Reprinted from

Eighth International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Crop Inventory and Monitoring

July 7-9, 1982

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

CAN CROP TYPES BE RESOLVED USING MIXTURE DISTRIBUTION COMPONENTS -- SOME INITIAL RESULTS AND IMPLICATIONS

R.K. LENNINGTON, C.T. SORENSEN
Lockheed Engineering and Management Services Company, Inc.
Houston, Texas

R.P. HEYDORN
National Aeronautics and Space Administration/Johnson Space Center
Houston, Texas

I. ABSTRACT

A fundamentally important problem in the analysis of remotely sensed data has been the characterization of the distribution of spectral measurements for cover types of interest. Such a characterization is an implicit or explicit part of the training of any classifier based on multispectral measurements. It also forms the basis for most unsupervised classification or clustering of such data. Proportions of cover types of interest are equal to proportions of distributions characterizing them.

For these reasons it appears natural to formulate the classification or proportion estimation problems in terms of a mixture of underlying distributions. This mixture would describe the whole image and would be composed of a sum of simpler distributions, each with some specificed proportion. The usual assumption has been that these underlying simpler distributions are multivariate normal. The questions in such a formulation are whether crop categories of interest may be well represented by a small number of underlying multivariate normal distributions and whether the underlying distributions themselves may be resolved from the overall mixture distribution.

The latter question has been answered in the affirmative for Landsat data using the CLASSY clustering algorithm developed by Lockheed at the Johnson Space Center. The former question has been the subject of much recent research. We have determined that when raw Landsat spectral measurements or even greenness and brightness transformed data are used, an important crop distribution, that of small grains, cannot always be represented by a subset of the CLASSY generated normal mixture components. This paper describes recent work devoted to examining this problem, using features derived from curves fitted to the temporal profile of greenness. Using such features and restricting the pixels examined to those that are reasonably pure, we show that small grains distributions may be well represented by a set of mixture component distributions.

Some implications of these results for future work in crop area estimation and classification are explored.

I. INTRODUCTION

The analysis of remotely sensed data frequently requires the design of a classifier for the purpose of either locating a ground cover class of interest or estimating the proportion of this ground cover class or possibly both. In any case, the design of such a classifier involves characterizing what the ground cover class of interest "looks like" in the feature space being used. It also involves a similar analysis for ground cover classes which resemble or are similar to the ground cover of interest. This training process may be viewed as an attempt to understand the distribution in measurement space of the class of interest and related confusion classes. The training may take the traditional form of selecting and labeling a representative set of samples from these classes. The difficulty with this approach, in the context of remotely sensed data, is that the class identities of samples are rarely known precisely if at all. Thus, any labeling which is done must include an unspecified number of mistakes or errors.

On the other hand, it may be reasonable to assume at the outset that the distribution of the samples from each class are of a specified parametric form, say $f_i(x)$. The distribution of all the data then becomes a mixture of these component distributions where the proportion

of each class, λ_i, becomes the weight for that class in the mixture. The form of the mixture density is then

$$f(x) = \sum_{i=1}^{M} \lambda_i f_i(x)$$

The advantages of this particular approach are two fold. First, the prior knowledge incorporated in the parametric model allows the form of the individual component densities to be determined without knowing the identities of any individual samples. In order to construct a classifier it is therefore only necessary to label the component distributions as being the class of interest or not. Typically this decision may be made based on the mean of each component distribution. Ambiguity in labeling individual samples in the tail areas where classes overlap is eliminated. A second advantage of the mixture distribution formulation is that the proportions of each component density may be estimated directly as a part of fitting the model. Classification is, thus, needed only if the location of the class of interest is desired.

Of course, for the mixture model to be effective two conditions must hold. The distribution of the class of interest must be similar to a known parametric distribution (or a mixture of a small number of such distributions). Also, the separability of the class of interest from other classes must not be two small. Lack of separability leads to difficulty in resolving the individual component distributions from the overall mixture. In order to assure these two conditions, features must be carefully selected. Section II discusses the manner in which features were selected for our particular experiment. Section III describes the results of estimating the proportion of small grains in ten Landsat data segments using the mixture model. Conclusions and implications are given in Section IV.

II. FEATURE SELECTION

In previous studies[1] we have used the mixture model to estimate the proportion of various classes using raw Landsat data and using Kauth-Thomas[2] transformed brightness and greenness data for several different acquisitions. The results of these studies were that direct proportion estimates exhibited more bias and variance than other types of estimators.

Recently a new set of features derived from multitemporal Landsat data has been described[3]. These features are determined from fitting a profile model to the greenness data for each picture element (pixel) collected over a number of acquisition dates. The profile model which we fitted has the form

$$g(t) = \begin{cases} g_0 & \text{for } t \leq t_0 \\ At^{\alpha} e^{-\beta t^2} & \text{for } t > t_0 \end{cases}$$

The model was fitted using a linear least squares procedure on the log of the data. Fitting was done separately to the left and fight of an assumed emergence date, t_0. The final value of t_0 was selected to minimize the total sum of squares. Figure 1 shows the form of this model

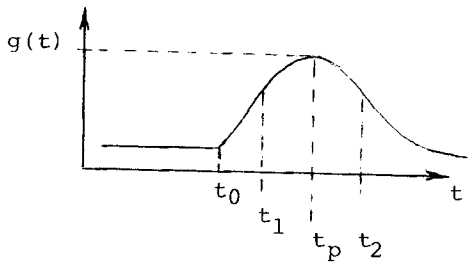

Figure 1

The profile model has a number of advantages. Not all of these will be discussed in this paper. However, it is clear that it serves to estimate the life history of a vegetative ground cover, as measured by greenness, and that this estimation allows the identification of important features in the life cycle of the vegetative cover which may not be observed on an individual Landsat acquisition.

A number of different features derived from fitted profiles were examined as a part of this study. These features were rated on their ability to separate small grains from other crop classes and on the normality of their distributional form. Normality was an important condsideration since a normal mixture model was to be used in proportion estimation.

Three features which were judged to be best in terms of separability and normality were selected. These features were the time of peak greenness, t_p, the value of greenness at the peak, $g(t_p)$, and the distance between the left and right inflection points, t_2-t_1. (see Figure 1)

III. MIXTURE MODEL RESULTS

A mixture model in which each component was assumed to be multivariate normal was fitted to the profile derived features for each of 10 Landsat segments. The fitting was done using the CLASSY program developed at the Johnson Space Center by Lockheed[4]. This algorithm uses an adaptive maximum likelihood procedure to estimate the number of components, the proportion of each component distribution, and the parameters describing each component distribution. The Landsat segments used are five by six nautical mile areas in the northern Great Plains. Table 1 gives the location of these segments and the Landsat acquisitions used. Only pixels which were determined to contain data from a single ground cover type were used. Mixed pixels were eliminated using existing ground truth data.

After the mixture distribution was fitted to the pure pixel data for each segment, component distributions were labeled as small grains or other using a maximum likelihood distribution labeling technique[1]. This technique infers a small grains or non-small grains label for each distribution using a set of 200 ground truth labeled pure pixels selected at random from the segment. Table 2 gives the number of small grains and non-small grains components for each segment.

To obtain a proportion estimate for small grains the estimated proportions for each small grains component distribution in the mixture model were added. Table 3 gives these mixture model proportion estimates and summary statistics.

IV. CONCLUSIONS AND IMPLICATIONS

The mixture model proportion estimates have a very low variance and coefficient of variation. The slight negative bias is the subject of further investigation. It is likely that this bias is caused by the inability of the normal mixture model to fit the asymmetric tails found in the feature distributions for some segments.

In general, the ability of the CLASSY algorithm to extract components from the mixture and the resulting fit of these components to histograms derived using the ground truth is exceptional. Figure 2(a) shows a typical small grains histogram for the feature t_p for segment 1899. This histogram was developed using ground truth labeled pixels. The sum of the small grains component distributions estimated from the unlabeled mixture is shown in Figure 2(b).

This work implies that the mixture model is a viable method for determining the distributions of classes of interest in remote sensing problems and in estimating the proportions of these classes directly. Further work needs to be done in applying the method to entire scenes where the effect of mixed or boundary pixels must be considered.

V. REFERENCES

1. Lennington, R. K., and Terrell, G. R. (1980) Evaluating the Use of Analyst Labels in Maximum Likelihood Cluster Proportion Estimation. AgRISTARS Technical Memorandum JSC-16538.

2. Kauth, R. J., and Thomas, G.S.: (1976) The Tasselled Cap -- A Graphic Description of the Spectral Temporal Development of Agricultural Crops As Seen by Landsat. Proceedings of the Symposium on the Machine Processing of Remotely Sensed Data, Purdue University.

3. Badhwar, G. D.: (1980) Crop Emergence Date Determination from Spectral Data. Photogrammetric Engineering and Remote Sensing, vol. 46, no. 3 pp. 369-377.

4. Lennington, R. K., and Rassbach, M. E.: (1978) CLASSY -- An Adaptive Maximum Likelihood Clustering Algorithm. Proceedings of the LACIE Symposium, pp. 671-689.

Segment Number	Crop Year	State	Acquisitions (Julian Dates)						
1544	78	Mont.	104	122	140	158	176	221	230
1394	78	N.D.	120	174	211	220	238		
1650	78	N.D.	156	191	209	218	236		
1920	78	N.D.	101	136	199	209	217	236	
1636	78	N.D.	135	154	190	207	226		
1663	77	N.D.	121	138	156	174	211		
1676	79	S.D.	120	165	184	211	237		
1566	78	Minn.	115	133	169	196	232		
1899	77	N.D.	122	140	157	175	193		
1825	78	Minn.	133	196	206	223	224		

Table 1. Landsat Segments and Acquisition Dates.

Segment Number	Number of Small Grains Distributions	Number of Non-Small Grains Distribution
1544	2	9
1394	2	6
1650	1	11
1920	4	12
1636	1	11
1663	5	7
1767	0	15
1566	2	8
1899	4	11
1825	2	6

Table 2. Number of Small Grains and Non-Small Grains Component Distributions.

Segment Number	Ground Truth Proportion (%)	Direct Proportion Estimate (%)
1544	26.81	26.40
1394	41.48	39.57
1650	13.73	10.70
1920	15.99	13.88
1636	50.16	50.42
1663	53.98	53.42
1676	7.06	0.0
1566	37.32	28.32
1899	67.51	59.03
1825	34.40	29.43

Avg. G.T. Prop. = 34.84

Bias = -3.75
Variance = 3.26

Relative Bias = -0.11
Coefficient of Variation = 0.09

Table 3. Proportion Estimates of Small Grains Obtained from the Mixture Model

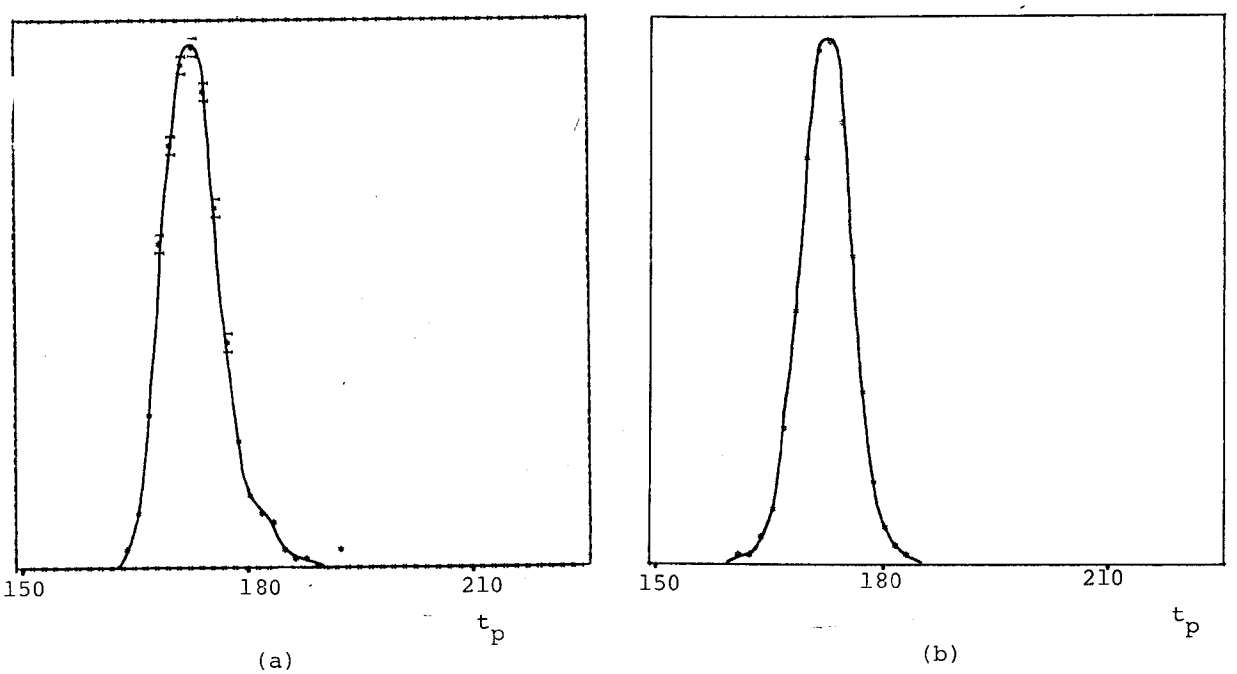

Figure 2. (a) Ground truth distribution for pure small grains pixels from segment 1899.

(b) CLASSY estimated distribution for small grains using all pure pixels from segment 1899.

R. Kent Lennington obtained the B.S.E.E. and M.S.E.E. at the University of Texas at Austin. His masters research was in the area of the analysis of bioelectric data. Following this, he obtained an M.A. degree in Statistics in 1971 and a Ph.D. in Electrical Engineering in 1975, also at the University of Texas. His dissertation research was devoted to the cluster analysis of biological populadata. After spending two years teaching Electrical Engineering at the University of Missouri at Kansas City, he joined Lockheed Engineering and Management Services Company in 1977. His work there has centered on applying pattern recognition techniques to the analysis of multispectral satellite data for the purpose of estimating the area planted to various crops. He is currently the supervisor of the Pattern Recognition Section at Lockheed.

Charles Sorensen received a B.S. in Mathematics and Chemistry in 1975 from Tarleton State University, and a M.S. in Statistics from Texas A&M University in 1976. Further graduate work in Statistics and Computing Science at Texas A&M University was ended in 1981. Mr. Soresnsen has been with Lockheed Engineering and Management Services Company since 1981. His current work is in the area of pattern recognition involving usage clustering and classification techniques for the estimation of cropland proportions using Landsat Imagery.

Richard P. Heydorn was born in Akron, Ohio, September 4, 1935. He received his B.S. degree in Electrial Engineering and his M.A. in Mathematics from the University of Akron. In 1971, he received his Ph.D. in Statistics from the Ohio State University. Dr. Heydorn has been actively working in Pattern Recognition and related research since 1959. He joined NASA/Johnson Space Center in 1974 where he worked on the Large Area Crop Inventory Experiment (LACIE). Later he became the Project Scientist of the LACIE transition project. Beginning in 1978, he has lead a series of research projects in Pattern Recognition under a program for Agricultural and Resources Inventory Surveys Through Aerospace Remote Sensing (AgRISTARS). Currently he is the NASA Science Manager of a fundamental research program in Mathematical Pattern Recognition and Image Analysis. Dr. Heydorn is a Registered Professional Engineer in the State of Ohio and a member of the American Statistical Association and Sigma Xi.