanding problem.

Still, all those unclassified pixels remain. Suppose we tentatively classify each pixel not in or bordering a patch willy-nilly by nearest cluster center. This gets all the pixels labelled, to be sure, but many of the labels will be incorrect. We confront two distinct problems: sharp boundaries with registration errors, and contaminated pixels. By "sharp"

we mean an abrupt spatial transition on the ground between one real class and another. By "contamination" we mean the contribution to spectral behavior by classes either not recognized or not spatially nearby. Let us consider the sharp boundary problem first.

Let P and Q be patches with mean vectors p and q, and suppose these are measurements from adjacent fields $\mathcal{P}$ and $\mathcal{Q}$. Let there be n measurements. A pixel $b = (b_1,\ldots,b_n)$ taken from the area of the sharp boundary between $\mathcal{P}$ and $\mathcal{Q}$ will not be a simple average between p and q. Not at all. Even if we assume the sample measurements in P and Q are constant, registration errors give b a completely foreign spectral appearance. Under this assumption, all we can say is that for some α_i, $0 \leq \alpha_i \leq 1$, we have $b_i = \alpha_i p_i + (1-\alpha_i)q_i$, $i = 1,\ldots,n$. Call such a vector b a registration-mixture of p and q. (A simple average would have all the α_i the same.) Since we intend to continue to utilize Euclidean distance, we compute

$$||b-p||^2 = \sum_{i=1}^{n} (\alpha_i p_i + (1-\alpha_i)q_i - p_i)^2$$

$$= \sum_{i=1}^{n} (1-\alpha_i)^2 r_i \text{ , where } r_i = (p_i-q_i)^2 \text{ , and}$$

$$||b-q||^2 = \sum_{i=1}^{n} \alpha_i^2 r_i .$$

We are interested in the case of equality between $||b-p||$ and $||b-q||$, for them any registration-mixture b will be no further than this to the nearer of p or q. In case of equality, we see

$$\Sigma (1-\alpha_i)^2 r_i = \Sigma \alpha_i^2 r_i \text{ , so that}$$

$$\Sigma\, r_i = 2\,\Sigma\, \alpha_i\, r_i \ , \text{ i.e.}$$

$$\Sigma\, \alpha_i\, r_i \,/\, \Sigma\, r_i = \frac{1}{2} \ .$$

The function $\phi(x) = x^2$ is convex, so that, by Jensen's inequality,[35]

$$\frac{1}{4} = (\Sigma\, \alpha_i\, r_i / \Sigma\, r_i)^2 \leq \Sigma\, \alpha_i^2\, r_i / \Sigma\, r_i \ .$$

That is, $\Sigma\, \alpha_i^2\, r_i \geq \frac{1}{4}\,\Sigma\, r_i = \frac{1}{4}\,||p-q||^2$. On the other hand, because $0 \leq \alpha_i \leq 1$, we have $\alpha_i^2 \leq \alpha_i$, so that (since $r_i \geq 0$)

$$\Sigma\, \alpha_i^2\, r_i / \Sigma\, r_i \leq \Sigma\, \alpha_i\, r_i / \Sigma\, r_i = \frac{1}{2} \ ;$$

thus

$$\Sigma\, \alpha_i^2\, r_i \leq \frac{1}{2}\,||p-q||^2 \ .$$

From these estimates we see that a spatial boundary pixel which is not contaminated must be closer to the nearer of the two classes it bounds than $\frac{1}{\sqrt{2}}$ times the distance between the two. We also see the estimate $0.7||p-q||$ is sharp in the sense that no better than $0.5||p-q||$ is possible. (In the absense of registration errors, $\frac{1}{\sqrt{2}}$ can, of course, be replaced by $\frac{1}{2}$.)

Return now to the contaminated pixel problem. We have determined that registration errors do not seriously compromise the nearest neighbor classification of mixture pixels on sharp boundaries, and have arrived at the upper bound $\frac{1}{\sqrt{2}}\,||p-q||$ for the distance of a mixture pixel to the nearest cluster p or q . If the data is as we assume it to be, a pixel further from a cluster center p than the largest possible $\frac{1}{\sqrt{2}}||p-q||$ (over all cluster centers q) must be contaminated. Let there be N clusters

with cluster centers $c_1,\ldots,c_N$ and define

$$r_i = \max\{\tfrac{1}{2}\,||c_i-c_j|| : j = 1,\ldots,N\} \quad .$$

The number r_i is the rejection threshold for pure pixels, and $\rho_i = \sqrt{2}\, r_i$ is the rejection threshold for mixed pixels. Let the patch means be denoted by p_j , $j = 1,\ldots,M$, and consider the patch classification problem. By definition the patches are pure, and so the p_j are uncontaminated. Therefore, if p_j is to be classified in c_i we must have $||p_j-c_i|| < r_i$. If this condition is always violated, then no cluster c_i fits patch mean p_j ; accordingly, the number N of clusters is incremented and new cluster $c_N = p_j$ is added. (Of course, this usually changes the rejection thresholds, which will be recomputed after each adjustment in the number of clusters.) Once all patch means have been successfully classified, the set of cluster centers is known. We return to the problem of understanding mixture pixels.

A first step in this direction is to classify each pixel not in or bordering a patch by nearest unrejected neighbor. If p is such a pixel, let $a_i = ||p-c_i||$, $i = 1,\ldots,N$, and label p with i provided $a_i < \rho_i$ and if $a_j < \rho_j$ then $a_i \leq a_j$, $j = 1,\ldots,N$. (If $a_i \geq \rho_i$ for each i , label p with 0 . We will cope with these pixels later. For the time being, it does not seem rational to assign such a pixel to any of the N classes.) The standard objection to this essentially per-pixel nearest neighbor classification is that points on spatial boundaries get misclassified. Suppose a misclassification results because the pixel is a spectral mixture. It seems unlikely that, in general, many of the neighbors of the pixel would be misclassified in the same

incorrect class. Certainly if no or one neighbor has the same label then the pixel's classification appears spatially to be an incorrect boundary-like misclassification. Note that registration errors and contamination seem to aid our cause: misclassification of a boundary will probably not always be to the same class because of these artifacts. A pixel with two or more neighbors having the same label seems much more likely to be correctly labelled. We therefore mark any classification with less than two neighbor agreement by replacing the label at that pixel by its negative.

The processing just described is order-dependent, but has the property that a 2×2 area all classified alike will be stable. Moreover, a pixel neighboring an area surviving this initial spatial check and having the same label will have its original label restored in the next step. For each pixel with a negative label, examine the classes (if any) to which the neighboring pixels are assigned, and reclassify the pixel in the nearest neighboring class provided, of course, the mixture rejection threshold test is met. Each of these spatial-spectral steps are repeated twice for each set of four rows of data. Most incorrect classifications of mixture pixels are corrected by this logic.

A final spatial processing step is to declassify each pixel with no neighbor in the same class. When registration errors and limited spatial resolution are taken into account, it seems impossible that an isolated classification of a pixel can be believed. These pixels are declassified (labelled with 0), and will be processed in the last step.

The Leader Algorithm

Pixels which remain unclassified after the above processing are a source of minor annoyance. For one thing, there aren't many of them in a scene dominated by agricultural activity. Still, it is simpler to give them labels than to explain why they were not labelled. A simple modification of the leader algorithm[36] is well suited for this purpose. For this algorithm, a tolerance is required which measures how distant samples can be and still be regarded as similar. The final value of the "elimination protection threshold" E introduced above serves as this tolerance. The algorithm proceeds as follows:

Step 1. The R pixels labelled with 0 are to be processed. Begin with pixel I = 1 , and clusters K = 1 . The first cluster center is pixel 1.

Step 2. Increment I. If I > R , stop, else set J = K .

Step 3. Find the distance d between pixel I and cluster J . If d < E , proceed to Step 4; else decrement J , and, if J > 0 repeat Step 3, else go to Step 5.

Step 4. Label pixel I with cluster J and go to Step 2.

Step 5. A new cluster has been found. Increment K , set the new cluster K equal to pixel I , and proceed to Step 2.

Summary

We now complete our discussion of the easy remote sensing problem: the problem of automatically finding natural classes in multitemporal cloud- and haze-free Landsat data taken from flat areas of Earth dominated by agricultural activity. In a theoretical sense, our study has not been

uniform: the analysis which leads to formation and classification of patches is based on solid theoretical ground. Much weaker is the spatial (logical) processing of per pixel classifications, although some theoretical work has been attempted even here. In any case, an automatic data analysis procedure has been described which was developed especially for currently available LANDSAT data and a particularly easy remote sensing problem.

Two main areas of investigation which remain are evaluation and generalization. The problem of comparison of two clustering methods is difficult.[37] What data set does one use? Certainly not all data; that would be prohibitively expensive. (The value of information should exceed its cost.) Much different from agricultural applications of remote sensing (and from each other) are applications to forestry, geography, geology, hydrology, meteorology and oceanography.[38] I doubt that the methodology developed for agricultural problems can be applied unchanged to any of these other areas. The approach taken here, however, may be generalized: these problems may be amenable to attack by methods which are philosophically alligned to the one taken here, although the methods must differ in detail. Remote sensing problems are data analysis and scene understanding problems, and not artificial intelligence problems. We should prefer the real stuff: an understanding of the meaning and relevance of the data structure in terms meaningful to a human analyst.

Acknowledgement

I appreciate the indulgence the National Aeronautics and Space Administration has shown me while these ideas were being developed.

References and Notes

1. The fourth Landsat spacecraft (Landsat-D) to be launched by NASA in early 1981 was the theme of the Plenary Session of the Fifth LARS Symposium. Blanchard and Weinstein discuss the design of the moving mirror assembly to minimize vibration and insure linearity of angle versus time. The thermal stability of the sensors and size of optical assembly are also discussed. Smith and Webb describe a separate data acquisition and processing center being planned at Goddard Space Flight Center. Zimmemerman discusses international use of the data, and Trichel and Erickson discuss agricultural applications. As Trichel and Erickson point out, Landsat-D will clearly alleviate many of the deficiencies which are discussed in this paper. They also, however, point to major research issues which were never directly addressed during LACIE.[4] The overview of the whole system was provided by:

V. V. Salomonson and A. B. Park, "An overview of the Landsat-D project with emphasis on the flight segment," in Proceedings of the Fifth Symposium on Machine Processing of Remotely Sensed Data, June 27-29, 1979, Laboratory for Applications of Remote Sensing, West Lafayette, Indiana, IEEE Catalog No. 79 CH 1430-B MPRSD, 1979, pp. 2-11.

2. The activities of the Earth Satellite Corporation in applying NASA technology to national resource development are described in:

R. Snelson, "Earth Sat," Future 8, No. 2 (February 1979), pp. 32-38.

3. The corn leaf blight watch, of course, made use of aircraft multispectral data. The epidemic was caused by the fungus Helminthosporium maydis, which causes corn in the field to look scorched as though killed by early frost. The epidemic was aided by the wide spread use of a hybrid which was particularly susceptible, and by favorable weather conditions. It seems unlikely that this particular epidemic will be repeated; yet, since ancient times, plant blights--major plant epidemics--have changed society, and it seems even more unlikely that the 1970 corn leaf blight will be the last.

Sample references are:

P. R. Day, "Corn blight," Frontiers of Plant Science 23 (No. 1), pp. 3, 8 (November, 1970).

R. D. Wennblom, Ed., "How bad is the corn blight?," Farm J., October 1970, pp. 21-23.

J. F. Fulkerson and J. M. Barnes, "A colloquy on corn blight," Agr. Sci. Res. 8(4), 1-10, Fourth Quarter, 1970.

The Corn Blight Watch Experiment was the first major agricultural application of remote sensing. Blighted corn is spectacularly different from normal vegetation, and the disease spread so rapidly that the week-to-week progress was of great interest. Although there were aspects of the blight which were more difficult to detect using aircraft data, and although the data itself was frequently poor and hard to manage, the simple identification problem could be done by a child. This is a completely different situation from the problem faced in LACIE[4] (and from the basic problem addressed in this paper), in which the separation between agricultural classes is often poor but the data is well documented and very well managed. See also:

"Corn Blight Watch Experiment Final Report, Experimental Results," NASA-JSC, Vol. III, 1974

4. Probably the most concise description of LACIE is:

R. B. MacDonald, F. G. Hall and R. B. Erb, "The use of Landsat data in a large area crop inventory experiment (LACIE)," in Proceedings, Symposium on Machine Processing of Remotely Sensed Data, June 3-5, 1975, The Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, IEEE Catalog No. 75 CH 1009-0-C, 1975.

The proceedings of the recent LACIE Symposium held at JSC-Houston (October 1978) contain a much more comprehensive summary of the goals and accomplishments of LACIE. The proceedings are in press.

5. C. M. Hay, J. B. Odenweller, R. W. Thomas and E. J. Sheffner, Advanced AI Aids and Keys Development, annual Progress Report for NASA Contract NAS9-14565, 15 November 1977 to 14 November 1978, University of California, Berkeley, Remote Sensing Research Program, Space Sciences Laboratory Series 19, Issue 66, November 1978.

6. Procedure 1 is described in:

W. W. Austin, "Detailed description of the wheat acreage estimation procedure used in the large area crop inventory experiment," Technical Memorandum 644-845, LEC-11497, Lockheed Electronics Company, Inc., Houston, Texas, 1978.

R. Heydorn, "Classification and mensuration approach," Proceedings of LACIE Symposium, NASA-JSC, Houston, Texas, October 1978 (in press).

Procedure 1 spawned a proliferation of "Procedures": Procedure B and M from ERIM, Ann Arbor, Michigan, and Procedure 2 from IBM Federal Systems Division, Houston, Texas, for example. Procedure 1 has as its aim the two class "bias correction" of AI labelling. Guseman and Heydorn (unpublished) extend the methodology to multiple classes, the problem Procedure M and Procedure 2 also address. I wonder, however, whether it might not be better to try to understand the source of analyst "bias," and furnish better products and aids to reduce this analyst imperfection.

7. T. N. Cornsweet, in Visual Perception, Academic Press, New York, 1970. See Chapter X, pp. 249-253.

8. T. G. Stockham, Jr., "The role of psychophysics in the mathematics of image science," in Symposium on Image Science Mathematics, Monterey, California, November 10-12, 1976, C. O. Wilde and E. Barret, eds., Western Periodicals Company, North Hollywood, California, 1977, pp. 57-66.

9. J. Bryant, "On the clustering of multidimensional pictoral data," Pattern Recognition 11, 115-126 (1979).

10. The problem that the modes are not really related to the means in any simple way may also be involved here. A complex mixture problem cannot be expected to have class means matching mixture population modes, not even in one dimensional data. The problems of false modes, perhaps caused by the digital range distortion, were communicated by W. Coberly in a 1977 quarterly review, NASA-JSC.

11. The G. E. Image 100 System has a parallelepiped classification mode which is virtually instantaneous in operation: thus it will be tried first. Experience with it on Landsat 1 data shows parallelepiped classification is an effective way to pick up scan line striping but not effective for classification. Parallelepiped classification can also be viewed as a weighted ℓ_∞ distance. For more on classification accuracy studied with changes in distance function or logical processing, see the following, where further references will be found:

M. E. Bauer, L. F. Silva, R. M. Hoffer and M. F. Baumgardner, "Agricultural scene understanding," LARS Contract Rep. 112677, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 1977.

W. Richardson and A. P. Pentland, "Evaluation of algorithms for estimating wheat acreage from multispectral scanner data," ERIM 109600-69-F, Environmental Research Inst. of Michigan, Ann Arbor, Michigan, 1976.

P. L. Odell and J. Pasa, "Effect of distance-measures on cluster-based classification procedures," Proceedings of the Milwaukee Symposium on Automatic Control, 1974.

12. R. J. Kauth and G. S. Thomas, "The tasselled cap--a graphic description of the spectral-temporal development of agricultural crops as seen by LANDSAT," LARS Symp.: machine processing of remotely sensed data, Purdue University, West Lafayette, Indiana, 1976. IEEE Catalogue No. 1976 CH 1103-1 MPRSD.

13. S. G. Wheeler, P. N. Misra and Q. A. Holmes, "Linear dimensionality of LANDSAT data," LARS Symp.: machine processing of remotely sensed data, Purdue University, West Lafayette, Indiana, 1976. IEEE Catalogue No. 1976 CH 1103-1 MPRSD.

14. In this paper, Hlavka et al. use a new approach to crop classification based on reconstructing the temporal pattern of the phenological growth state the crop displays. It is unclear whether the "noiser" measurements they prefer work better than brightness-greenness because of something in the data or because of the dynamic programming technique employed to capture the signatures from the samples. See:

C. A. Hlavka, S. M. Carlyle, R. Yokoyoma and R. M. Haralick, "Multitemporal classification of winter wheat using a growth state model," LARS Symp.: machine processing of remotely sensed data, Purdue University, West Lafayette, Indiana, 1979, pp. 105-116. IEEE Catalogue No. 79 CH 1430-8 MPRSD.

15. P. F. Lambeck, "Implementation of the XSTAR haze correction algorithm and associated steps for Landsat data," ERIM Memo No. IS-PFL-1272, rev. ERIM Memo No. IS-PFL-1916, 1977.

16. R. M. Haralick, "Glossary and index to remotely sensed image pattern recognition concepts," Pattern Recognition 5, 391-403 (1973).

17. G. H. Ball and D. J. Hall, "ISODATA, a novel method of data analysis," AD699616, Stanford Res. Inst., Menlo Park, California, 1965.

18. R. M. Haralick and I. Dinstein, "A spatial clustering procedure for multi-image data," IEEE Trans. on Circuits and Systems, CAS-22, 440-450 (1975).

19. L. G. Roberts, "Machine perception of three-dimensional solids," in Optical and Electro-Optical Processing of Information, MIT Press, Cambridge, Mass., 1965, pp. 159-197.

20. Scores of good papers on boundary detection could be cited at this point, but, since this is a digression at best, I will refrain. Still, the boundary finding problem (and I mean thin boundaries) is irresistible. It is so easy to "see" a boundary in a scene which a computer can't find . . . we wonder why. One reason which is emerging is that extensive cerebral processing finds linear features, as well as areas of high curvature, and that this highly parallel processing lies below our "consciousness." We should not forget the lesson of Lowell's Mars: we see the canals, but they are not, as we know now, really on Mars. Although this deepens the digression, I would suggest the book (the entire book, although Chatper 20 is involved in this discussion):

I. S. Shklovskii and C. Sagen, Intelligent Life in the Universe. Holden-Day, San Francisco, 1966.

Also, general probelms of perception are addressed philosophically in:

N. R. Hanson, Patterns of Discovery. Cambridge University Press, Cambridge, 1958.

There is another issue, of course, a very practical one: Even if the computer can find the boundaries, can we afford the cost? Certainly the currently popular adaptive techniques will be costly when applied to multi-image data.

21. H. C. Andrews, Introduction to Mathematical Techniques in Pattern Recognition, Wiley-Interscience, New York, 1972.

22. I. Gitman, "A parameter-free clustering model," Pattern Recognition 4, 307-315 (1972).

23. G. Leboucher and G. E. Lowitz, "What a histogram can really tell the classifier," Pattern Recognition 10, 351-358 (1978).

24. G. E. Lowtiz, "Stability and dimensionality of Karhunen-Loêve multispectral image expansions," Pattern Recognition 10, 359-364 (1978).

25. S. K. Jenson and F. A. Waltz, "Principal components analysis and canonical analysis in remote sensing, paper presented at ASP-ACSM meeting March 18-23, 1979, Washington, D.C.

26. C. R. Rao, "The use and interpretation of principle components analysis in applied research," Sankhyā Ser. A 26, 329-358 (1964).

27. T. Kaneko, "Color composite pictures from principal axis components of multispectral scanner data," IBM J. Res. Devlop. 22, 386-392 (1978).

28. J. Bryant and L. F. Guseman, Jr., "Distance preserving linear feature selection," Pattern Recognition, 1979 (to appear).

29. J. Bryant and S. Jenson, "Dimensionality reduction and data compression," in preparation.

30. H. Borko, D. A. Blankenship and R. C. Burket, On-line Information Retrieval Using Associative Indexing, RADC-TR-68-100, AD 670195, Systems Development Corp., Santa Monica, California, 1968, pp. 62-80.

31. W. M. Rand, "Objective criteria for the evaluation of clustering methods," J. Amer. Statist. Assoc. 66, 846-850 (1971).

32. P. E. Green and V. R. Rao, "A note on proximity measures and cluster analysis," J. Marketing Res. 6, 359-364 (1969).

33. R. Dubes and A. K. Jain, "Clustering techniques: the user's dilemma, Pattern Recognition 8, 247-260 (1976).

34. M. R. Anderberg, Cluster Analysis for Applications, Academic Press, New York, 1973.

35. G. H. Hardy, J. E. Littlewood and G. Pólya, Inequalities, Cambridge, 1934; 2nd ed. 1952.

36. J. A. Hartigan, Clustering Algorithms, Wiley series in probability and mathematical statistics, John Wiley and Sons, New York, 1975.

37. Even the simpler problem of comparing the three specific clustering programs ISOCLS, CLASSY and AMOEBA seems to be very difficult. ISOCLS is a modification of Ball and Hall's famous ISODATA. CLASSY and AMOEBA are based on a model for the data: the underlying assumption for

CLASSY is essentially that the pixels are independent samples from a mixture of multivariate normal populations. The model of AMOEBA is the one discussed in this paper. It is easy to see that these three programs are not really comparable, except perhaps on a subjective basis. Of course, one can compare how they perform on any given data set (for example, the clustering can be compared to ground truth); but how does one select the data sets on which to evaluate the programs? Three reports bear on this question:

P. L. Odell, "A plan for comparing clustering programs," Contract Report 17, Texas A&M University, NASA Contract #NAS-9-14689-8S, 1978.

R. K. Lennington and H. Malek, "The CLASSY algorithm--description, evaluation and comparison with the iterative self-organizing clustering system (ISOCLS)," Technical Memorandum 642-2600, Lockheed Electronics Company, Inc., Houston, Texas, 1978.

C. B. Chittineni, "Research plan for developing and evaluating classifiers," Technical Memorandum JSC-14849, Lockheed Electronics, Inc., Houston, Texas, 1979.

It may be that the problem of evaluation is really a combination of two not unrelated problems which have bothered humanity since ancient times: one is the problem of sampling. How do we cope with the obviously impossible task of testing all possible programs on all possible data sets? A far reaching essay bearing on the sampling question is:

O. Kempthorne, "Some aspects of statistics, sampling and randomization," in Contributions to survey sampling and applied statistics, H. A. David, ed., Academic Press, Inc., New York, 1978.

The other problem is equally bothersome. How do we know when we are right? The answer may be that a rigorous interpretation of this simple question is impossible. This is discussed at length in:

G. Pólya, Mathematics and Plausible Reasoning, Vol. II: Patterns of Plausible Inference. Princeton University Press, Princeton, N.J., 1954. See also the introduction to Vol. I: Induction and Analogy in Mathematics.

38. A workshop on scanner systems and applications of remote sensing was held December 11-15, 1972, under Goddard Space Flight Center sponsorship. Landsat-D incorporates many of the technological advances foreseen at that time. Also, the data requirements for various applications are discussed in depth. See:

Advanced Scanners and Imaging Systems for Earth Observations, National Aeronautics and Space Administration, NASA SP-335, prepared by Goddard Space Flight Center, available from Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 20402, xvii + 604 pp., 1973.

The Cramer-Rao Lower Bound As A Criteria For A Large Data Reduction System Such as LACIE

Patrick L. Odell
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, Texas 75080

Report #21

Prepared For

Earth Observations Division
NASA/Johnson Space Center
Houston, Texas
Contract NAS-9-14689-9S

September, 1979

D3

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
The Cramer-Rao Lower Bound As A Criteria For Evaluating A Large Data Reduction System Such As LACIE	Sept., 1979
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Patrick L. Odell University of Texas at Dallas	21
	10. Work Unit No.

9. Performing Organization Name and Address	11. Contract or Grant No.
Department of Mathematics Texas A&M University College Station, Texas 77843	NAS-9-14689-9S
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address Earth Observations Division NASA/Johnson Space Center Houston, Texas 77058	Unscheduled Technical
	14. Sponsoring Agency Code

15. Supplementary Notes
Principal Investigator: L. F. Guseman, Jr.

16. Abstract

When using a remote sensing system which produces large quantities of multi-dimensional data containing spectral, temporal, and spatial information, it is difficult for a decision maker to know whether or not

(1) the sensors are performing adequately; that is, there is sufficient information in the data to adequately perform the task for which the sensors were designed;

(2) the data analysis systems (including software, models, computation) are optimal with respect to well-defined criteria.

This report presents a theoretical approach which might serve as a basis for an implementable decision strategy for answering the above questions.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Cramer-Rao lower bound, likelihood functions, LACIE system evaluation	

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

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

THE CRAMER-RAO LOWER BOUND AS A CRITERIA FOR EVALUATING A LARGE COMPLICATED DATA REDUCTION SYSTEM SUCH AS LACIE

Patrick L. Odell
University of Texas at Dallas

1. INTRODUCTION

The LACIE (Large Area Crop Inventory Experiment) data reduction system when viewed in its totality is large and complicated. When one considers that ultimately, when perfected, this system or one of its successors will process large numbers of vector-valued observations to produce estimates of total annual production of several crop classes. Each 1.1 acre of land will generate a vector of data. The aims include world surveys of various crops; hence the amount of data to be processed is large.

In developing the LACIE system specific assumptions are made which in some cases cannot be tested for validity. This in turn affects the effectiveness of the system and the accuracy of the final estimates. These situations and various ad hoc decisions made when developing the software give reasons to question the optimality of the total system; hence responsible administrators may never really know how well the development task has been performed.

If it can be determined that the data lacks sufficient information for performing the task effectively even when the software system is the best (conceptually), then one can judge that the data taking sensors or some part of that system is inadequate. The method for analysis formulated

in this paper gives an objective method for making the decision. The method is based on the Cramer-Rao Lower Bound, a well-known result in the theory of statistics [1].

The analysis in this paper is restricted to spectral data or spectral and temporal data. Spatial information cannot be evaluated using this technique in LACIE data. It may be true and more evidence is mounting that the major source of information for estimating proportions lies in the spatial information in the scene and the spectral information is valuable only in that it allows a spatial analysis (see Bryant [2]).

2. THE PROBLEM

The wheat acreage estimation problem is essentially the problem of estimating α_i in the mixture probability density function

$$p(x) = \sum_{i=1}^{m} \alpha_i p_i(x) \tag{1}$$

where α_i denotes the a priori probability that a pixel contains crop class Π_i $i = 1,2,\ldots,m$. Note that

$$\sum_{i=1}^{m} \alpha_i = 1 \tag{2}$$

and

$$\alpha_i > 0 \tag{3}$$

for each $i = 1,2,\ldots,m$. Also, in the LACIE problem, the value of m (the number of crop classes in the region) is not known hence also must be estimated in practice; [3], [4].

Suppose that N_i denotes the number of pixels in a region which contains crop Π_i and

$$\sum_{i=1}^{m} N_i = N \tag{4}$$

then

$$\alpha_i = N_i/N \ .$$

Also, if A denotes the area of a single pixel (approximately 1.1 acre) then the total wheat acreage assigned to the i^{th} crop is given by

$$A_i = AN_i = NA\alpha_i \tag{5}$$

which can be estimated by

$$\hat{A}_i = NA\hat{\alpha}_i \tag{6}$$

where $\hat{\alpha}_i$ is an estimator for α_i . Hopefully, the estimates are unbiased; that is,

$$E[\hat{A}_i] = NAE[\hat{\alpha}_i] = NA\alpha_i = A_i \tag{7}$$

and the variance

$$V[\hat{A}_i] = (NA)^2V[\hat{\alpha}_i] = (NA)^2(E[\hat{\alpha}_i])(1 - E[\hat{\alpha}_i])[1 - \tfrac{n}{N}]/n \tag{8}$$

is sufficiently small. Unfortunately, in practice the estimator for α_i is not unbiased and the mean square error

$$MSE(\hat{A}_i) = E[\hat{A}_i - A_i]^2 \tag{9}$$

can be large. Note that if one processes every pixel in the region that $V[\hat{A}_i] = \Phi$, the null matrix and

$$E[\hat{A}_i - A_i]^2 = B^2(\hat{A}_i) \ , \tag{10}$$

where

$$B^2(A_i) = \{E([\hat{A}_i] - A_i\}^2 \tag{11}$$

is the bias of the estimator A_i . However, if $E[\hat{A}_i] = A_i$ that is $\hat{A}_i$ is unbiased, then $B^2(\hat{A}_i)$ is zero and the $V(A_i) = MSE(\hat{A}_i)$.

The rationale for the method of evaluation is the following:

(a) Compute the C-R Lower Bound for the unbiased estimator of α_i when the p.d.f. of the observations is defined by (1),

(b) Estimate the C-R Lower Bound using LACIE data,

(c) Compare the variance of various estimators now being used in LACIE and estimate their efficiency with respect to the C-R bound,

(d) Determine whether or not any estimator whose variance attains the C-R Lower Bound is sufficient for the task of effectively estimating the total acreage of a specified crop class.

Note if one is given an estimator which attains the C-R Lower Bound (a measure of the best one can do given the data from (1)), and that estimator is not "good enough", then this is equivalent to saying that the "data is not good enough" or "the hardware is not good enough". Conversely, if an estimator that takes on the C-R Lower Bound is "good enough" which means that the data and the hardware are "good enough"; yet the estimator being used fails to attain that bound, then the "software is not good enough".

A decision maker who must evaluate the total system can use the logic described above to determine and differentiate the levels of effectiveness of the "hardware system" and the "software system".

3. THE CRAMER-RAO LOWER BOUND

Let $x_1, x_2, \ldots x_n$ denote a random sample from a multivariate (p-variate) population whose p.d.f. is defined by (1). The likelihood function is

$$L(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} p(x_i) \tag{12}$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_j p_j(x_i) \quad ; \quad \sum_{j=1}^{m} \alpha_j = 1 \ . \tag{13}$$

If we denote the parameter vector $\alpha = (\alpha_1, \ldots, \alpha_{m-1})$ then by the C-R theory the $m-1 \times 1$ random vector

$$S = \partial \ln L / \partial \alpha = \frac{\ln L}{\partial \alpha} \tag{14}$$

is such that

$$E[S] = \Phi$$

and

$$E[SS^T] = -E\left[\frac{\partial^2 \ln L}{\partial \alpha \partial \alpha}\right] = -E\left\{\frac{\partial^2 \ln L}{\partial \alpha_i \partial \alpha_j}\right\} = \Lambda^{-1} \tag{15}$$

where Λ^{-1} is the Cramer-Rao Lower Covariance matrix bound for unbiased estimators of the vector α. That is, if $\hat{\alpha}$ is any unbiased estimator for α, then the covariance matrix $V(\hat{\alpha})$ will never be less than Λ^{-1}.

Note that if A and B are two positive definite matrices of the same size and $A - B$ is positive semi-definite, then we say B is less than A.

From (1) and (2) it follows that

$$p(x) = \sum_{j=1}^{m-1} \alpha_j p_j(x) + [1 - \sum_{j=1}^{m-1} \alpha_j] p_m(x) \tag{16}$$

$$= \sum_{j=1}^{m-1} \alpha_j [p_j(x) - p_m(x)] + p_m(x) \ . \tag{17}$$

then from (13),

$$\ln L = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{m-1} \alpha_j \ p_j(x_i) - p_m(x_i) \ + p_m(x_i) \right] \ . \tag{18}$$

It follows from (14), (18) and (1) that

$$S_j = \sum_{i=1}^{n} \frac{[p_j(x_i) - p_m(x_1)]}{[\sum_{j=1}^{m} \alpha_j p_j(x_i)]} \tag{19}$$

$$= \sum_{i=1}^{n} \frac{[p_j(x_i) - p_m(x_i)]}{p(x_i)} \tag{20}$$

and

$$\frac{\partial S_j}{\partial \alpha_k} = - \sum_{i=1}^{n} \frac{[p_j(x_i) - p_m(x_i)][p_k(x_i) - p_m(x_i)]}{[p(x_i)]^2} \ . \tag{21}$$

Hence, from (15) the C-R Lower Bound covariance matrix is given by the $(m-1)x(m-1)$ matrix

$$\Lambda = \{\Lambda_{jk}\} = \left\{ -E\left[\frac{\partial S_j}{\partial \alpha_k} \right] \right\}^{-1} \tag{22}$$

when $m = 2$.

$$\Lambda_{11} = \left[E \sum_{i=1}^{n} \left[\frac{p_1(x_i)-p_m(x_i)}{p(x_i)} \right]^2 \right]^{-1} = \left[\frac{n}{\alpha_1\alpha_2} \left\{ 1 - \int_{-\infty}^{\infty} \frac{p_i(x)p_2(x)}{p(x)} \, dx \right\} \right]^{-1}$$

such that

$$0 \leq \left[\int_{-\infty}^{\infty} \frac{p_1(x)p_2(x)}{p(x)} \, dx \right] \leq 1 \quad ,$$

a result first obtained by Hill [5]. Note also that when $m = 2$, the "inside-out rule" yields

$$\Lambda_{11} \geq \left[E \sum_{i=1}^{n} \left[\frac{p_i(x_i)-p_m(x_i)}{p(x_i)} \right]^2 \right]^{-1}$$

4. ON ESTIMATING THE LOWER BOUND

Note that it is necessary that the probability density functions

$$p_1(x), p_2(x), \ldots, p_m(x) \tag{23}$$

be known in order for (22) to be computed. Also, (22) is a function of the sample size, n, and the values of $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{m-1})$. Since in practice one knowns neither α or the p.d.f.'s in (23), one must estimate these quantities. Let

$$\hat{p}_1(x), \hat{p}_2(x), \ldots, \hat{p}_m(x) \tag{24}$$

denote valid estimates of the p.d.f.'s in (23) and let

$$\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{m-1} \tag{25}$$

C-2

denote estimates of the α_i's , $i = 1,2,\ldots,m-1$. Then by substituting these estimators in (22) and taking the appropriate expectation, one can approximate the lower bound.

In order to perform the approximation, one must have training data to estimate $p_j(x)$ $j = 1,2,\ldots,m$; a random sample of unlabelled data to estimate α_i $i = 1,2,\ldots,m-1$; and finally a method to perform the expectation integration in (22).

The expectation integral will in most cases have to be evaluated using Monte Carlo techniques [6]. Note also that the procedure will need to be repeated over a specified range of values of n , the sample size. Since the vector α is unknown, it is desirable to vary the values about $\hat{\alpha}$ to study the sensitivity of the lower bound to small changes in α . Clearly, to evaluate the lower bound is a relatively large undertaking when the probability density functions and the parameters must be estimated.

5. A C-R LOWER BOUND FOR ESTIMATORS USED IN LACIE

In an attempt to avoid knowing the p.d.f.'s defined in (23) and to model the problem closer to the actual situation in LACIE, the following formulation appears valid.

In the LACIE application one classifies the unlabelled samples

$$x_1, x_2, \ldots, x_n \tag{26}$$

into crop classes $\Pi_1,\Pi_2,\ldots,\Pi_m$ and then counts the number of pixels classified into each class. One can use this information to estimate α_j . A naive but biased estimate of α_j for $j = 1,2,\ldots,n$ is

$$\hat{g}_j = n_j/n$$

where n_j is the number of pixels classified as Π_j and

$$\sum_{j=1}^{m} n_j = n \ . \tag{27}$$

This motivates the formulation of a C-R lower bound based on a multinomial probability density function which may be derived by introducing a vector of random variables

$$Y_1,Y_2,\ldots,Y_m \tag{28}$$

where

$$Y_j(x_i) = \begin{cases} 1 & \text{if } x_i \text{ is assigned to } \Pi_j \ , \\ 0 & \text{if } x_i \text{ is not assigned to } \Pi_j \ . \end{cases} \tag{29}$$

Note that

$$\Pr\left[Y_j(x_i) = 1\right] = \sum_{k=1}^{m} \alpha_j P(j|k) \quad j = 1,\ldots,m \tag{30}$$

where

$$P(i/j) = \int_{R_i} p_j(x)dx$$

the probability of classifying an observation X from the j^{th} class Π_j into the i^{th} class, Π_i and $R = (R_1, R_2, \ldots, R_m)$ are the classification decision regions defined by some classification decision rule (not necessarily a Bayes rule).

Let

$$g_i = \sum_{j=1}^{m} \alpha_j P(i|j) = E[\hat{g}_j] \ . \tag{32}$$

Then, the likelihood function for a random sample defined in (26) can be written

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m} g_j^{Y_j(x_i)} = \prod_{j=1}^{m} g_j^{n_j} \tag{33}$$

where

$$n_j = \sum_{i=1}^{n} Y_j(x_i) \ . \tag{34}$$

It follows that

$$\ln L = \sum_{L=1}^{m} n_i \ln g_i$$

$$= \sum_{i=1}^{m-1} n_i \ln g_i + \left[n - \sum_{i=1}^{m-1} n_i \right] \ln \left[1 - \sum_{i=1}^{m-1} g_i \right]$$

since

$$\sum_{i=1}^{m} g_i = 1 \ .$$

Also, from (31) and $\sum_{i=1}^{m} \alpha_i = 1$ that

$$g_i = \sum_{j=1}^{m-1} \alpha_i \left[P(i|j) - P(i|m) \right] + P(i|m) \tag{35}$$

and

$$\frac{\partial g_i}{\partial \alpha_j} = P(i|j) - P(i|m) \ . \tag{36}$$

From (14) then $S_j = \partial \ln L / \partial \alpha_j$ it follows that

$$S_j = \sum_{L=1}^{m} n_i \frac{1}{g_i} \left[P(i|j) - P(i|m) \right]$$

or in matrix notation

$$S = [\Delta_{ij}]^T \, G^{-1} \, \bar{n} \tag{37}$$

where the $(m-1)xm$ matrix $[\Delta_{ij}]^T$ is defined by its elements

$$\Delta_{ij} = P(i|j) - P(i|m) \ , \tag{38}$$

$$G = \begin{bmatrix} g_1 & 0 \ldots 0 \\ 0 & g_2 \ddots 0 \\ \vdots & \ddots \\ 0 & 0 \ldots \end{bmatrix} \ , \tag{39}$$

and

$$\bar{n} = (n_1, n_2, \ldots, n_m)^T \ . \tag{40}$$

Note that by the C-R theory the expected value of S is the zero vector which we will verify directly.

$$\begin{aligned} E[S] &= E \ [\Delta_{ij}]^T \, G^{-1} \, \bar{n} \\ &= [\Delta_{ij}]^T \, G^{-1} \, E[\bar{n}] \\ &= [\Delta_{ij}]^T \, G^{-1} \, (ng) \ . \end{aligned} \tag{41}$$

where

$$g = (g_1, g_2, \ldots, g_m)^T$$

or

$$g = GJ \tag{42}$$

where

$$J = (1, 1, \ldots, 1)^T \quad .$$

It follows from

$$\sum_{i=1}^{m} P(i|j) = 1 \tag{43}$$

for $j = 1,2,\ldots,m$ that

$$[\Delta_{ij}]J = \Phi \tag{44}$$

and in turn from (41) and (42) that

$$E[S] = n[\Delta_{ij}]\, G^{-1}\, GJ = \Phi \tag{45}$$

The covariance matrix $V(S)$ of S can now be computed using (37) and (45), that is

$$V(S) = [\Delta_{ij}]^T\, G^{-1}\, V(\bar{n})\, G^{-1}\, [\Delta_{ij}] \tag{46}$$

where $V(\bar{n})$ is the covariance matrix of the $\bar{n} = (n_1, n_2, \ldots n_m)$, a multinomial vector variate; that is,

$$V(n) = n[G - GJJ^TG]$$
$$= n\, G(I - JJ^TG] \qquad (47)$$
$$= n[G - P\alpha\alpha^TP]$$

where

$$G = \begin{bmatrix} P_1\alpha & 0 & \cdots & 0 \\ 0 & P_2\alpha & & 0 \\ \vdots & & \ddots & \\ 0 & & & P_{m-1}\alpha \end{bmatrix}$$

From (44), (46), and (47)

$$V(S) = n[\Delta_{ij}]^T\, G^{-1}[\Delta_{ij}] \ , \qquad (48)$$

the inverse of the desired C-R lower bound

$$\Lambda = [V(S)]^{-1} \ .$$

For exemplary purposes consider the case when $m = 2$, then since

$$[\Delta_{ij}]^T = [P(1|1)-P(1|2),\ P(2|1)-P(2|2)] \ ,$$

$$G = \begin{bmatrix} g_1 & 0 \\ 0 & g_2 \end{bmatrix} \ ,$$

$$g_1 = 1-g_2 \ ,$$

$$P(1|1) = 1 - P(2|1), \text{ and}$$

$$P(2|2) = 1 - P(1|2) \ ,$$

then

$$\Lambda_{11} = \frac{1}{n} \frac{g_1 g_2}{[P(1|1)-P(1|2)]^2} \qquad (49)$$

Suppose further, that if there are no errors in classification, that is

$$P(1|1) = P(2|2) = 1$$

then

$$g_1 = \alpha_1 \quad \text{and} \quad g_2 = \alpha_2$$

and

$$\Lambda = \frac{1}{n} g_1 g_2 = \frac{1}{n} g_1(1-g_1) \quad .$$

The variance of a sufficient statistic $\hat{\alpha}_1 = \frac{n_1}{n}$ for the α_1 in a binomial probability density function.

Odell and Chhikara [5] defined an estimator for α_1 when $m = 2$ as

$$\hat{\alpha}_1 = [n_1/n - P(1|2)]/(P(1|2) - P(1|1))$$

whose

$$E[\hat{\alpha}_1] = \alpha_1$$

and whose variance was (48). Hence $\hat{\alpha}_1$ is a minimal variance unbiased estimator for $\hat{\alpha}_1$.

Fortunately in the LACIE application we have estimates for the matrix of probabilities of misclassification

$$P = [P_{ij} = P(i|j)] \tag{50}$$

hence we can study the lower bound defined by (48) as a function of the matrix P and the matrix G .

A question that the author has failed to resolve is the relation between the C-R bound computed in (22) and the one computed using (48). The conjecture is that Λ defined by (22) is less than or equal to that defined by (48), and that equality may indeed be the only case.

6. LACIE ACREAGE ESTIMATORS

In the LACIE activity three general estimators for $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{m-1})$ have been posed and are used. These are

$$\hat{\alpha} = \overline{n}/n = \hat{g} \tag{51}$$

$$\tilde{\alpha} = P^{+}\hat{g} \tag{52}$$

$$= [\hat{P}]^{+}\hat{g} \qquad \text{Odell and Chhikara [7]} \tag{52b}$$

$$= [P]^{+}g \tag{52c}$$

$$\overline{\alpha} = Q\hat{g} \tag{53a}$$

$$= \hat{Q}g \qquad \text{Wheeler et al. [8]} \tag{53b}$$

$$= \hat{Q}\hat{g} \tag{53c}$$

where

$$Q = \Pr(\Pi_i | C_j) \tag{54}$$

and

$$\Pr(\Pi_i | C_j) \tag{55}$$

is the probability an observation X classified as being from the Π_j really belong to the class Π_i. Note also, that in this notation that (31) can be rewritten as

$$P(i|j) = \Pr(C_i | \Pi_j) \ . \tag{56}$$

Note that the matrices P and Q are related since

$$Pr(C_j|\Pi_i) = Pr(\Pi_i|C_j)P(C_j)/Pr(\Pi_i) \tag{57}$$

where

$$\alpha_i = Pr(\Pi_i) \tag{58}$$

and

$$Pr(C_j) = g_j \tag{59}$$

Let

$$A = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \alpha_m \end{bmatrix}$$

then

$$A = A\, P^T\, G^{-1} \tag{61}$$

and

$$P = G\, Q^T\, A^{-1} \tag{62}$$

The advantage for using the model defined by (61) over that used by (62) is that one need not invert the matrix P to obtain an unbiased estimate for α as implied in (52b). Note that since

$E[\hat{g}] = g$,

$E[\hat{Q}] = Q$,

$E[\hat{P}] = P$, and

$E(\hat{P}^{-1}] = \hat{P}^{-1} + B(\hat{P}^{-1})$

it follows that (53a), (53b), and (53c) are unbiased estimators for α if $\hat{Q}$ and $\hat{g}$ are independent. In general, (52b) and (52c) are biased estimators for α. However, the estimator (53b) is especially interesting in that one can process the whole region, then g is known exactly (no sampling error) and by estimating Q by $\hat{Q}$ an unbiased estimator can be achieved. Since $P^{+}P\alpha$ may not equal α (52a) may also be biased; however, if $P^{+} = P^{-1}$ then (52a) is unbiased.

7. ON ESTIMATING P AND Q

Let X_i, $i = 1,2,\ldots,n_j$ $j = 1,2,\ldots,m$ denote m independent random samples from m crop classes, then

$$\hat{P}(i|j) = \hat{P}(C_i|\Pi_j) = n(i|j)/n_j \tag{63}$$

where

$$n_j = \sum_{i=1}^{m} n(i|j)$$

and $n(i|j)$ are the number of pixels from Π_j and classified as Π_i. Then

$$\hat{P} = [\hat{P}(i|j)] \quad . \tag{64}$$

Note that there is no information in this sample concerning α_j, $j = 1,2,\ldots,m$.

Let X_{ij} $i = 1,2,\ldots,n_j'$ $j = 1,2,\ldots,m$ denote a independent random sample from m a cluster of all (or a random sample of) classified pixels. Then using ground truth one can determine

$$n'(\Pi_i|C_j) = n'(i|j) \ ,$$

the number of pixels classified as C_j which in reality came from the crop class Π_i . Then

$$Q_{ij} = n'(i|j)/n'_j \quad . \tag{65}$$

Note that there is no information about g_i in this data.

However, if X_{ij} $i = 1,2,\ldots,n_j$ $j = 1,2,\ldots,m$ where $n = \sum_{J=1}^{m} n_j$ is the total sample size and X_{ij} is a random sample from the mixture defined in (1), then $n_j, j = 1,2,\ldots,m$ are random variables and

$$P(i,j) = P(C_i|\Pi_j)P(\Pi_j) = P(i|j)\alpha_j$$

and

$$P(i,j) = P(\Pi_j|C_i)P(C_i) = Q(j,i)g_i$$

Then

$$\hat{P}(i|j) = n_{ij}/n_{\cdot j} \tag{66}$$

where

$$n_{\cdot j} = \sum_{i=1}^{m} n_{ij}$$

and

$$\hat{Q}(j,i) = n_{ij}/n_{i\cdot} \tag{67}$$

where

$$n_{i\cdot} = \sum_{J=1}^{m} n_{ij} \quad .$$

It is for this sample that the relations (61) and (62) hold. An estimator for P when test data is available is given by (64) and the estimator for Q results when one classifies the scene then takes a random sample from the classified pixels to ground truth. It is important to note that one must already have ground truth to "train the classifier" in order to produce the clusters of classified pixels. Then one takes a random sample of these pixels from ground truth to obtain the estimate for Q. Estimates of α_J, $J = 1,2,\ldots,m$ are immediate then by (53b) since

$$g_i = N_i/N$$

where N_i is the total number of pixels classified in Π_i and

$$N = \sum_{i=1}^{m} N_i$$

is the total number of pixels in the region.

8. THE VARIANCES OF THE ESTIMATES FOR α

The estimators listed in (52) and (53) are associated with moments as follows:

$$E[P^+\hat{g}] = [P^+P\alpha]$$

$$E[\hat{P}^+]g = E[\hat{P}^+]P\alpha$$

$$E[\hat{P}^+]\hat{g} = E[\hat{P}^+]E[\hat{g}] = E[\hat{P}^+]P\alpha$$

$$E[Q\hat{g}] = Qg = \alpha$$

$$E[\hat{Q}g] = E[Q]g = Qg = \alpha$$

$$E[\hat{Q}\hat{g}] = E[\hat{Q}]E[\hat{g}] = Qg = \alpha$$

and since from (47)

$$V(\hat{g}) = \frac{1}{n}\,[G-GJJ^{T}G]$$

we can compute

$$V[P^{+}\hat{g}] = P^{+}V[\hat{g}]P^{+^{T}}$$

$$= \frac{1}{n}\,P^{+}[G-GJJ^{T}G]P^{+^{T}}$$

$V[\hat{P}^{+}g]$ = (no simple form has been derived)

$V[\hat{P}^{+}g]$ = (no simple form has been derived)

$$V(Q\hat{g}) = QV(\hat{g})Q^{T} = \frac{1}{n}Q[G-GJJ^{T}G]Q^{T}$$

$V(\hat{Q}g)$ = (no simple form has been derived)

$V(\hat{Q}\hat{g})$ = (no simple form has been derived).

One should note that since

$$g = P\alpha = GJ$$

that

$$V(\hat{g}) = \frac{1}{n}[G-GJJ^{T}G]$$

$$= \frac{1}{n}\left\{\begin{bmatrix} P_1\alpha & 0 & \cdots & 0 \\ 0 & P_2\alpha & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & P_m\alpha \end{bmatrix} - P\alpha\alpha^{T}P\right\}$$

a function of probabilities of misclassification and the values of α. Also one should note that mean square error is a better measure of variability since several of the estimators are biased.

The task is to compute the mean square error for these estimators and compare them with the C-R Lower Bound. This will give a measure of the efficiency of the software subsystem of LACIE.

8. CONCLUDING REMARKS

In order to obtain results from which evaluation and planning discussions can be made, it appears that a large amount of data processing must take place in order to evaluate the two forms of the C-R lower bounds, that is (22) and (48). Fortunately, one can use (48) without large amounts of data processing by using matrices of probabilities of misclassification already observed in earlier studies, but to evaluate (22) will require data analysis plus some Monte Carlo work to evaluate expectations.

In order to evaluate the moments of the estimators of α one will have to process some data but do a considerable amount of Monte Carlo work.

9. REFERENCES

[1] Kendall, M. G, and Stuart, A. The Advanced Theory of Statistics. Chas. Griffin and Co., Ltd. London, 1961, pp. 9-44.

[2] Bryant, J., "On the Clustering of Multidimensional Pictoral Data," Pattern Recognition, Vol. 11, No. 2, pp. 115-125.

[3] Goldberg, M. and Shreen, S., "A Four-Dimensional Histogram Approach to Clustering of Landsat Data," Canadian Journal of Remote Sensing, Vol. 2, April 1976, pp. 1-11.

[4] Basu, Rekha, Effect of Model Specification in Pattern Recognition, unpublished doctoral dissertation, University of Houston, 1979.

[5] Hill, Bruce M., "Information for Estimating the Porportions in Mixtures of Exponential and Normal Distribution," Journal of American Statistical Assoc., Vol. 58, No. 305, pp. 918-932.

[6] Newman, T. G. and Odell, P. L., The Generation of Random Variates, Griffin's Statistical Monographs and Courses, No. 28, Hafner Publishing Col, N.Y., 1971.

[7] Chhikara, R. S. and Odell, P. L. "Estimation of a Large Area Crop Inventory Using Remote Sensing Technology," Statistical Theory and Methodology for Remote Sensing Data Analysis, UT-Dallas Annual Report, 1974, pp. 1-48.

[8] Wheeler, S. G.; Heydorn, R. P.; Misra, P. M.; Lee, W., Jr.; and Smart, R. T., "An evaluation of Procedure 1," Proceedings of the Technical Session, Vol. 2, p. 825-842, LACIE Symposium, Oct., 1978, NASA, Johnson Space Center, Houston, Texas. JSC 16015.


FEB 1 1980
RECEIVED
UNIVERSITY OF CALIFORNIA
BERKELEY
ACCESSIONS