AgRISTARS

SR-T1-04112
NAS-9-14689

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

Supporting Research

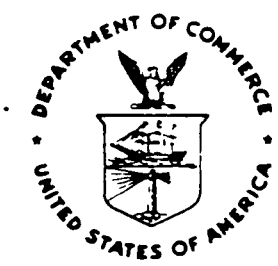
December 1980

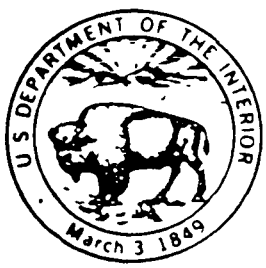
FINAL REPORT

DEVELOPMENT OF ADVANCED ACREAGE ESTIMATION METHODS

By: L. F. Guseman, Jr.

NASA





Department of Mathematics
Texas A&M University
College Station, Texas

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

4 Title and Subtitle	5 Report Date
FINAL REPORT: Development of Advanced Acreage Estimation Methods	December 1980
	6 Performing Organization Code
7 Author(s)	8. Performing Organization Report No
L. F. Guseman, Jr.	
9 Performing Organization Name and Address	10 Work Unit No
Department of Mathematics Texas A&M University College Station, Texas 77843	11. Contract or Grant No NAS-9-14689
	13 Type of Report and Period Covered FINAL (11/1/79-10/31/80)
Sponsoring Agency Name and Address Earth Observations Division NASA/Johnson Space Center Houston, Texas	14 Sponsoring Agency Code

5 Supplementary Notes

6 Abstract

Work carried out under the above contract was concerned with:

Development of Multi-Image Color Images

Spectral-Spatial Classification Algorithm Development

Spatial Correlation Studies

Evaluation of Data Reduction Systems

17 Key Words (Suggested by Author(s))	18 Distribution Statement

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

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

FINAL REPORT

DEVELOPMENT OF ADVANCED

ACREAGE ESTIMATION METHODS

Contract NAS-9-14689

November 1, 1979 - October 31, 1980

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 work reported herein was 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, 1979 to October 31, 1980. The investigations were carried out by personnel at Texas A&M University, University of Houston, University of Tulsa, and Dr. P. L. Odell, U. T. Dallas (Consultant).

L. F. Guseman, Jr.
Principal Invesigator

DEVELOPMENT OF ADVANCED ACREAGE ESTIMATION METHODS

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 a starting point for other algorithms which produce acreage estimates. 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.

Work carried out under this contract dealt with algorithm development, theoretical investigations, and empirical studies. The algorithm development tasks centered around the use of the AMOEBA clustering/classification

algorithm as a basis for both a color display generation technique and maximum likelihood proportion estimation procedure. Theoretical results were obtained which form a basis for the maximum lieklihood estimation procedures. An approach to analyzing large data reduction systems was formulated. An exploratory empirical study of spatial correlation in LANDSAT data was also carried out. Specifically, investigations were carried out in the following areas:

Development of Multi-Image Color Images

Spectral-Spatial Classification Algorithm Development

Spatial Correlation Studies

Evaluation of Data Reduction Systems

Each of these investigations is discussed in turn in the sequel.

2. DEVELOPMENT OF MULTI-IMAGE COLOR IMAGES

In a crop inventory 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 multi-dimensional data (multispectral measurements) we also have "multi-pictures" of the subject. The availability of this spatial aspect of the data and attempts to preserve the spatial integrity were the basis for investigations carried out in previous contract periods (see [1] and the references therein). These investigations led to the development of the AMOEBA spatial clustering/classification algorithm ([2]) and a distance preserving algorithm for dimensionality reduction ([3]).

The above mentioned algorithms were combined with a model for human color vision to formulate a technique for generating a single color image from a multi-image. The formulation and results of the technique are presented in the attached report:

Jack Bryant and Gary Breaux, Multi-Image Display for Human Understanding, Contract NAS-9-14689, SR-T1-04080, Report #22, Department of Mathematics, Texas A&M University, August, 1980.

3. SPECTRAL-SPATIAL CLASSIFICATION ALGORITHM DEVELOPMENT

The objective of this study was to formulate and test algorithms based on a likelihood function which respected the integrity of some predetermined structure in the data.

For purposes of these investigations, the "pure field data" (patches) determined by the AMOEBA algorithm ([2]) were used as the predetermined structure. A maximum likelihood parameter estimation procedure (HISSE) was designed to respect (take into account) field integrity.

A mathematical description and implementation of the procedure, along with results from preliminary tests appears in the attached report:

Charles Peters and Frank Kampe, Numerical trials of HISSE, Contract NAS-9-14689, SR-HO-00477, Department of Mathematics, University of Houston, August, 1980.

Theoretical results underlying the approach used in the HISSE algorithm are discussed in the attached report:

Charles Peters, On the existence, uniqueness, and asymptotic normality of a consistent solution of the likelihood equations for nonidentically distributed observations--applications to missing data problems, Contract NAS-9-14689, SR-HO-00492, Department of Mathematics, University of Houston, September, 1980.

Additional theoretical results were obtained which address the convergence of a particular iterative form of the likelihood equations in the case of a mixture of densities from (possibly distinct) exponential

families. These results appear in the attached report:

Richard A. Redner, An iterative procedure for obtaining maximum likelihood estimates in a mixture model, Contract NAS-9-14689, SR-T1-04081, Division of Mathematical Sciences, University of Tulsa, September, 1980.

4. SPATIAL CORRELATION STUDIES

The objective of this study was to gain some insight into the nature of the spatial correlation of pixels in Landsat data. In particular, an empirical study of neighboring pixels (along scan lines) was carried out in an attempt to understand the characteristics of spatial correlation for boundary or mixed pixels. Results of this study appear in the attached report:

W. A. Coberly, Spatial correlation in LANDSAT: An empirical study, Contract NAS-9-14689, SR-T1-04082, Division of Mathematical Sciences, University of Tulsa, November, 1980.

5. EVALUATION OF DATA REDUCTION SYSTEMS

Data reduction systems which utilize multi-temporal 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 proportion estimates cannot (in many instances) be carried out. In addition, when the software system is (conceptually) the best it may be that properties 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:

Virgil R. Marco, Jr. and P. L. Odell, Information in remotely sensed data for estimating proportions in mixture densities, Contract NAS-9-14689, SR-T1-04083, Program in Mathematical Sciences, University of Texas at Dallas, November, 1980.

REFERENCES

1. L. F. Guseman, Jr., Development and evaluation of clustering procedures, Contract NAS-9-14689-9S, SR-T9-00402, Department of Mathematics, Texas A&M University, November, 1979.

2. Jack Bryant, On the clustering of multidimensional pictorial data, Pattern Recognition 11, 115-125 (1979).

3. Jack Bryant and L. F. Guseman, Jr., Distance preserving linear feature selection, Pattern Recognition 11, 347-352 (1979).

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

4 Title and Subtitle	5 Report Date
Multi-Image Display for Human Understanding	August, 1980
	6 Performing Organization Code

7 Author(s)	8 Performing Organization Report No
Jack Bryant Gary Breaux	22
	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-
	13 Type of Report and Period Covered
12 Sponsoring Agency Name and Address	Unscheduled Technical
Earth Observations Division Johnson Space Center Houston, Texas 77058	14 Sponsoring Agency Code

15 Supplementary Notes

Principal Investigator: L. F. Guseman, Jr.

16 Abstract

Three recently discovered techniques are combined to produce subjectively appealing color displays of multi-temporal Landsat imagery. The first technique selects prototypes by use of an unsupervised clustering program. These are used to find a linear dimensionality reduction such that the inter-prototype separation in the original space is nearly preserved in three dimensions. The third technique produces red, green, and blue values for an image in which normal human interpretation of color differences closely matches the Euclidean distances within the three dimensional pre-image.

17 Key Words (Suggested by Author(s))	18 Distribution Statement
Clustering, Linear feature selection, Landsat, Color display, Human vision, Multi-imagery	

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

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

MULTI-IMAGE DISPLAY FOR HUMAN UNDERSTANDING

By

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

and

Gary Breaux
Department of Mathematics
Texas A&M University
College Station, Texas

Report #22

Prepared For

Earth Observations Division
NASA/Johnson Space Center
Houston, Texas
Contract NAS-9-14680-11S

August 1980

MULTI-IMAGE DISPLAY FOR HUMAN UNDERSTANDING

Jack Bryant* and Gary Breaux*

Abstract. Three recently discovered techniques are combined to produce subjectively appealing color displays of multi-temporal Landsat imagery. The first technique selects prototypes by use of an unsupervised clustering program. These are used to find a linear dimensionality reduction such that the inter-prototype separation in the original space is nearly preserved in three dimensions. The third technique produces red, green, and blue values for an image in which normal human interpretation of color differences closely matches the Euclidean distances within the three dimensional pre-image.

Clustering Linear feature selection Landsat
Color display Human vision Multi-imagery

*The authors were partly supported by the National Aeronautics and Space Administration, Contract NAS-9-14689, principal investigator, L. F. Guseman, Jr.

Consider the imagery shown in Fig. 1. Each scene of about 23,000 picture elements (pixels) is a Landsat remotely-sensed image taken from the North American Great Plains. The images have been corrected geometrically to be in close spatial registration to one another. Each was acquired on a different date: in May, June, August, and September, 1976. The August acquisition is shown in Plate 1A, the standard false-color product produced at Johnson Space Center, Houston, Texas. The two Landsat infra-red bands have no color; the standard product is somewhat like color infra-red film. The images of Fig. 1 are small, but the digital data set is not, for each pixel is a 16-vector (4 components for each acquisition).

The high dimensionality of the space in which these data are embedded is a common problem in pattern recognition. Most data analysis techniques such as clustering or classification require computer time at least in proportion to the dimension, and some (e.g. maximum likelihood classification) require time porportional to the square. Thus a common motive for dimensionality reduction is computational complexity. Another is human understanding: the presentation of the multi-image in the form of Fig. 1 (as four images) is not ideal. Yet there seems to exist no better way to present high dimensional imagery for human analysis. This is exactly the problem we tackle: is there a way to display the imagery of Fig. 1 while retaining the spatial and spectral-temporal structure?

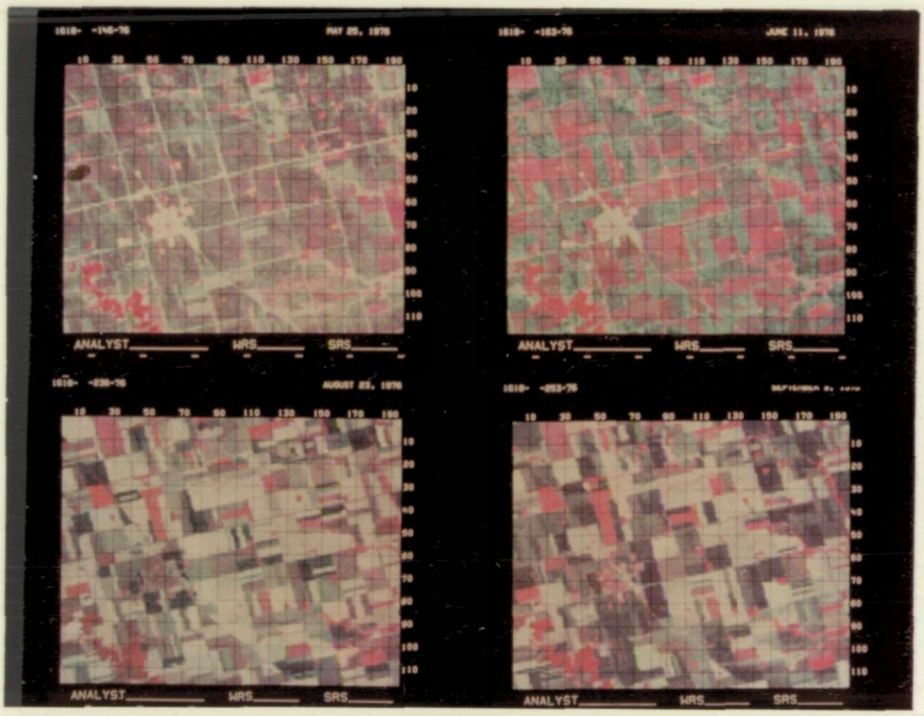


Fig. 1 Four Pass Landsat Imagery

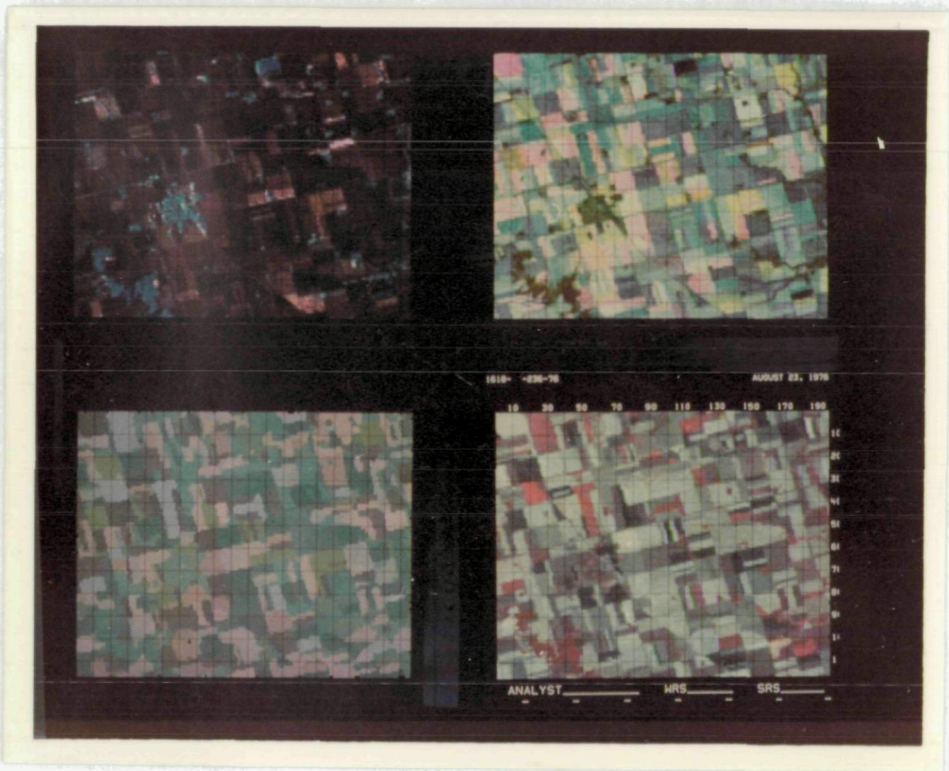


Plate 1. Color Products:
A. JSC Product 1
B. AMOEBA Clustering of Fig. 1
C. Principle Components Display
D. Distance Preserving Display

WHAT IS STRUCTURE?

By spatial structure we mean the spatial relationship between objects in the scene. To preserve spatial structure we produce a single image which is pixel-by-pixel registered to the multi-imagery. It is not so clear what spectral-temporal structure means. It will surely mean different things to different people. Our view is that the structure is represented by the Euclidean distances (in the high dimensional space) between typical measurement-space samples. Structure is preserved when these distances are accurately reproduced in the lower dimensional space. A new technique[1] for linear feature selection has as its objective the preservation of distances between samples (prototypes). Rather than obtain the prototypes at random, we use the spatial clustering program AMOEBA.[2] Plate 1B shows the clustering of the data in Fig. 1 we obtain. Note that this cluster map is not an image in the usual sense of a picture of a scene. Some of the spatial structure is clearly lost, particularly the pattern of roads so easily seen in Fig. 1.

Because of the spectral overlap between the measurements in any one acquisition (and present in the scene), the intrinsic dimensionality of a given acquisition is less than the number of measurements.[3] Thus we know some of the spectral structure, and use a four-to-two brightness-greenness transformation.[4] This converts the 16-dimensional data of Fig. 1 to 8-dimensional data. This is the data we cluster to produce Plate 1B.

WHAT IS COLOR PERCEPTION?

A method for reducing dimensionality (and a measure of success) is only helpful is we can display the reduced data so it can be understood. As an example, suppose the data could be represented in one dimension. Then it is natural to produce a gray-scale or black-and-white image. Since we know that normal human gray (i.e. non-color) vision has a logarithmic response, we prepare an image so that the perceived brightness (not the actual brightness) is linearly proportional to the transformed data (with, perhaps, a bias to translate the transformed data). That is, we consider the physiology of human vision in preparing our image.

Unfortunately, the multi-imagery of Fig. 1 is not one dimensional spectrally: nor is any single acquisition. As we shall see, however, the data can be reduced to three dimensions with small errors. Color images can be produced with three colors, which suggests color vision is at most three dimensional. The easy way to get a color dispaly (reduce dimensionality to three, display one red, one green, and one blue) is not appropriate for the same reason that we would have been wrong to produce a black-and-white image with the flux viewed linearly proportional to the transformed data. Namely, this display fails to take into account the physiology of human color vision. Indeed, imagery produced in this way is disappointing (Hay et al.[5]). Instead, we should produce a color image in which human perception of color difference matches distances between the objects being displayed. To this end, we need to model visual perception. We begin with a red-green-blue digital image and follow the processing of this image by the visual system. We use the notation of Faugheras.[6]

A model for the combined video or photographic system and pigmented cone photochemical response gives a linear transformation U to produce cone output signals L, M, and S. A model for retinal receptor response produces the (nonlinear) transformation by the logarithm function to L*, M*, and S*. Next a model for the Ganglion neural connections gives a final linear transformation P to signals A, C_1, and C_2. Signal A is brightness and C_1 and C_2 are chromaticity signals: these go to the visual cortex. (We are ignoring spatial effects.) Faugheras notices that each of these transformations is invertible and uses this to transmit color imagery over a noisy channel with lower bit rate (or better perceived signal-to-noise ratio). He reports[6, p. 91] a reduction in the average bit-rate by a factor of 27.

A comprehensive survey of color image perception and a bibliographical guide is found in Hall[7, Chapter 2]. Hall gives a block diagram (p. 42) of the monocular visual system (but gives no numeric parameters). Faugheras's work is based on a slightly simpler model (for light-adapted (or photoptic) vision). To use his work, one need only determine U. He has determined P by psychovisual experiments. There is another approach to this problem, outlined by Hall[7, pp. 21-22] and followed by Juday[8] and Kaneko[9]. We prefer the approach based on a model, although we do not know the exact U for the film product used. This problem is being studied, but our requirements are not severe: we do not need strict color fidelity. The major problems left are: first, how much of subjective color space can we occupy without exceeding the film color gamut? Second, how do we scale the output image so that it can be displayed on

a given digital system? We found experimentally that twenty-five levels of brightness A and thirteen levels of each chromaticity channel C_1 and C_2 could be displayed. The details of how to scale everything are less interesting and are relegated to the Appendix.

Let's now review the end-to-end process. We obtain our connection between measurement space and perception space by the following steps:

1. Using feature selection techniques,[10,11] reduce the dimensionality to three. We use here the principle components map and the distance preserving map.[1]
2. Apply suitable scaling (see the Appendix) and apply P^{-1}, exponential, and U^{-1} to the transformed image.
3. Again scale, and display the result on a color monitor or as color film. These products make up Plate 1C (the principle components map) and Plate 1D (the distance preserving map).

DISCUSSION

Observers, viewing Plate 1, uniformly prefer the color image 1D. The cluster map 1B is rejected because it is not a picture in the same sense that 1A, 1C, and 1D are pictures, although the clustering shown might be a helpful aid to a human analyst. Plate 1C is not favored because obviously distinct classes are colored the same. This is certainly not the case in 1D. We observe that 1C is "too dark," yet it was produced by the same method as led to 1D; only the feature selection method was different. This finding which discredits the principle components approach is new but not entirely unexpected. See, for example, the

imagery shown in Lowitz.[12, Fig. 1, p. 360] The seventh (of seven) principle components image contains significant structural information. Here we find that the principle components map from 8 to 3 dimensions identifies distinct classes, a flaw which goes against our underlying purpose. If B is 3 x 8 matrix and $y_1,\ldots,y_p$ are the prototypes, let $P = p(p-1)/2$, let

$$f(B) = \sum_{1\le i<j\le p} (||By_i - By_j|| - ||y_i - y_j||)^2,$$

and let

$$N(B) = \left(\frac{1}{P} f(b)\right)^{1/2} .$$

For the principle components map B, $N(B) \cong 9.78$, and for the distance preserving map $N(B) \cong 0.95$. The two are shown in Table 1.

The main open problem is to make the colors reproducible. The experiment reported here used 32 prototypes. In another, using the same data and procedure, we let AMOEBA find the natural number of clusters rather than the forced number 32. It found 12, and their centers were used as prototypes. The resulting image was as satisfactory as 1D, but red and green were interchanged. Clearly the process does not lead to stable color assignments in any absolute sense. Another problem: should the spatial aspects of color vision be taken into account? We suspect not if one is to view the composite as an image. Image enhancement by spatial filtering is another matter. The three perception space channels A, C_1, and C_2 have different modulation transfer functions.[6, pp. 58-74]

Table 1.

Principle Components Transformation

-.6454	-.2910	.0362	.0120	.3973	-.2734	-.4939	-.1442
.2356	.1264	-.0406	.0470	-.2396	.2934	-.8290	.3065
.4714	.3878	.0495	-.1812	.7280	-.1414	-.1366	-.1530

Transformation which Minimizes f

-.4441	-.2485	-.0040	-.5235	.5668	-.1261	-.5492	-.4266
.2721	.1634	-.1447	-.2517	-.6681	.0082	-.7080	.3029
.2802	.2787	.1073	.7353	.1515	-.3301	-.5169	-.2412

The underlying psychovisual experimentation is incomplete in that the interaction of perception and filtering A, C_1, and C_2 differently has not been resolved. Is linear filtering (as by spatial convolution) even the appropriate operation in perception space? Results we have obtained so far with image enhancement in perception space have been disappointing.

One sees, on viewing Plate 1D, that no saturated red is present. This results from our avoidance of the boundary of the color gamut. It is safe, but does leave many displayable colors unused. Can these colors be used without identifying classes which must be projected onto the boundary of the gamut to be displayed?

SUMMARY

Linear feature selection and a model for human color vision are combined to obtain a connection between multi-imagery and the human visual system. The overall objective is to preserve the spatial structure of the data as a single image, with perceived color separation matching multi-dimensional Euclidean separation in the original measurement space. The principle components feature selection technique is found to fail to separate classes obviously separated in the original data. A new distance-preserving linear map is tested and is found to accurately represent the measurement-space structure of the data. Color products are reproduced to illustrate the results. Several open problems are mentioned. An appendix giving all key details of the method is included.

6. REFERENCES

[1] MacDonald, R. B. and Hall, F. G., "Global Crop Forecasting," Science, Vol. 208, May 1980, pp. 670-679.

[2] Cramer, H., Mathematical Methods of Statistics, Princeton University Press, Princeton, N.J., 1946.

[3] Rao, C. R., Linear Statistical Inference and Its Applications, John Wiley and Sons, New York, 1973.

[4] Toussaint, G. T., "On Some Measures of Information and Their Application to Pattern Recognition", Proceedings of Conference on Measures of Information and Their Applications, Indian Institute of Technology, Bombay, India, Aug. 16-18, 1974.

[5] Toussaint, G. T., "On the Divergence Between Two Distributions and the Probability of Misclassification of Several Decision Rules", Proceeding of the Second International Joint Conference on Pattern Recognition, Copenhagen, Aug., 1974.

[6] Trouborst, P. M., and et al., "New Families of Probabilistic Distance Measures", Proceedings of the Second International Joint Conference on Pattern Recognition, Copenhagen, Aug., 1974.

[7] Devijver, P. A., "On a New Class of Bounds on Bayes Risk in Multihypothesis Pattern Recognition", IEEE Trans. Inform. Theory, Vol. C-23, No. 1 Jan., 1974, pp. 70-79.

[8] Cover, T. M. and Hart, P.E., "Nearest Neighbor Pattern Classification", IEEE Trans. Inform. Theory, Vol. IT-13, pp. 21-27, Jan., 1967.

REFERENCES

[1]Jack Bryant and L. F. Guseman, Jr., Distance preserving linear feature selection, Pattern Recognition 11, 347-352 (1979).

[2]Jack Bryant, On the clustering of multidimensional pictorial data, Pattern Recognition 11, 115-125 (1979).

[3]S. G. Wheeler, P. N. Misra and Q. A. Holmes, Linear dimensionality of Landsat agricultural data with implications for classification, Proceedings of the Purdue LARS Symp. Machine Processing of Remotely Sensed Data, 29 June - 1 July, 1976. Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, Indiana (1976).

[4]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, Proceedings of the Purdue LARS Symp. Machine Processing of Remotely Sensed Data, 29 June - 1 July, 1976. Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, Indiana (1976).

[5]C. M. Hay and R. W. Thomas, et al., Development of techniques for producing strata maps and development of photointerpretive methods based on multitemporal Landsat data, Final Report for NASA Contract NAS9-14565, Remote Sensing Research Program, Berkeley, California, December, 1977.

[6]Olivier D. Faughras, Digital color image processing and phychophysics within the framework of a human visual model, Ph.D. Dissertation, University of Utah, 1976.

[7]Ernest L. Hall, Computer Image Processing and Recognition, Academic Press, New York, 1979.

[8]Richard D. Juday, Colorimetric principles as applied to multichannel imagery, NASA Technical Memorandum 58215, NASA, Lyndon B. Johnson Space Center, Houston, Texas, July, 1979.

[9]T. Kaneko, Color composite pictures from principal axis components of multispectral scanner data, IBM Journal of Research and Development 22 386-392 (1978).

[10]L. Kanal, Patterns in pattern recognition, IEEE Trans. Pattern Recognition, IT-20, 697-722 (1974).

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

[12]G. E. Lowitz, Stability and dimensionality of Karhunen-Loéve of multispectral image expansions, Pattern Recognition 10, 359-363 (1978).

APPENDIX

Let the prototypes selected by AMOEBA (or by some other method) be denoted by $y_1,\ldots,y_p$. Let A be a linear feature selection matrix to three dimensions, and let $x_i = Ay_1$. The transformed prototypes preserve some aspect of the data structure in lower dimensional space, depending, of course, on the feature selection technique. Let x_M be the mean vector of the transformed prototypes, and let $z_i = x_i - x_M$. We first determine a scale factor s_p for the prototypes. For any s_p, let $w_i = s_p z_i$. Determine s_p so that each w_i is in the parallelepiped $[-12,12] \times [-6,6] \times [-6,6]$, and at least one w_i is on a face of this parallelpiped. Let $S = s_p P^{-1}$, where P is the transformation determined by psychovisual experiments.[6] Let $u_{ij} = \exp(w_{ij})$, $i = 1,\ldots,p$ and $j = 1,2,3$. (We use the second subscript to indicate the j-th component of the vector u_i.) Let $v_i = U^{-1}u_1$. Usually v_1 would now be translated and scaled to fit the range of the display device. The imaging system we use*, however, makes transmission density linearly proportional to input rather than to the logarithm,[8, pp. 5-6] so we compute $t_{ij} = \log v_{ij}$, $j = 1,2,3$. Now determine a scale factor s_D and a display bias b such that if $d_{1J} = s_D t_{iJ} + b$ then each d_{1J} is in [0,255] and at least one d_{1J} has the value 0 and another has the value 255.

---

*The Information International FR-80 at Johnson Space Center, Houston, Texas. The machine gives transmission density linearly proportional to input in a channel with zero input on the other two channels. Transmission density is the logarithm of the ratio of the transmitted flux with and without the sample's presence in the light beam.

We are now prepared to define the transformation by which all data (not just the prototypes) is mapped to perception space. Let $E \cdot E^3 \to E^3$ be defined by $Ep_j = \exp(p_j)$, $j = 1,2,3$. Let $d = \exp(-b/s_D)$ and define $L^+ : E^3 \to E^3$ by $L^+ p_J = \log p_j$ if $p_j > d$, $L^+ p_J = -b/s_D$ if $p_j \leq d$. Finally, let $M : E^3 \to E^3$ be defined by $M(p_J) = [\min\{p_J, 255\}]$, $J = 1,2,3$. The transformation from input multi-imagery I to gun values G is

$$G = M(s_D\, L^+ U^{-1} ES(AI - x_M) + b).$$

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

4 Title and Subtitle

Numerical Trials of HISSE

5 Report Date
5 August, 1980

6 Performing Organization Code
SR-HO-00477

7 Author(s)

Charles Peters and Frank Kampe

8 Performing Organization Report No
75

9 Performing Organization Name and Address

University of Houston
Department of Mathematics
Houston, TX
77004

10 Work Unit No

11 Contract or Grant No
NAS9-14689

12 Sponsoring Agency Name and Address

National Aeronautics and Space Administration
Lyndon B. Johnson Space Center
Houston, TX 77058 Task Monitor: Dale Browne

13 Type of Report and Period Covered*
Technical Report

14 Sponsoring Agency Code

15 Supplementary Notes

16 Abstract

This paper addresses the mathematical description and implementation of the statistical estimation procedure known as the Houston Integrated Spatial/Spectral Estimator (HISSE). HISSE is based on a normal mixture model and is designed to take advantage of spectral and spatial information of LANDSAT data pixels, utilizing the initial classification and clustering information provided by the AMOEBA algorithm. HISSE calculates parametric estimates of class proportions which reduce the error inherent in estimates derived from typical "classify and count" procedures common to non-parametric clustering algorithms. HISSE also singles out spatial groupings of pixels which are most suitable for labeling classes. These calculations are designed to aid the analyst/interpreter in labeling patches with a crop class label. Finally, we report HISSE's initial performance on an actual LANDSAT agricultural ground truth data set.

17 Key Words (Suggested by Author(s))	18 Distribution Statement
acreage estimation mixture density estimation spatial/spectral clustering	

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

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

SR-HO-00477
NAS9-14689

TECHNICAL REPORT

NUMERICAL TRIALS OF HISSE

BY

Charles Peters and Frank Kampe

This report describes research in acreage estimation performed for the Supporting Research Project.

University of Houston
Department of Mathematics
Houston, TX 77004

August 5, 1980

Numerical Trials of HISSE

by

Charles Peters and Frank Kampe

1. Introduction.

The Houston Integrated Spatial/Spectral Estimator (HISSE) is a statistical estimation procedure based on a normal mixture model which is designed to take advantage of spatial associations of LANDSAT data pixels produced by an automated spatial/spectral clustering algorithm. The clustering algorithm used in this experiment is the AMOEBA algorithm developed at Texas A & M University, which is based on the three assumptions listed below [1]. AMOEBA detects spatially connected sets of LANDSAT pixels, called patches, whose elements are characterized by spectral similarity, within certian tolerances, to their neighbors.

Assumption 1: Real classes exist.

Assumption 2: Each patch contains pixels from one and only one real class.

Assumption 3 Each real class is represented by at least one patch

No absolute commitment to the agricultural nature of real classes is expressed in [1], however, there is an indication of a high degree of purity of patches with respect to ground truth labels when AMOEBA patches are plotted on ground truth maps. A more complete study, with the same conclusion, is reported in [5]. Therefore, we feel justified in identifying the real classes with ground truth labels. In addition to the three assumptions just given,

HISSE requires the following assumption.

Assumption 4. The data from each patch is normally distributed with mean and covariance depending only on the class to which it belongs.

Assumption 4 has been challenged, some might say refuted, in [2]. However, we take the position that the proper question to ask is whether assumption 4 is close enough to the truth to be useful in estimating class proportions and labeling classes with ground truth labels. The clustering portion of AMOEBA may be described as a k-means algorithm which respects patch integrity (see Assumption 2) with a novel way of determining the correct number of clusters. As such, it contains no way of compensating for the confusion arising from classes with overlapping spectral characteristics. Thus, Assumption 4 may be regarded as a step toward mitigating the error in proportion estimation which is unavoidable with the classify and count method. Henceforth, pixels contained in patches will be called pure pixels, and all others boundary pixels.

2. Mathematical Description.

It is assumed that there are m real classes, labelled $1, \cdots, m$, and p patches represented by independent random vectors $(X_1,\theta_1), \cdots, (X_p,\theta_p)$ where $\theta_J \in \{1,\cdots,m\}$ is the unknown real class to which patch J belongs and $X_J = (X_{J1},\cdots,X_{JN_J})$ is a set of N_J n-vectors representing the spectral data from the Jth patch. The θ_J are i.i.d. with $\alpha_\ell = \text{Prob}[\theta_J=\ell]$ unknown and, given that $\theta_J = \ell$, X_J is a random sample from an n-variate normal distribution $N_n(\mu_\ell,\Omega_\ell)$ with unknown mean and covariance. Notice that α_ℓ is the expected

fraction of patches belonging to class ℓ and for a given scene may be quite different from the fraction of pure pixels belonging to class ℓ, which we denote by ϕ_ℓ. The random variable ϕ_ℓ is directly related to the total acreage of the patches belonging to class ℓ.

The log likelihood function for the parameters $\alpha_\ell, \mu_\ell, \Omega_\ell$ is

1) $$L = \sum_{J=1}^{P} \log f(X_J)$$

where

2) $$f(X_J) = \sum_{\ell=1}^{m} \alpha_\ell f_\ell(X_J)$$

and $f_\ell(X_j)$ is the N_J-fold product normal density

3) $$f_\ell(X_J) = \prod_{k=1}^{N_J} N_n(X_{Jk}; \mu_\ell, \Omega_\ell).$$

Despite the apparent complexity of L, it depends on the data only through the patch means

4) $$m_J = \frac{1}{N_j} \sum_{k=1}^{N_j} X_{Jk}$$

and scatter matrices

5) $$S_J = \sum_{k=1}^{N_j} (X_{Jk}-m_J)(X_{Jk}-m_J)^T$$

Once the m_J's and S_J's are computed and stored, HISSE has no further need for the pure data.

The numerical procedure used in HISSE for finding a maximum of the likelihood function is defined by iteratively substituting into the likelihood equations, viz.

$$\alpha_\ell^{(k+1)} = \frac{1}{p} \sum_{J=1}^{p} \frac{\alpha_\ell^{(k)} f_\ell(X_J)}{f(X_j)} \tag{6}$$

$$\mu_\ell^{(k+1)} = \sum_{j=1}^{p} N_j \frac{f_\ell(X_j)}{f(X_j)} m_j \Big/ \sum_{j=1}^{p} N_j \frac{f_\ell(X_j)}{f(X_j)} \tag{7}$$

$$\Omega_\ell^{(k+1)} = \sum_{j=1}^{p} \frac{f_\ell(X_j)}{f(X_j)} R_J \Big/ \sum_{j=1}^{p} N_J \frac{f_\ell(X_J)}{f(X_j)} - \mu_\ell^{(k+1)} \mu_\ell^{(k+1)T}, \tag{8}$$

where $R_j = S_J + N_J m_J m_J^T$ is the noncentral scatter of the jth patch. The values of the parameters used in evaluating the ratios $\frac{f_\ell(X_J)}{f(X_J)}$ are those at the preceding kth step of the algorithm. It is shown in [6] that there is a unique strongly consistent solution of the likelihood equations in a neighborhood of the true parameters as $p \to \infty$ and that the iteration procedure (6)-(8) converges to the consistent solution if the starting values are near it.

Let $N = N_1 + \cdots + N_p$ be the total number of pure pixels. It is easy to show that $\mathcal{E}[\phi_\ell] = \alpha_\ell$ and $\text{var}(\phi_\ell) \leq \frac{1}{4N^2} \sum_{J=1}^{p} N_J^2$. Thus, if the patches are nearly uniform in size, the MLE of α_ℓ can be used as a predictor of ϕ_ℓ. However, the least MSE predictor of ϕ_ℓ based on the observed data (assuming that the parameters are known) is

$$\beta_\ell = \mathcal{E}[\phi_\ell | X_1, \cdots, X_p] = \frac{1}{N} \sum_{J=1}^{p} N_J \frac{\alpha_\ell f_\ell(X_J)}{f(X_j)} . \tag{9}$$

Therefore, we take β_ℓ evaluated with the maximum likelihood estimates of the parameters as our estimate of ϕ_ℓ.

In processing the boundary pixels, which typically constitute 60-70% of the scene, we assume that the boundary data consist of an independent sample from a mixture

10) $$\sum_{\ell=1}^{m} \bar{\alpha}_\ell N_n(\mu_\ell, \Omega_\ell)$$

where the component normal distributions are the same class distributions represented in the pure data, plus observations from a contaminant class (possibly corresponding to the "not in field" ground truth label) in the tails of the $N_n(\mu_\ell, \Omega_\ell)$. In other words, we assume that a boundary observation which is spectrally unlike all of the pure classes is much more likely to be from the contaminating class than an outlier from one of the pure classes. Therefore we classify as a contaminant each boundary observation X which satisfies

11) $$(x-\mu_\ell)^T \Omega_\ell^{-1} (x-\mu_\ell) > \chi^2_\alpha$$

for all $\ell = 1, \cdots, m$, where the μ_ℓ's and Ω_ℓ's are the previously estimated pure data class means and covariances and χ^2_α is a size α critical value for χ^2 with n degrees of freedom. In this experiment we chose $\alpha = .1$.

Let $Y_1, \cdots, Y_M$ denote the boundary observations remaining after rejecting those classified as contaminants. We treat $Y_1, \cdots, Y_M$ as an independent sample from the mixture density (10), with unknown mixing proportions $\bar{\alpha}_1, \cdots, \bar{\alpha}_m$

but known components $N_n(\mu_\ell, \Omega_\ell)$, and obtain a MLE of $\bar{\alpha}_1, \cdots, \bar{\alpha}_m$ by successively substituting into (6). Obviously, $Y_1, \cdots, Y_M$ is, at best, a truncated sample from the mixture (10), so that the MLE of $\bar{\alpha}_1, \cdots, \bar{\alpha}_m$ is asymptotically biased. We do not expect this effect to be a reason for serious concern. After obtaining the MLE for $\bar{\alpha}_1, \cdots, \bar{\alpha}_m$, we use as our final estimate of the number of pixels corresponding to class ℓ, the quantity $N\beta_\ell + M\bar{\alpha}_\ell$, where β_ℓ is given by (9).

3. *Implementation.*

The number of classes assumed in this experiment is determined by AMOEBA subroutines PAINT and CLASFY. PAINT produces the pure/boundary division of a 5 × 6 mile LACIE segment, an array LABELS containing a patch description for each of the pure pixel locations, and a map of the scene showing the pure and boundary pixels. CLASFY produces an array CLASS containing the final cluster designation of each of the patches. A subroutine STAT2 has been attached to AMOEBA which calculates and saves patch sizes (N_J), patch means (m_J) and noncentral patch scatters (R_J). These statistics are then passed to STAT3 which uses the CLASS array to compute the fraction (α_ℓ^0) of patches assigned to each cluster, the fraction of pure pixels assigned to each cluster, and cluster means (μ_ℓ^0) and covariances (Ω_ℓ^0) for the pure data only. These cluster statistics are used as initial estimates of the parameters for the iteration procedure described by (6)-(8). CLASFY occasionally produces a cluster with such a small number of pure pixels that an initial covariance estimate cannot be calculated. In this case the initial Ω_ℓ^0 in HISSE is obtained by averaging the cluster sample covariance with a multiple of the identity so as to insure that the condition number of Ω_ℓ^0 is no greater than 16.

After initialization HISSE produces iterative estimates $\alpha_\ell^{(k)}, \mu_\ell^{(k)}, \Omega_\ell^{(k)}$ of the parameters until a convergence criterion is satisfied, after which the estimates β_ℓ are computed in the manner described in Section 2 and stored.

The boundary pixels are identified from the LABELS array output by AMOEBA. For each one, the quadratic forms $(x-\mu_\ell)^T\Omega_\ell^{-1}(x-\mu_\ell)$ are computed and tested against the threshold value of χ^2_α, as in (11). For those boundary pixels not rejected by the thresholding procedure, the likelihood ratios $f_\ell(x)/f_k(x)$ are computed and stored in a temporary disc file for use in the iteration procedure for estimating $\bar{\alpha}_1, \cdots, \bar{\alpha}_m$. Although the number of boundary pixels processed is much greater than the number of patches, the cost is comparable to that of processing the pure data because the iteration procedure (6) can be carried out simply by accessing the temporary file.

For the purpose of labeling classes HISSE identifies for each class ℓ, the three patches j which have the highest posterior probability $\frac{\alpha_\ell f_\ell(X_j)}{f(X_j)}$ in that class. The spatial coordinates of pixels in these labeling patches are obtained from the LABELS array. Thus, in using HISSE, the analyst would be required to make a judgement concerning the identity of each class based on his ability to label the labeling patches.

4. Numerical Results.

The results tabulated in this section are from four passes over LACIE segment 1618 acquired in May, June, August and September of 1976. The data was preprocessed by premultiplying each single pass 4-dimensional data vector by the LANDSAT I transformation to brightness-greenness space

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -1 & 1 & 1 \end{pmatrix}$$

and stacking the brightness-greenness vectors to obtain 8-dimensional data vectors. The results of the AMOEBA run were 7500 pure pixels, organized into 310 patches. The number of clusters estimated by NUMCLU was 13. HISSE required 19 iterations to estimate the parameters of the pure data mixture model. Of the 15290 boundary pixels, the thresholding procedure rejected 5575. The number of passes through the remaining 9725 boundary pixels required to produce estimates of the boundary mixing proportions $\bar{\alpha}_1, \cdots, \bar{\alpha}_{13}$ was 8. The total cost of running AMOEBA and HISSE together is much less than that of running UHMLE or CLASSY on the full scene.

Figures 1-4 show the scatter plots in brightness-greenness space, corresponding to each of the passes, of the means of the patches determined by AMOEBA. Particularly in the fourth pass, the tasseled cap configuration described in [4] is visible. Figures 5, 6, and 7 show the plotted trajectories of the estimated class means from pass to pass on the same coordinate system used in the 4th pass scatter plot. The trajectories of the means of the pure data clusters produced by AMOEBA would be nearly indistinguishable. It is interesting that the class means trajectories eventually given a small grains label exhibit a characteristic triangular shape. Obviously, this characteristic can be used as an aid in labeling the classes (see [3], for a discussion of this idea).

Figure 8 tabulates the initial cluster means, cluster variances, and patch membership proportions obtained from AMOEBA's clustering of the pure data. Figure 9 tabulates class means, variances and patch memebership probabilities (the α's) estimated by HISSE. Figure 10 compares the estimates derived from AMOEBA and HISSE of the fraction of pure pixels belonging to each cluster (class). Notice that in Figure 10, there is a significant difference between the two estimates, particularly in the more populous classes. These classes happen to be the most

spectrally confused classes. There is also an appreciable difference seen in Figures 8 and 9 between the respective estimates of the α's, although the difference is not as pronounced.

Figure 11 shows the AMOEBA boundary map for segment 1618 with the three labeling patches corresponding to each class outlined. A ground truth map was used to attach ground truth labels to the labeling patches and hence to the classes. Most of the classes were given a single ground truth label by this procedure. Classes 2, 5, 6, 7, were not assigned a single ground truth label and appeared to be made up of more than one type of small grains. However, each of these classes was clearly small grains. Class 1 was the only really difficult class to label; each of its labeling patches represented small grains ground truth labels as well as such labels as beans and fallow. In other words, the labeling patches for class 1 were spurious. For the purpose of obtaining an aggregate small grains estimate, it was assumed that class 1 was a mixture of 1/3 small grains, 1/3 beans, and 1/3 fallow acreage.

Figure 12 shows the final acreage estimate for each of the 13 classes in the mixture model, the acreage of the set C of boundary pixels rejected as outliers or contaminants, and the crop labels (including "small grains") assigned to each class. The aggregate small grains acreage estimate is 15,288. The small grains acreage from the ground truth tape is 15,465, an error of only 1.1%. If class 1 is labelled all small grains, the error is 15%. If none of class 1 is classified small grains, the error is 9.2%. It should be emphasized that the problem of labeling cluster #1 from AMOEBA is also serious, since cluster 1 is centered near the means of the spurious patches used to label class 1.

The thresholding of boundary outliers makes a pronounced difference in the

estimate. The small grains acreage estimate derived from HISSE without thresholding would be 19,230, comparable to the estimate of 20,336 derived from AMOEBA's cluster map.

5. Conclusions.

The accuracy with which HISSE estimated the small grains acreage in segment 1618 was impressive, to say the least, but of course the procedure must be tested on other segments for which ground truth is available. Also, as we mentioned in Section 4, the accuracy of the estimate depends on the classification given to the labeling fields for class 1, the problem class. The procedure we used-dividing the class evenly among competing ground truth labels - seems fair; however, in an operational situation the class would be labeled by an analyst looking at a film product and it seems unlikely that he would apportion the class in such a way. In any case, the greatest possible relative error was 15%, still a marked improvement over the accuracy obtained by labeling AMOEBA's clusters and counting the cluster assignments, or that achieved by HISSE without the thresholding procedure.

The performance of HISSE, or AMOEBA, depends in large part upon the purity with respect to ground truth labels of the patches found by AMOEBA, which is influenced by the user defined "percent in fields" parameter in AMOEBA. In this experiment we defined the parameter as 50%; that is, we conservatively estimate that 50% of the pixels in the scene should be found in fields. By reducing the size of this parameter, we expect to produce a higher degree of patch purity and thus alleviate the problem of having a class represented by labeling patches which should not be patches at all. We hope that this will not aggravate another

problem, namely that the ground truth map for segment 1618 shows a few large fields representing important classes (such as barley) in which no patches were found.

Finally, we note that although the aggregated small grains acreage was very accurately estimated, the individual estimates for the various small grains classes (spring wheat, barley, oats, and millet) were not nearly as accurate. Indeed, several of the HISSE classes could not be given a single one of these labels, although they clearly represented small grains. Moreover, there was one significant crop class (beans) without a small grains label which was seriously underestimated. Thus, the usefulness of HISSE in a multicrop inventory cannot yet be determined.

REFERENCES

1. Jack Bryant, On the clustering of multidimensional pictorial data, Pattern Recognition, 11(2), pp. 115-125, 1979.

2. W. D. Davenport and S. G. Wheeler, Multispectral data analysis based on ground truth crop classes. IBM Report No. RES 23-83, IBM Federal Systems Division, Houston, Texas, May 1980.

3. J. L. Engvall, D. Tubbs, and Q. A. Holmes, Pattern recognition of Landsat data based upon temporal trend analysis. Remote Sensing of the Environment, 6, 303-314, (1977).

4. R. J. Kauth and G. Thomas, The tasseled cap-a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat, Purdue/LARS Symposium on Machine Processing of Remotely Sensed Data, Purdue University, 1976.

5. R. M. Myers, Spatial-spectral procedure development: the purity experiment, IBM Report No. SR-IO-00445, IBM Federal Systems Division, Houston, Texas, April 1980.

6. B. C. Peters, Jr., On the consistency of the maximum likelihood estimate of normal mixture parameters for a sample with field structure. Report No. 74, Department of Mathematics, University of Houston, Sept 1979.

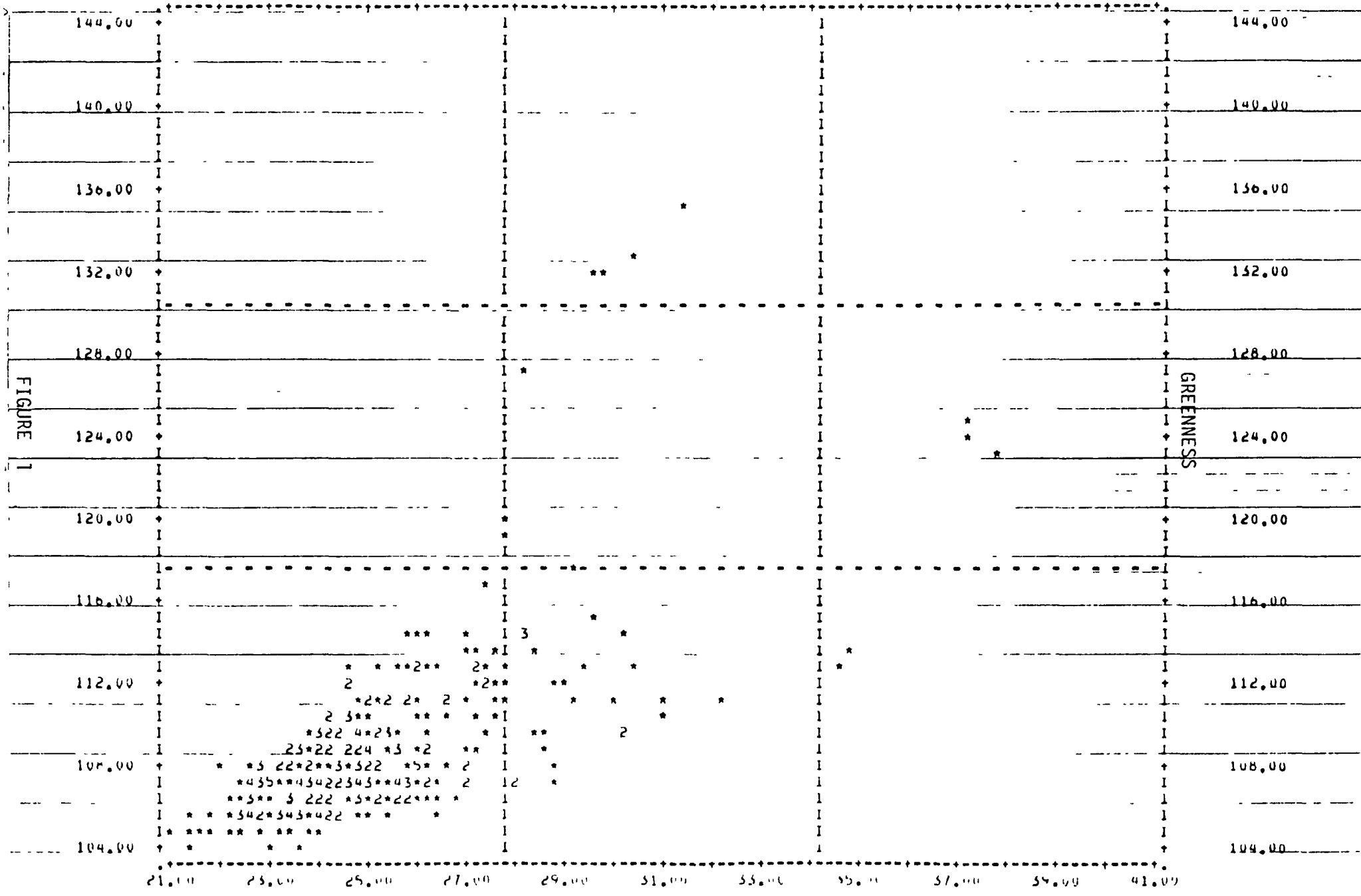

FILE ... (C-RATIO DATE = 08-15-80)
SCATTERGRAM OF (...) Y
(AL-USS) X
22.00 24.00 26.00 28.00 30.00 32.00 34.00 36.00 38.00 40.00
144.00
140.00
136.00
132.00
128.00
124.00
120.00
116.00
112.00
108.00
104.00
21.00 23.00 25.00 27.00 29.00 31.00 33.00 35.00 37.00 39.00 41.00
GREENNESS
BRIGHTNESS

FIGURE 1

1618

PASS 2

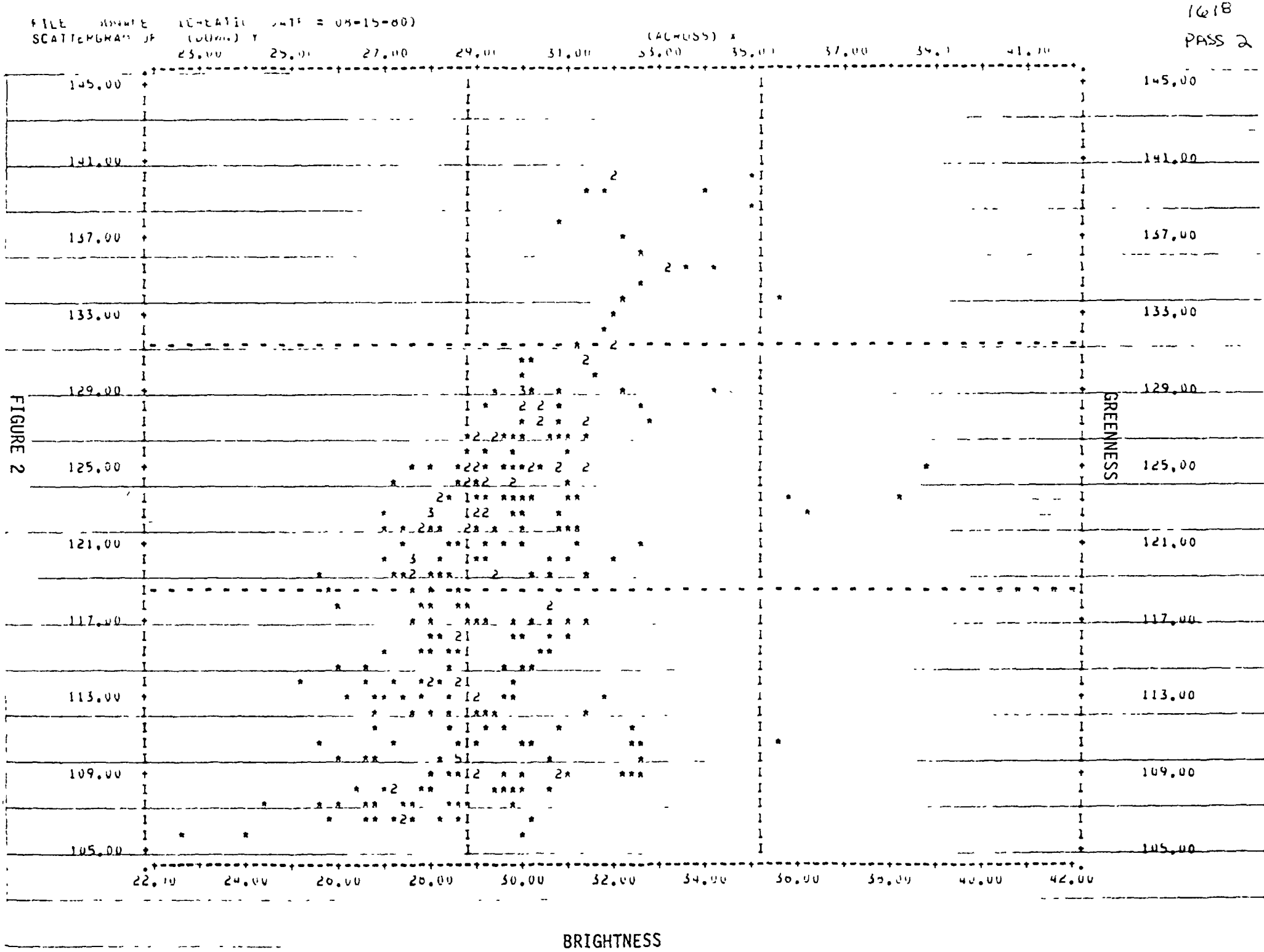


FIGURE 2

PASS 3

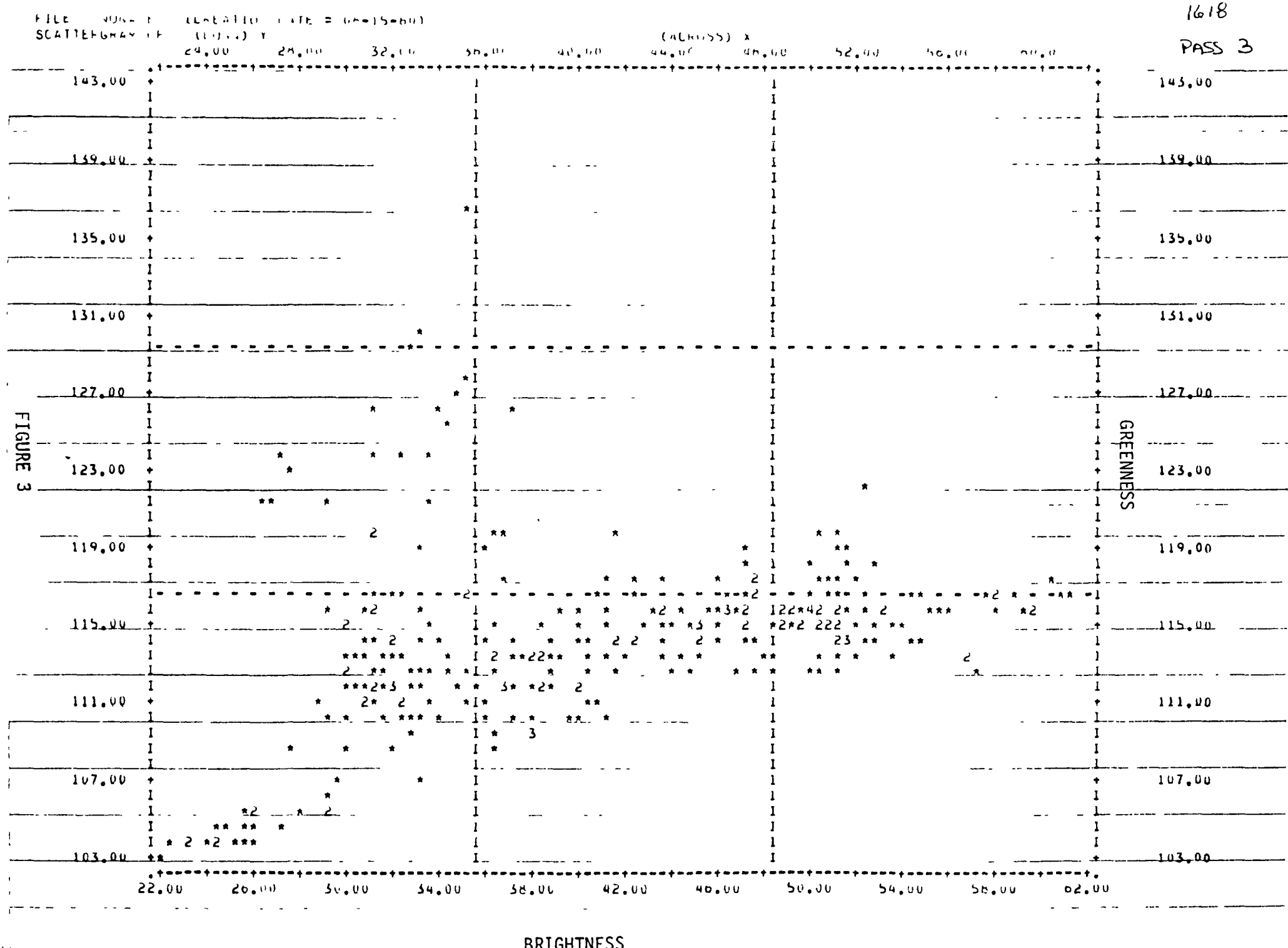


FIGURE 3

1618

PASS 4

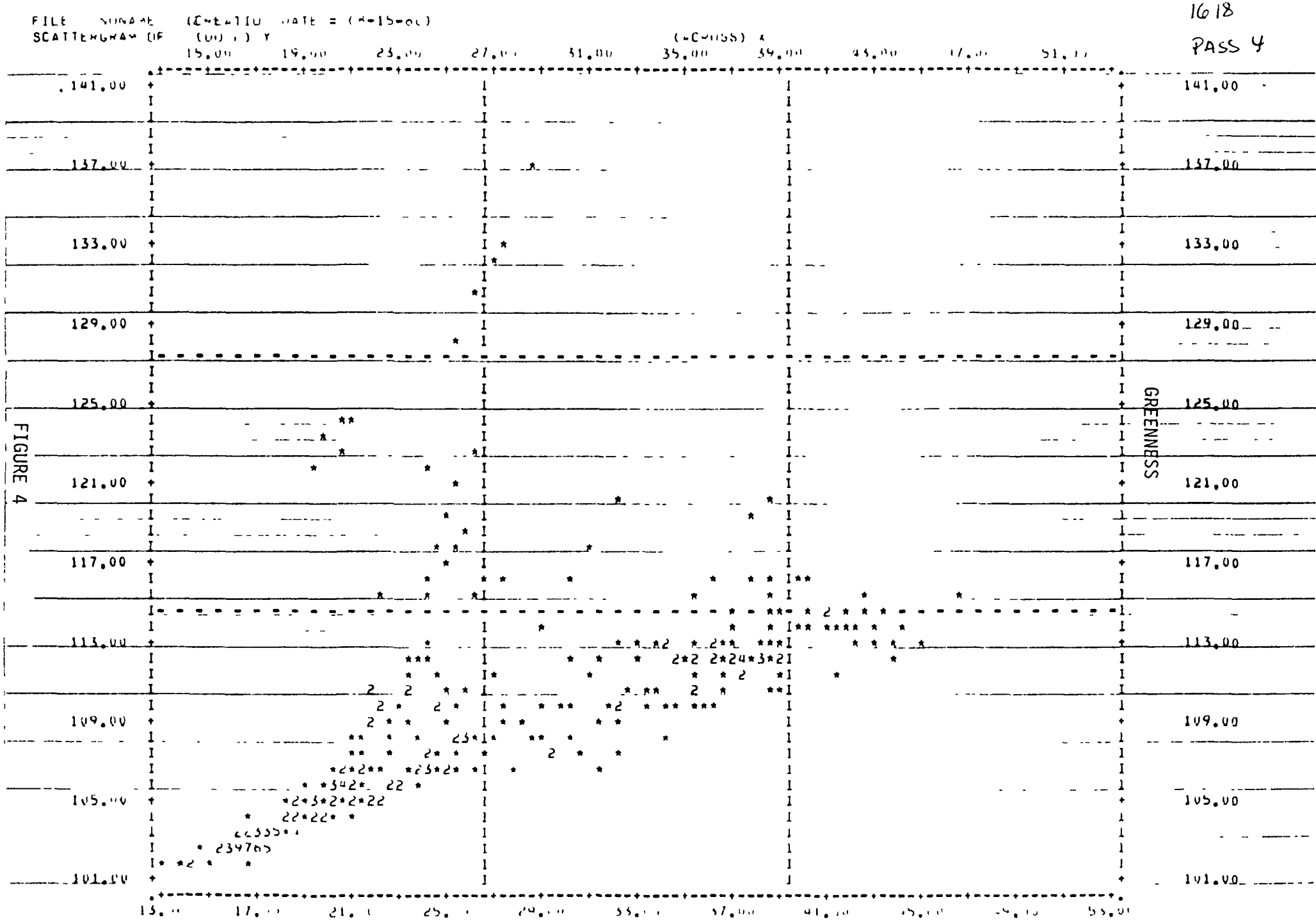


FIGURE 4

FINAL CLASS TRAJECTORIES

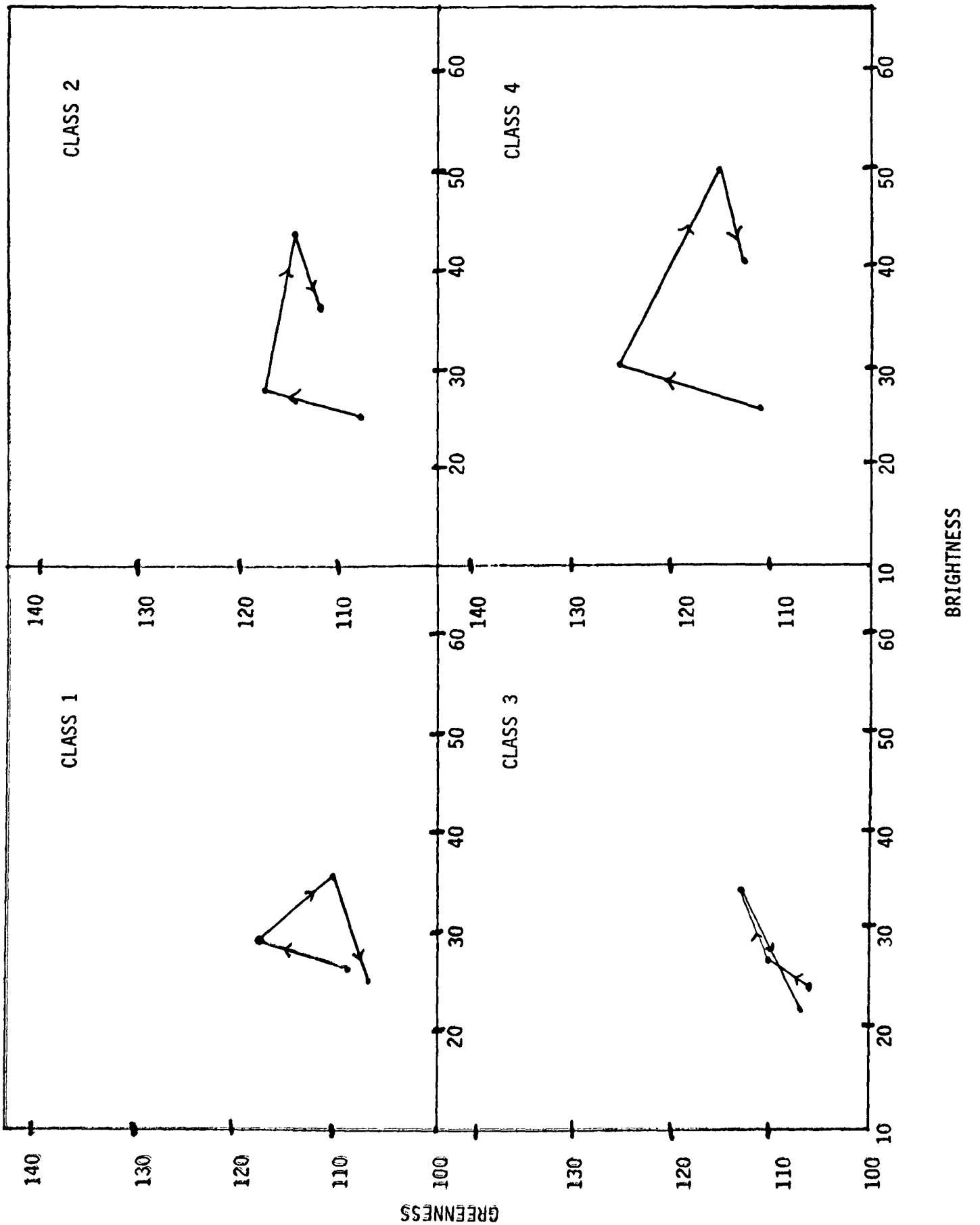


FIGURE 5

FINAL CLASS TRAJECTORIES

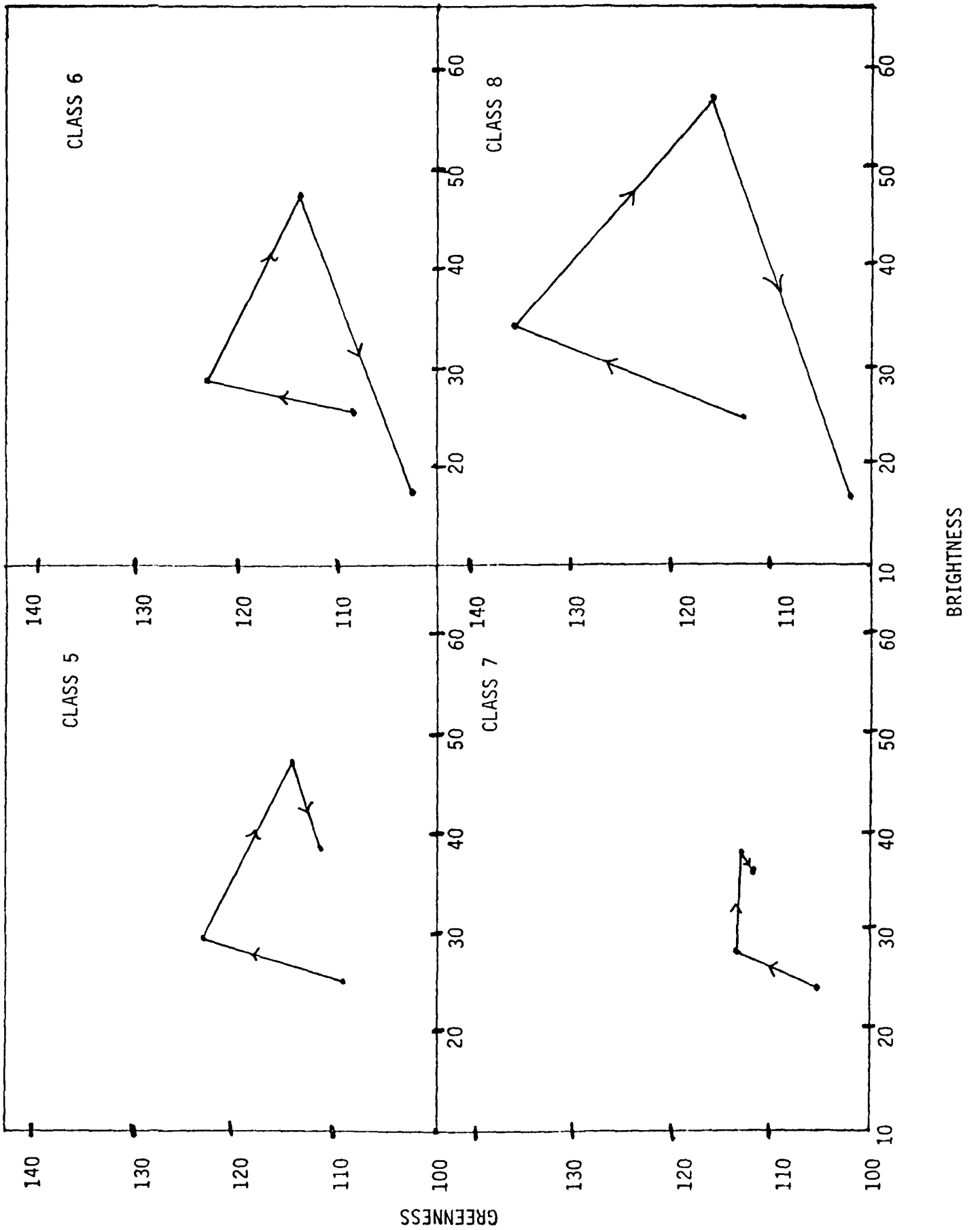


FIGURE 6

FINAL CLASS TRAJECTORIES

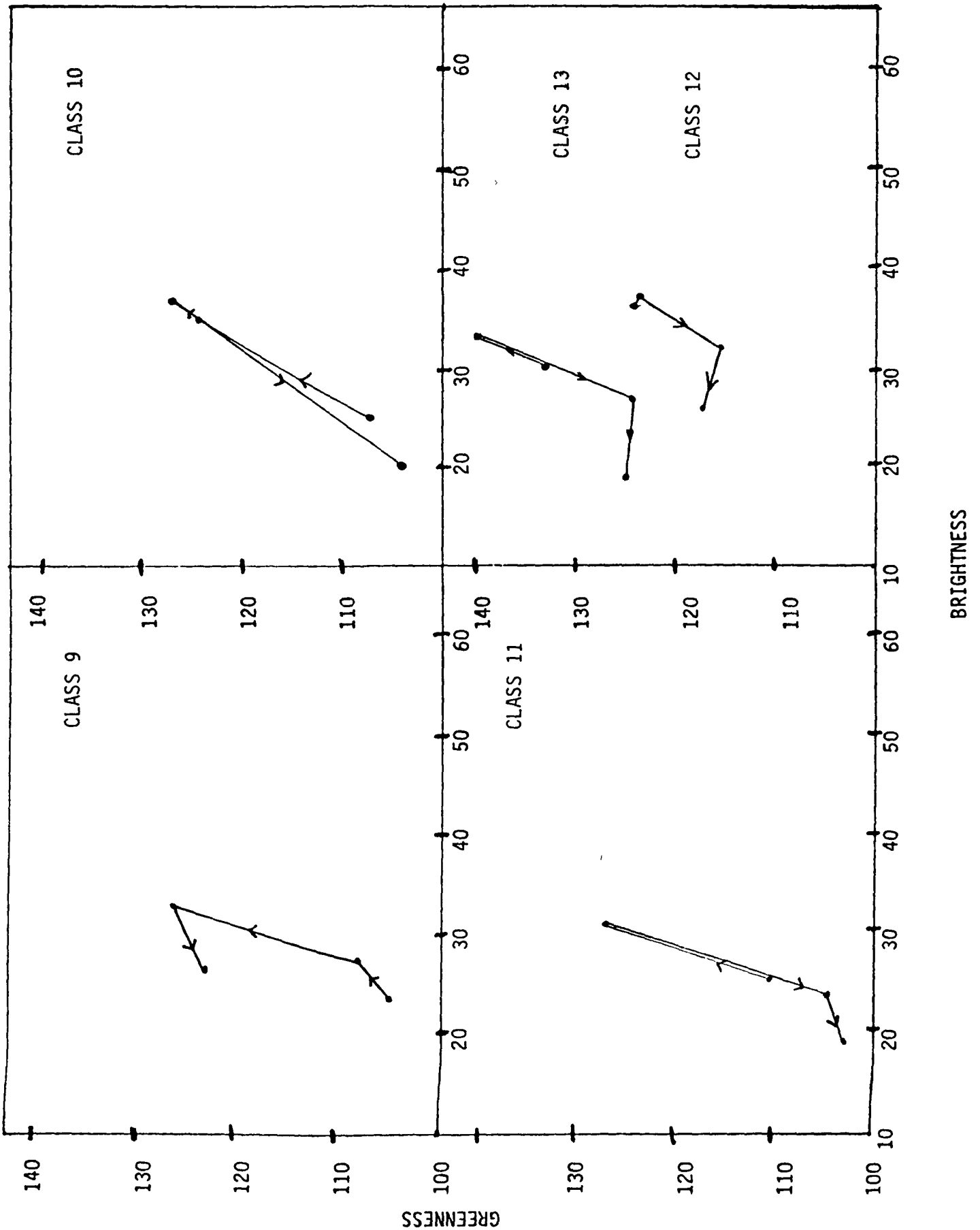


FIGURE 7

INITIAL CLASS STATISTICS (AMOEBA)

CLUSTER #	CLUSTER MEAN								PATCH PROPORTION
1	26.84	110.39	29.79	121.70	36.49	111.02	26.44	108.04	.077
2	24.99	108.48	28.17	117.42	44.25	115.57	34.05	112.63	.094
3	24.80	106.86	28.82	111.90	32.59	111.73	21.69	107.00	.271
4	25.51	111.64	30.29	127.63	50.08	115.15	39.10	113.13	.094
5	25.46	108.75	29.26	122.53	48.90	114.94	36.61	111.77	.100
6	25.09	109.24	29.35	123.39	48.80	114.94	18.15	103.83	.158
7	23.90	106.14	28.76	113.53	38.15	113.07	37.15	112.73	.058
8	25.05	112.20	33.45	135.38	56.52	116.32	17.19	102.97	.026
9	23.26	105.98	29.02	108.48	34.30	125.54	25.91	121.94	.048
10	25.50	107.50	35.75	123.25	37.25	126.50	20.25	104.75	.003
11	25.49	110.83	30.71	128.90	24.92	104.16	19.04	104.01	.045
12	37.60	123.64	37.76	123.44	31.92	116.60	25.48	118.12	.010
13	30.16	132.47	31.80	139.64	27.37	123.07	20.68	123.83	.016

CLUSTER VARIANCE

1	7.98	10.82	3.22	36.25	51.31	16.82	32.68	10.60
2	6.09	10.51	3.25	25.33	33.50	8.50	23.18	18.36
3	7.87	5.24	7.29	32.49	29.88	18.48	17.25	12.48
4	4.54	18.49	2.48	15.77	32.80	7.96	16.41	5.97
5	9.11	4.70	3.13	21.46	27.59	6.43	19.92	6.90
6	4.64	8.34	4.26	38.13	44.59	6.00	11.12	6.22
7	4.74	2.60	6.14	22.52	15.73	11.22	37.19	7.90
8	1.50	3.18	3.61	12.71	15.00	1.84	3.43	1.59
9	2.90	3.42	5.40	11.30	11.44	24.02	8.12	53.75
10	4.25	0.25	0.69	35.19	11.19	4.25	1.19	3.69
11	4.00	5.83	5.35	33.79	5.26	1.55	8.07	3.38
12	3.28	2.56	2.90	3.69	1.43	3.61	3.93	3.95
13	1.75	9.97	1.38	5.20	1.31	2.81	1.09	3.41

FIGURE 8

FINAL CLASS STATISTICS (HISSE)

CLASS #	CLASS MEAN	PATCH PROBABILITY
1	26.91 109.19 29.64 117.57 35.07 110.50 25.53 107.45	.126
2	24.62 108.52 27.91 117.84 44.68 115.93 35.13 113.58	.083
3	24.11 106.34 28.61 110.87 33.73 113.30 21.65 107.51	.221
4	25.58 111.88 30.23 126.89 50.83 115.51 39.97 113.64	.084
5	25.30 108.73 29.41 123.19 48.09 114.35 35.83 111.28	.108
6	25.10 109.25 29.36 123.38 48.73 114.95 18.20 103.89	.170
7	23.89 106.13 28.78 113.49 38.08 113.06 37.04 112.70	.061
8	25.06 112.25 33.47 135.41 56.65 116.35 17.13 102.93	.023
9	23.26 105.98 29.02 108.48 34.30 125.55 25.91 121.94	.048
10	25.50 107.50 35.75 123.25 37.25 126.50 20.25 104.75	.003
11	25.25 110.37 29.80 127.20 24.86 104.14 19.07 103.99	.048
12	37.60 123.64 37.76 123.44 31.92 116.60 25.48 118.12	.010
13	30.16 132.47 31.80 139.64 27.37 123.07 20.68 123.83	.016

CLASS VARIANCE

1	9.56	10.44	5.08	51.15	72.18	24.81	44.57	12.14
2	3.76	10.02	2.71	23.47	35.32	8.02	15.39	17.05
3	4.66	3.29	6.93	25.02	21.94	9.55	14.74	13.05
4	4.78	20.68	2.74	19.15	39.22	7.15	16.30	4.31
5	9.48	4.02	2.98	26.54	20.81	4.94	18.06	6.76
6	4.60	8.04	4.29	38 42	44 64	5.61	11.24	6.09
7	4.66	2.34	6.15	22.65	15.92	11.02	37.65	7.82
8	1.53	3.19	3.62	12.65	14.57	1.81	3.33	1.50
9	2.89	3.24	5.36	11.27	11.47	23.77	8.18	53.66
10	4.25	0.26	0.69	35.20	11.19	4.25	1.19	3.70
11	3.78	5.89	8.48	42.06	4.79	1.75	6.84	2.88
12	3.07	3.24	3.00	3.00	1.31	3.32	4.07	3.66
13	1.64	9.20	1.49	5.16	1.30	2.49	0.99	3.85

FIGURE 9

PURE PIXEL PROPORTIONS(Φ_K)

CLUSTER #	AMOEBA ESTIMATE	CLASS #	HISSE ESTIMATE(β_k)
1	.054	1	.143
2	.136	2	.107
3	.259	3	.188
4	.101	4	.089
5	.109	5	.123
6	.171	6	.174
7	.067	7	.068
8	.021	8	.021
9	.034	9	.034
10	.001	10	.001
11	.031	11	.038
12	.003	12	.003
13	.012	13	.012

FIGURE 10

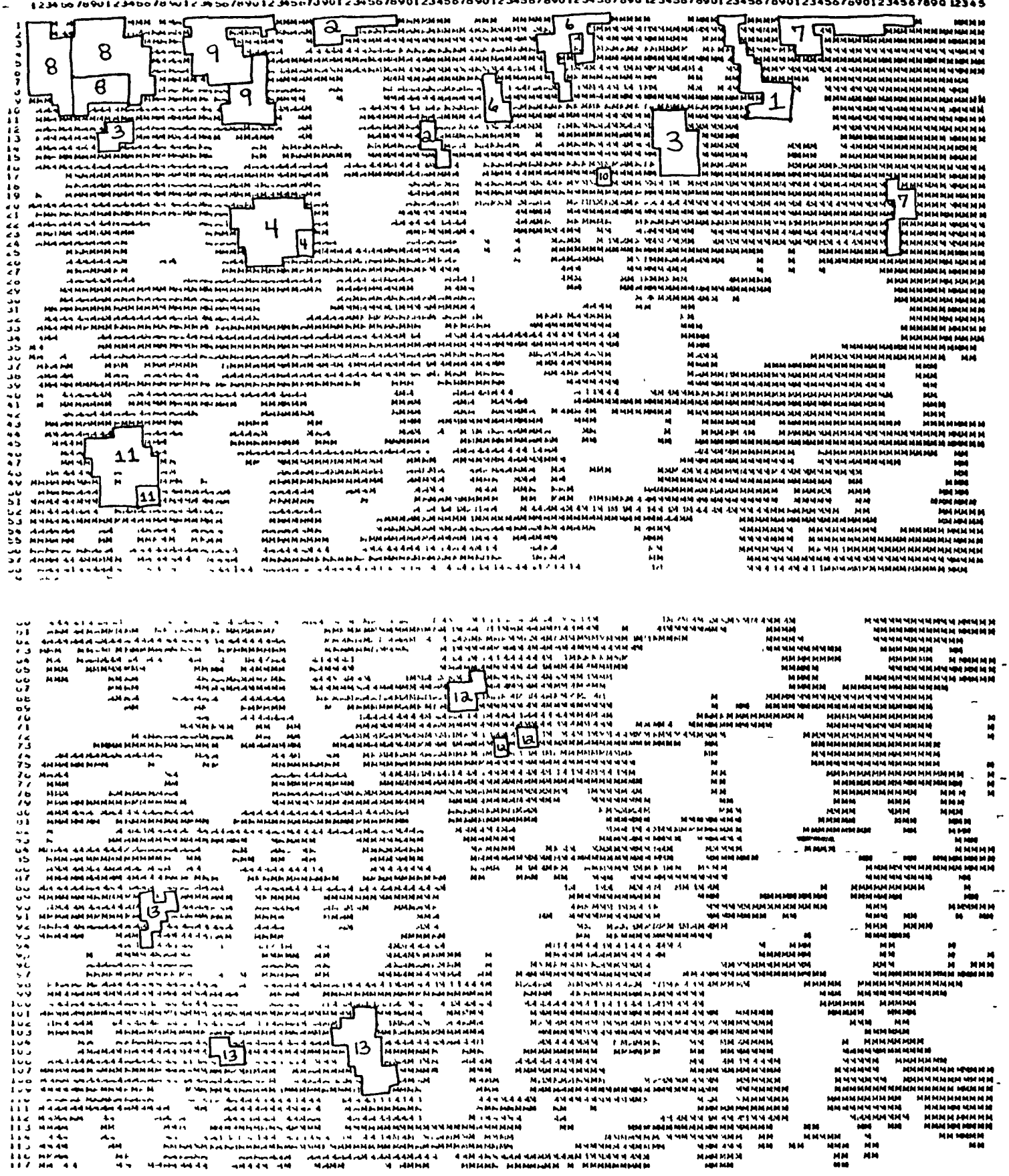


FIGURE 11 (PART 1)

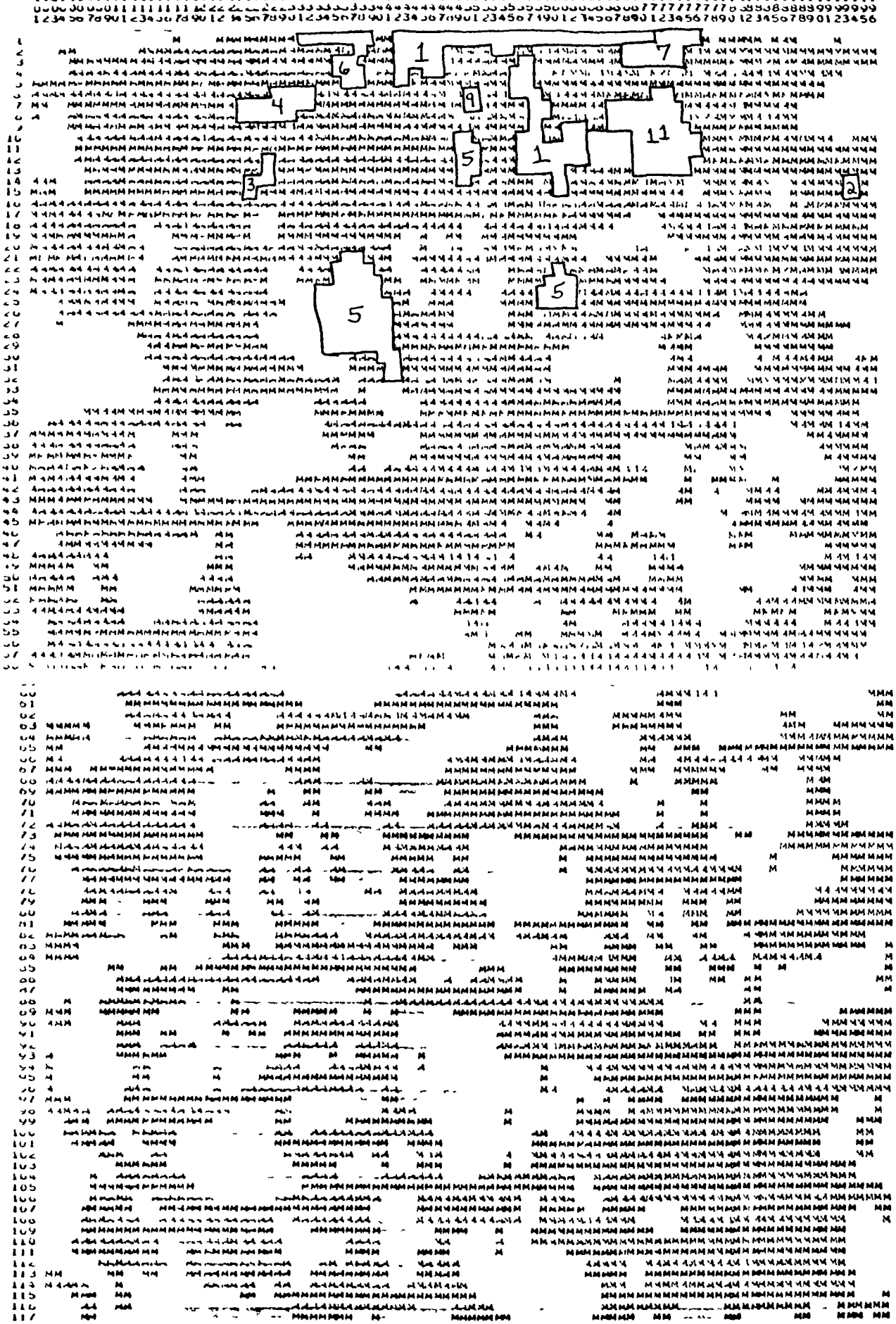


FIGURE 11 (PART 2)

CLASS ACREAGE ESTIMATES

CLASS	ACREAGE	CROP LABEL
1	3764	Small Grains Beans Idle Fallow
2	1550	Small Grains
3	3560	Spring Wheat
4	1237	Spring Wheat
5	2253	Small Grains
6	3257	Small Grains
7	1218	Small Grains
8	262	Spring Wheat
9	917	Idle Cover Crop
10	4	Flax
11	697	Barley
12	49	Homestead
13	171	Trees
C	6124	Contaminated Data

FIGURE 12

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

4 Title and Subtitle	5 Report Date
On the Existence, Uniqueness, and Asymptotic Normality of A Consistent Solution of The Likelihood Equations For Nonidentically Distributed Observations-Applications to Missing Data Problems	September 10, 1980
	6 Performing Organization Code SR-HO-00492
7 Author(s) Charles Peters	8 Performing Organization Report No 76
	10 Work Unit No
9 Performing Organization Name and Address University of Houston Department of Mathematics Houston, TX 77004	11 Contract or Grant No NAS9-14689
	13 Type of Report and Period Covered Technical Report
12 Sponsoring Agency Name and Address National Aeronautics and Space Administration Lyndon B. Johnson Space Center Houston, TX 77058 Task Monitor: Dale Browne	14 Sponsoring Agency Code

Supplementary Notes

Abstract

A general theorem is given which establishes the existence and uniqueness of a consistent solution of the likelihood equations given a sequence of independent random vectors whose distributions are not identical but have the same parameter set. In addition, it is shown that the consistent solution is a MLE and that it is asymptotically normal and efficient. Two applications are discussed: one in which independent observations of a normal random vector have missing components, and the other in which the parameters in a mixture from an exponential family are estimated using independent homogeneous sample blocks of different sizes.

Key Words (Suggested by Author(s))	18 Distribution Statement
Strong consistency of maximum likelihood estimates; asymptotic normality; asymptotic efficiency; missing data; exponential families; mixture densities	

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

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

SR-HO-00492
NAS9-14689

TECHNICAL REPORT

ON THE EXISTENCE, UNIQUENESS, AND ASYMPOTIC NORMALITY OF A CONSISTENT SOLUTION OF THE LIKELIHOOD EQUATIONS FOR NONIDENTICALLY DISTRIBUTED OBSERVATIONS - APPLICATIONS TO MISSING DATA PROBLEMS

BY

CHARLES PETERS

UNIVERSITY OF HOUSTON
DEPARTMENT OF MATHEMATICS
HOUSTON, TX 77004

SEPTEMBER 10, 1980

1. Introduction

This paper is concerned with the existence, uniqueness, and asymptotic properties of a strongly consistent local maximizer of the likelihood function for a vector parameter in the case of nonidentically distributed samples and without prior assumptions which insure the existence of a global MLE. Well known results pertaining to scalar parameters and i.i.d. samples date back to theorems of Cramér [5] and Huzurbazar [11], while results concerning the consistency of the MLE, under assumptions that insure a unique MLE, may be found in Wald [17], Wolfowitz [19], and LeCam [12]. Somewhat more recently, Silvey [15] has dealt with the asymptotic properties of the MLE without independence. Surprisingly however, a correct proof of the multidimensional version of the combined results of Cramér and Huzurbazar on the existence of a unique consistent solution of the likelihood equations when multiple roots occur did not appear until 1977 in a note by Foutz [10], (see also Tarone & Gruenhage [16], Chanda [3], and Peters and Walker [14,Appendix].) Examples 1 and 2 which follow illustrate the need for a consistency theorem along these lines which relaxes the assumption of identically distributed observations.

Example 1 (Observations with missing components): Let X_1, X_2, ... be independent random vectors in R^n whose common density is one of a parametric family $\{q(x|\theta)\}_{\theta\in\Theta}$, where Θ is a subset of R^ν. Suppose that instead of the X_i we observe only certain subvectors B_1X_1, B_2X_2, ..., where $\{B_i\}$ is a given sequence of $n_i \times n$ matrices obtained by deleting $n - n_i$ rows from the identity. Clearly we can assume that components are missing at random provided that the B_i's are independent of the X_i's. Under what conditions is there a unique

strongly consistent (and asymptotically efficient) local MLE of θ based on the observations B_1X_1, B_2X_2, ...?

A recent paper by Dahiya and Korwar [6] illustrates that even for a bivariate normal sample, with several simplifying restrictions on the sample and on the parameters, the likelihood equation for Example 1 has multiple roots and requires numerical methods for its solution.

Example 2 (Estimating mixture density parameters with sample blocks of varying sizes): Let $f(x|\tau_1)$, $f(x|\tau_2)$, ..., $f(x|\tau_m)$ be unknown but distinct members of a multivariate parametric family $\{f(x|\tau)\}_{\tau\in T}$, and let α_1, ..., α_m be the unknown positive probabilities corresponding to a discrete mixing distribution supported on $\{\tau_1, \ldots, \tau_m\}$. The number m is known. Under what conditions will there be a unique consistent MLE of the parameter $\theta = (\alpha_1, \ldots, \alpha_{m-1}, \tau_1, \ldots, \tau_m)$ describing the mixture density $q(x|\theta) = \sum_{i=1}^{m} \alpha_i f(x|\tau_i)$, based on a sample of the type X_1, X_2, ..., where the X_i are independent and each X_i is itself a random sample $X_i = (X_{i1}, \ldots, X_{iNi})$ of known size from an unknown component density $f(x|\tau_i)$? In this example the parameter θ is only locally identifiable. Moreover, it can easily occur that the likelihood function is unbounded [9]; hence, the need for a consistency theorem for local maximizers is especially clear.

The practical importance of Example 2 is indicated by the fact that estimation of mixture density parameters is often proposed as an alternative to the clustering of large amounts of multivariate data [18]. The asymptotic properties of the MLE are of interest because of the prevalence of large sample considerations in judging cluster validity [8], even though it may be difficult to argue for a statistical basis for a given clustering problem. The presentation of the data in blocks of varying size may occur when the primary sampling units are grouped by physical or spatial associations (see [2] and [13] for an

application of this idea in the analysis of pictorial data.)

Finally we remark that the existence and uniqueness of a consistent solution of the likelihood equations bears on the numerical problem of obtaining the estimate. Each of Examples 1 and 2 is a missing data problem (in Example 2 the random variables which indicate the component population of origin are missing); thus, a natural numerical procedure for obtaining a MLE is one derived from the generalized EM procedure of Dempster, Laird, and Rubin [7]. Such a procedure increases the value of the likelihood at each iterative step; however, this is no guarantee of convergence, since the likelihood function may be unbounded. Generally speaking it is possible to show that the Hessian of the log likelihood is negative definite near the consistent solution of the likelihood equations. Thus, the generalized EM procedure is convergent to it given a good enough starting value (see [14] for a thorough discussion of numerical properties in the case of a mixture of multivariate normal distributions.)

Throughout this paper the symbol E_θ will denote expectation with respect to a distribution determined by a parameter θ and D_u, $D^2_{u,v}$ etc. will denote differentiation or partial differentiation with respect to scalar or vector variables u, v. For a scalar valued function, ∇_u will denote the gradient with respect to an inner product which will usually be understood from the context. Given an inner product $\langle \cdot | \cdot \rangle$ and a vector σ, the symmetric k-linear form $f(\eta_1, \ldots, \eta_k) = \prod_{j=1}^{k} \langle \sigma | \eta_i \rangle$ will be denoted by $\langle \sigma | \cdot \rangle^k$. Thus, for example, we may write the covariance of a statistic S as $\mathrm{Cov}_\tau(S) = E_\tau\{\langle S - E_\tau(S) | \cdot \rangle^2\}$. The largest and smallest eigenvalues of a symmetric positive definite operator A will be denoted respectively by $\rho(A)$ and $\sigma(A)$.

2. A General Consistency Theorem. Let Θ be an open subset of R^{ν} and for each positive integer r and each $\theta \in \Theta$, let $q_r(\cdot|\theta)$ be an N_r-variate density with respect to some fixed σ-finite measure λ_r on R^{N_r}. Let $\theta^o \in \Theta$ and let $X_1, \ldots, X_p, \ldots$ be a sequence of independent random vectors with X_r having density $q_r(\cdot|\theta^o)$. For $\theta \in \Theta$ define

$$L_p(\theta) = \sum_{r=1}^{p} \log q_r(X_r|\theta)$$

Theorem 1: Suppose

(i) $$\int_{R^{N_r}} D_\theta\, q_r(x|\theta^o)\, d\lambda_r(x) = 0 ,$$

(ii) $$\int_{R^{N_r}} D_\theta^2\, q_r(x|\theta^o)\, d\lambda_r(x) = 0 ,$$

and that there is a constant M, functions f_r, a neighborhood Ω of θ^o and λ_r-null sets A_r in R^{N_r} such that for all r, $\theta \in \Omega, x \notin A_r$,

(iii) $$|D^3_{\theta_i, \theta_j, \theta_k} \log q_r(x|\theta)| \le f_r(x) \qquad i, j, k = 1, \ldots, \nu$$

(iv) $$E_{\theta^o}\{f_r(X_r)^2\} \le M$$

(v) $$E_{\theta^o}\{[D_{\theta_i} \log q_r(X_r|\theta^o)]^4\} \le M \qquad i = 1, \ldots, \nu$$

(vi) $$E_{\theta^o}\left\{ \frac{1}{q_r(X_r|\theta^o)^2} [D^2_{\theta_i,\theta_j} q_r(X_r|\theta^o)]^2\right\} \le M \qquad i, j = 1, \ldots, \nu$$

and

(vii) there exists $\epsilon > 0$ such that $\frac{1}{p}\sum_{r=1}^{p} J_r(\theta^o) \ge \epsilon I_\nu$ for sufficiently large p,

where $J_r(\theta^o) = E_{\theta^o}\{\nabla_\theta \log q_r(X_r|\theta^o)\, \nabla_\theta^T \log q_r(X_r|\theta^o)\}$, I_ν is the identity on R^{ν}, and the ordering is the usual one on symmetric operators. Then there is a neighborhood Ω^o of θ^o such that with probability 1 there is an integer p_1 such that for $p \ge p_1$ there is a unique solution θ^p in Ω^o of the likelihood equation

$D_\theta L_p(\theta) = 0$. Furthermore, $\theta^p \to \theta^o$ as $p \to \infty$ and θ^p is a maximum likelihood estimate. The consistent estimator θ^p is asymptotically normal and asymptotically efficient.

Proof: In the proof we make repeated use of the following version of the strong law [4, p. 103]: let $Z_1, Z_2, \ldots$ be uncorrelated random variables such that the variances of the Z_i are bounded. Then $\frac{1}{n}\sum_{j=1}^{n}(Z_u - E[Z_j]) \to 0$ a.s. as $n \to \infty$.

Let $S_p(\theta) = \frac{1}{p}\sum_{r=1}^{p} D_\theta \log q_r(X_r|\theta)$. By (i) $E_{\theta^o}\{S_p(\theta^o)\} = 0$ and by (v) $S_p(\theta^o) \to 0$ a.s. as $p \to \infty$. Consider the $\nu\times\nu$ matrix $D_\theta S_p(\theta^o)$ whose i, j^{th} element is

$$\frac{1}{p}\sum_{r=1}^{p} D^2_{\theta_i,\theta_j} \log q_r(X_r|\theta^o) = \frac{1}{p}\sum_{r=1}^{p} \frac{1}{q_r(X_r|\theta^o)} D^2_{\theta_i,\theta_j} q_r(X_r|\theta^o)$$

$$- \frac{1}{p}\sum_{r=1}^{p} D_{\theta_i} \log q_r(X_r|\theta^o) D_{\theta_j} \log q_r(X_r|\theta^o).$$

By (ii) the expected value of the first term on the right is zero. Hence, by (v) and (vi)

$$D_\theta S_p(\theta^o) + \frac{1}{p}\sum_{r=1}^{p} J_r(\theta^o) \to 0$$

a.s. as $p \to \infty$. Thus, with probability 1, if $0 < \eta < \epsilon/2$ there is $p_o \in N$ so that for $p \geq p_o$

$$D_\theta S_p(\theta^o) \leq -2\eta I .$$

Without loss of generality we can assume Ω is convex. For $\theta \in \Omega$,

$$\frac{1}{p}\sum_{r=1}^{p} |D^2_{\theta_i,\theta_j} \log q_r(X_r|\theta) - D^2_{\theta_i,\theta_j} \log q_r(X_r|\theta^o)|$$

$$\leq \frac{1}{p}\sum_{r=1}^{p}\sum_{k=1}^{\nu} |\theta_k - \theta_k^o| \int_0^1 |D^3_{\theta_i,\theta_j,\theta_k} \log q_r(X_r|\theta^o + t(\theta - \theta^o))|dt$$

$$\leq \frac{1}{p}\sum_{r=1}^{p}\sum_{k=1}^{\nu} |\theta_k - \theta_k^o| f_r(X_r)$$

With probability 1, for large p

$$\frac{1}{p}\sum_{r=1}^{p} f_r(X_r) \le 1 + \frac{1}{p}\sum_{r=1}^{p} E_{\theta^o}\{f_r(X_r)\}$$

$$\le 1 + M^{\frac{1}{2}} .$$

It follows that for any particular norms on R^ν and on the symmetric $\nu\times\nu$ matrices there is a constant $\bar{M}$ such with probability 1 there is a positive integer p_1 such that for $p \ge p_1$, $\theta \in \Omega$,

$$||D_\theta S_p(\theta) - D_\theta S_p(\theta^o)|| \le \bar{M}||\theta - \theta^o|| .$$

Thus there is a convex neighborhood Ω^o of θ^o such that

$$D_\theta S_p(\theta) \le - \eta I$$

for all $\theta \in \Omega^o$, $p \ge p_1$. It now follows that for $p \ge p_1$ S_p is one to one on Ω^o and that the image under S_p of the sphere $\Omega_\delta(\theta^o)$ at θ^o of small radius δ contains the sphere $\Omega_{\eta\delta}(S_p(\theta^o))$ at $S_p(\theta^o)$ of radius $\eta\delta$. Since 0 is eventually in $\Omega_{\eta\delta}(S_p(\theta^o))$ there is a unique solution of $D_\theta S_p(\theta) = 0$ in $\Omega_\delta(\theta^o)$. Since $D_\theta S_p(\theta)$ is negative definite, this solution is a MLE.

Let $\Sigma_p = \frac{1}{p}\sum_{r=1}^{p} J_r(\theta^o)$. The Cramér-Rao lower bound for p observations is verified without difficulty to be $(p\,\Sigma_p)^{-1}$. By (v), (vii), and Liapounovs Theorem [4, p. 200], $p^{\frac{1}{2}}\,\Sigma_p^{-\frac{1}{2}}\,S_p(\theta^o)$ is asymptotically distributed as $N_\nu(0, I)$. Moreover, in a neighborhood of θ^o we may write

$$S_p(\theta) = S_p(\theta^o) + A(\theta)(\theta - \theta^o)$$

where $A(\theta) \to D_\theta S_p(\theta^o)$ as $\theta \to \theta^o$. It follows that with probability 1.

$$p^{\frac{1}{2}}\,\Sigma_p^{\frac{1}{2}}\,(\theta^p - \theta^o) = - \Sigma_p^{\frac{1}{2}}\,A(\theta^p)^{-1}\,\Sigma_p^{\frac{1}{2}}\,p^{\frac{1}{2}}\,\Sigma_p^{-\frac{1}{2}}\,S_p(\theta^o)$$

for large p. Since $D_\theta S_p(\theta^o) + \Sigma_p \to 0$ and $A(\theta^p) \to D_\theta S_p(\theta^o)$ with probability 1,

the expression $-\Sigma_p^{\frac{1}{2}} A(\theta^p)^{-1}\Sigma_p^{\frac{1}{2}}$ converges almost surely to the identity. Therefore, $p^{\frac{1}{2}} \Sigma_p^{\frac{1}{2}} (\theta^p - \theta^o)$ is asymptotically $N_\nu(0,I)$ and θ^p is asymptotically efficient. This concludes the proof.

3. Applications.

Suppose that in Example 1 the X_i have a common n variate normal distribution $N_n(\mu, \Sigma)$ and it is desired to estimate μ, Σ by maximum likelihood based on the observed components B_1X_1, B_2X_2, ..., B_pX_p. The likelihood equations for μ and Σ are

$$\sum_{r=1}^{p} B_r^T(B_r \Sigma B_r^T)^{-1} B_r\mu = \sum_{r=1}^{p} B_r^T(B_r\Sigma B_r^T)^{-1}B_rX_r \ . \tag{3.1}$$

and

$$\sum_{r=1}^{p} B_r^T(B_r\Sigma B_r^T)^{-1}B_r = \sum_{r=1}^{p} B_r^T(B_r\Sigma B_r^T)^{-1}B_r(X_r - \mu)(X_r - \mu)^T B_r^T(B_r\Sigma B_r^T)^{-1}B_r \ . \tag{3.2}$$

and have no explicit solution, although for given Σ (3.1) may be solved explicitly for μ provided that the matrix an the left of (3.2) is invertible.

Components i and j are <u>paired</u> in the observation B_rX_r if both the i^{th} and j^{th} columns of B_r contain a 1. Let $\phi(i, j, p)$ denote the relative frequency with which the i^{th} and j^{th} components are paired in the first p observations B_1X_1, ..., B_pX_p, and let $\phi_1(i, j) = \lim_{p \to \infty} \phi(i, j, p)$.

<u>Theorem 2</u>: Let X_1, X_2, ... be independent, identically distributed according to $N_n(\mu, \Sigma)$. If $\phi_1(i, j) > 0$ for all $i, j = 1, \ldots, n$, then there is a unique strongly consistent solution of the likelihood equations (3.1) and (3.2), which has the asymptotic properties given in Theorem 1.

<u>Proof</u>: The only one of conditions (i) - (vii) in Theorem 1 which poses any

difficulty is number (vii). For $\theta = (\mu, \Sigma)$, the information matrix $J_r(\theta)$ corresponding to the density of $B_r X_r$,

$$q_r(\cdot|\theta) = N_{n_r}(B_r\mu, B_r\Sigma B_r^T) ,$$

is

$$J_r(\theta) = \left[\begin{array}{c|c} U_r(\theta) & 0 \\ \hline 0 & U_r(\theta) \otimes U_r(\theta) \end{array}\right] , \tag{3.3}$$

where $U_r(\theta) = B_r^T(B_r\Sigma B_r^T)^{-1}B_r$, and the Kronecker product $U_r(\theta) \otimes U_r(\theta)$ represents the symmetric operator on $n \times n$ real symmetric matrices S (with trace inner product) defined by $U_r(\theta)SU_r(\theta)$. Thus (vii) is satisfied if for each Σ there exists $\epsilon = \epsilon(\Sigma)>0$ such that for all p sufficiently large

$$\frac{1}{p}\sum_{r=1}^{p} Z^T B_r^T(B_r\Sigma B_r^T)^{-1}B_r Z \geq \epsilon Z^T Z \tag{3.4}$$

and

$$\frac{1}{p}\sum_{r=1}^{p} \mathrm{Tr}[B_r^T(B_r\Sigma B_r^T)^{-1}B_r S]^2 \geq \epsilon \mathrm{Tr}S^2 \tag{3.5}$$

for all $Z \in R^n$ and symmetric S. However, (3.5) implies (3.4), as can be seen by taking $S = ZZ^T$. Hence, it suffices to establish (3.5) under the stated hypotheses.

Now,

$$\begin{aligned} &\mathrm{Tr}[B_r^T(B_r\Sigma B_r^T)^{-1}B_r S]^2 \\ &= \mathrm{Tr}[(B_r\Sigma B_r^T)^{-1}(B^r S B_r^T)]^2 \\ &= \mathrm{Tr}[(B_r\Sigma B_r^T)^{-\frac{1}{2}}(B_r S B_r^T)(B_r\Sigma B_r^T)^{-\frac{1}{2}}]^2 \\ &\geq \sigma[(B_r\Sigma B_r^T)^{-\frac{1}{2}} \otimes (B_r\Sigma B_r^T)^{-\frac{1}{2}}]\ \mathrm{Tr}[B_r S B_r^T]^2 \end{aligned}$$

But,

$$\sigma[(B_r\Sigma B_r^T)^{-\frac{1}{2}} \otimes (B_r\Sigma B_r^T)^{-\frac{1}{2}}] = 1/\rho[(B_r\Sigma B_r^T)^{\frac{1}{2}} \otimes (B_r\Sigma B_r^T)^{\frac{1}{2}}]$$

and

$$\rho[(B_r\Sigma B_r^T)^{\frac{1}{2}} \otimes (B_r\Sigma B_r^T)^{\frac{1}{2}}] = \sup_{Tr\Delta^2\le 1} Tr(B_r\Sigma B_r^T)^{\frac{1}{2}}\,\Delta\,(B_r\Sigma B_r^T)\,\Delta\,(B_r\Sigma B_r^T)^{\frac{1}{2}}$$

$$= \sup_{Tr\Delta^2\le 1} Tr[(B\Sigma B_r^T)\Delta]^2$$

$$= \sup_{Tr\Delta^2\le 1} Tr\Sigma B_r^T\Delta B_r\Sigma B_r^T\Delta B_r$$

$$= \sup_{Tr\Delta^2\le 1} Tr[\Sigma^{\frac{1}{2}}B_r^T\Delta B_r\Sigma^{\frac{1}{2}}]^2$$

$$\le \rho[\Sigma^{\frac{1}{2}} \otimes \Sigma^{\frac{1}{2}}] \sup_{Tr\Delta^2\le 1} Tr[B_r^T\Delta B_r]^2$$

$$= \rho[\Sigma^{\frac{1}{2}} \otimes \Sigma^{\frac{1}{2}}] \ .$$

The last equation follows from $B_rB_r^T = I_{n_r}$. Hence,

$$Tr[B_r^T(B_r\Sigma B_r^T)^{-1}B_rS]^2 \ge \sigma[\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}}]\ Tr[B_rSB_r^T]^2$$

$$= \sigma[\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}}]\ Tr[B_r^TB_rSB_r^TB_r]^2 \quad .$$

Therefore,

$$\frac{1}{p}\sum_{r=1}^{p} Tr[B_r^T(B_r\Sigma B_r^T)^{-1}B_rS]^2 \ge \sigma[\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}}] \cdot \frac{1}{p}\sum_{r=1}^{p} Tr[B_r^TB_rSB_r^TB_r]^2$$

$$\ge \sigma[\Sigma^{-\frac{1}{2}} \otimes \Sigma^{-\frac{1}{2}}]\sigma[\frac{1}{p}\sum_{r=1}^{p}(B_r^TB_r) \otimes (B_r^TB_r)]TrS^2$$

Since eventually

$$\sigma[\frac{1}{p}\sum_{r=1}^{p}(B_r^TB_r) \otimes (B_r^TB_r)] > \frac{1}{2}\min_{i,j}\phi_1(i,j) \quad ,$$

(vii) follows upon taking $c = \frac{1}{2}\min_{i,j}\phi_1(i,j) \cdot \rho[\Sigma^{\frac{1}{2}} \otimes \Sigma^{\frac{1}{2}}]$. QED.

The second application of Theorem 1 is to the problem outlined in Example 2. We assume that the unknown component densities $f(x|\tau_i)$ are from a regular exponential family (see [1] for definitions) with minimal canonical representation

(3.6) $$f(x|\tau) = C(\tau) \exp \langle \tau | F(x) \rangle \qquad (\tau \in T)$$

with respect to a σ-finite measure λ, where T is an open subset of a finite dimensional space V with inner product $\langle \cdot | \cdot \rangle$. We also assume that for distinct $\tau_1, \ldots, \tau_m$, the functions $e^{\langle \tau_1 | F(x) \rangle}, \ldots, e^{\langle \tau_m | F(x) \rangle}$, together with any components of $F(x)e^{\langle \tau_1 | F(x) \rangle}, \ldots, F(x)e^{\langle \tau_m | F(x) \rangle}$ are linearly independent $[\lambda]$. The joint density of $X_r = (X_{r1}, \ldots, X_{rNr})$, given that X_r is a sample from $f(x|\tau_\ell)$ is

(3.7) $$p_r(x_r|\tau_\ell) = \gamma_r(\tau_\ell) \exp \langle \tau_\ell | G_r(x_r) \rangle$$

where $$x_r = (x_{r1}, \ldots, x_{rNr})$$

$$\gamma_r(\tau_\ell) = C(\tau_\ell)^{N_r}$$

and

$$G_r(x_r) = \sum_{j=1}^{N_r} F(x_{rj}) \quad .$$

The log-likelihood for the parameter $\theta = (\alpha_1, \ldots, \alpha_{m-1}, \tau_1, \ldots, \tau_m)$ of Example 2, based on the sample $X_1, \ldots, X_p$ is

(3.8) $$L_p(\theta) = \sum_{r=1}^{p} \log q_r(X_r|\theta) \quad ,$$

where

(3.9) $$q_r(X_r|\theta) = \sum_{\ell=1}^{m} \alpha_\ell \, p_r(X_r|\tau_\ell)$$

and $p_r(X_r|\tau_\ell)$ is given by (3.7). The following lemma collects some facts about exponential families which we require. For proofs, see Barndorff-Nielsen [1] .

Lemma 1: Let (1) be a canonical representation of an exponential family. For $\tau \in T$ let $\kappa(\tau) = -\ln.C(\tau) = \ln \int_{R^n} \exp\langle \tau | F(x) \rangle d\lambda(x)$

Then

(i) For each $\tau \in T$, $F(x)$ has moments of all orders with respect to $f(x|\tau)$;

(ii) $\kappa(\tau)$ has derivatives of all orders which may be obtained by differentiating under the integral sign. $D_\tau^k \kappa(\tau)$ may conveniently be represented as a symmetric k-linear form on V whose coefficients are polynomials in the first k moments of F. In particular,

(iii) $$D_\tau \kappa(\tau) = \langle E_\tau(F) | \cdot \rangle = \int_{R^n} \langle F(x) | \cdot \rangle f(x|\tau) d\lambda(x)$$

and

(iv) $$D_\tau^2 \kappa(\tau) = \mathrm{cov}_\tau(F) = \int_{R^n} \langle F - E_\tau(F) | \cdot \rangle^2 f(x|\tau) d\lambda(x)$$; $D_\tau^2 \kappa(\tau)$ is positive definite.

(v) $\kappa(\tau)$ is strictly convex on T.

We are now ready to establish consistency of the MLE in Example 2.

Theorem 3: If the numbers $\{N_r\}$ are bounded and $L_p(\theta)$ is given by (3.8) then with probability 1 there is a unique consistent solution of $D_\theta L_p(\theta) = 0$ which, moreover, is a MLE of the parameter $\theta^o = (\alpha_1^o, \ldots, \alpha_{m-1}^o, \tau_1^o, \ldots, \tau_m^o)$ and is asymptotically normal and efficient.

Proof: Write $\mu_r(\tau_\ell) = E_{\tau_\ell}(G_r)$; $\mu(\tau_\ell) = E_{\tau_\ell}(F)$. Using Lemma 1, the nonzero derivatives of $q_r(x_r|\theta)$ up to order 2 are:

(3.10) $$D_{\alpha_\ell} q_r(x_r|\theta) = p_r(x_r|\tau_\ell) - p_r(x_r|\tau_m) \ , \quad 1 \le \ell \le m-1$$

(3.11) $$D_{\tau_\ell} q_r(x_r|\theta) = \alpha_\ell p_r(x_r|\tau_\ell) \langle G_r(x_r) - \mu_r(\tau_\ell) | \cdot \rangle \ , \quad 1 \le \ell \le m$$

$$(3.12)\qquad D^2_{\tau_\ell,\alpha_\ell} q_r(x_r|\theta) = p_r(x_r|\tau_\ell)\langle G_r - \mu_r(\tau_\ell)|\cdot\rangle \ , \quad 1 \le \ell \le m-1$$

$$(3.13)\qquad D^2_{\tau_m,\alpha_\ell} q_r(x_r|\theta) = -p_r(x_r|\tau_m)\langle G_r - \mu_r(\tau_m)|\cdot\rangle \ , \quad 1 \le \ell \le m-1$$

$$(3.14)\qquad D^2_{\tau_\ell} q_r(x_r|\theta) = \alpha_\ell p_r(x_r|\tau_\ell)\{\langle G_r - \mu_r(\tau_\ell)|\cdot\rangle^2 - \mathrm{cov}_{\tau_\ell}(G_r)\} \ , \quad 1 \le \ell \le m \ .$$

Conditions (i) and (ii) of Theorem 1 follow immediately from (3.10) - (3.14). Similarly, using Lemma 1 and the boundedness of $\{N_r\}$, conditions (iii) - (vi) of Theorem 1 are readily verified. It remain to verify (vii). We may write $J_r(\psi)$ in matrix form as

$$J_r(\theta) = \begin{bmatrix} I_1 & 0 \\ 0 & N_r^{\frac{1}{2}} I_2 \end{bmatrix} E_\theta \begin{bmatrix} A_r & B_r \\ B_r^* & C_r \end{bmatrix} \begin{bmatrix} I_1 & 0 \\ 0 & N_r^{\frac{1}{2}} I_2 \end{bmatrix}$$

where I_1 and I_2 are respectively the identity operators on R^{m-1} and V^m and

$$A_r = \left(\frac{[p_r(X_r|\tau_\ell) - p_r(X_r|\tau_m)][p_r(X_r|\tau_k) - p_r(X_r|\tau_m)]}{q_r(X_r|\theta)^2} \right) \qquad \ell, k = 1, \ldots, m-1$$

$$B_r = \left(\frac{\alpha_k p_r(X_r|\tau_k)[p_r(X_r|\tau_\ell) - p_r(X_r|\tau_m)]}{q_r(X_r|\theta)^2} N_r^{-\frac{1}{2}} \langle G_r - \mu_r(\tau_k)|\cdot\rangle \right) \qquad \begin{matrix} \ell = 1, \ldots, m-1 \\ k = 1, \ldots, m \end{matrix}$$

$$C_r = \left(\frac{\alpha_\ell \alpha_k p_r(X_r|\tau_\ell) p_r(X_r|\tau_k)}{q_r(X_r|\theta)^2} N_r^{-1} (G_r - \mu_r(\tau_k))\langle G_r - \mu_r(\tau_\ell)|\cdot\rangle \right) \qquad k, \ell = 1, \ldots, m.$$

The assumptions concerning the linear dependence of the functions $\exp\langle\tau|F(x)\rangle$ and $F(x)\exp\langle\tau|F(x)\rangle$ insure that $J_r(\theta)$ is positive definite for each r. Condition (vii) will be established once it is shown that the smallest eigenvalue of $J_r(\theta)$ is bounded away from zero as $N_r \to \infty$.

Clearly,

$$\sigma(J_r(\theta)) \geq \sigma\left(E_\theta\begin{bmatrix} A_r & B_r \\ B_r^* & C_r \end{bmatrix}\right) .$$

Observe that

$$\frac{p_r(X_r|\tau_\ell)}{p_r(X_r|\tau_k)} = \exp\{-N_r[\kappa(\tau_\ell) - \kappa(\tau_k) - \langle \tau_\ell - \tau_k | \frac{1}{N_r} G_r \rangle]\} .$$

If X_r is a sample from $f(x|\tau_k)$, then the expression in square brackets converges to

$$\kappa(\tau_\ell) - \kappa(\tau_k) - \langle \tau_\ell - \tau_k | E_{\tau_k}(F) \rangle = \kappa(\tau_\ell) - \kappa(\tau_k) - \kappa'(\tau_k) \cdot (\tau_\ell - \tau_k)$$

which is positive by the strict convexity of κ. Hence,

$$\frac{p_r(X_r|\tau_\ell)}{p_r(X_r|\tau_k)} \to 0 \quad \text{as } N_r \to \infty .$$

Therefore,

$$E_\theta\left[\frac{p_r(X_r|\tau_\ell)p_r(X_r|\tau_k)}{q_r(X_r|\theta)^2}\right] = E_{\tau_k}\left[\frac{p_r(X_r|\tau_\ell)}{q_r(X_r|\theta)}\right] .$$

converges to 0 if $\ell \neq k$ and $\frac{1}{\alpha_k}$ if $\ell = k$ as $N_r \to \infty$. Thus,

$$E_\theta[A_r] \to \left(\frac{1}{\alpha_m^2} + \frac{\delta_{\ell k}}{\alpha_k^2}\right) \quad \text{as } N_r \to \infty .$$

Given that X_r is from $f(x|\tau_k)$, $N_r^{-\frac{1}{2}}(G_r - \mu_r(\tau_k))$ converges in distribution to a normal random variable Z with mean zero and covariance $\text{cov}_{\tau_k}(F)$. Hence,

$$\frac{p_r(X_r|\tau_\ell)}{q_r(X_r|\theta)} N_r^{-\frac{1}{2}}(G_r - \mu_r(\tau_k))$$

converges in distribution to 0 if $\ell \neq k$ and $\frac{1}{\alpha_k} Z$ if $\ell = k$.

Let Λ be any element of V and consider

$$[N_r^{-\frac{1}{2}} <G_r - \mu_r(\tau_k)|\Lambda>]^4 = N_r^{-2}[\sum_{j=1}^{N_r} <F(X_{rj}) - E_{\tau_k}(F)|\Lambda>]^4$$

After expanding and taking expectation with respect to τ_k, it will be seen that the only nonvanishing terms are those of the form

$$E_{\tau_k}[<F(X_{rj}) - E_{\tau_k}(F)|\Lambda>^2 <F(X_{r\ell}) - E_{\tau_k}(F)|\Lambda>^2]$$

of which there are $N_r + \binom{N_r}{2} = O(N_r^2)$. Thus

$$E_{\tau_k}[N_r^{-\frac{1}{2}} <G_r - \mu_r(\tau_k)|\Lambda>]^4$$

is bounded as $N_r \to \infty$. It follows from a standard theorem on convergence of moments [4, p. 95] that

$$E_{\tau_k}\left[\frac{p_r(X_r|\tau_\ell)}{q_r(X_r|\theta)} N_r^{-\frac{1}{2}} (G_r - \mu_r(\tau_k))\right] \to 0 \quad \text{as } N_r \to \infty .$$

Thus $E_\theta(B_r) \to 0$. Similar reasoning shows that

$$E_\theta(C_r) \to (\delta_{k\ell} \operatorname{cov}_{\tau_k}(F))$$

as $N_r \to \infty$. Therefore $\dot{\sigma}(J_r(\theta))$ is bounded away from 0 and this concludes the proof.

4. Concluding Remarks.

Theorem 3 remains true under weake assumptions then the boundedness of the sample sizes N_r, but nothing like the approach embodied in Theorem 1 will work without some restrictions on N_r. Nevertheless, it is far from

intuitively clear that restrictions are needed for the existence of a consistent MLE. Similarly, it seems plausible that the assumption in Theorem 2 that components be paired with nonzero asymptotic frequency might also be weakened. In certain cases, e.g., when a normal mean is to be estimated from data with missing components and the covariance is the identity, the existence of a consistent MLE with desirable asymptotic properties can be shown under weaker hypotheses than those derived from Theorem 1. The condition in Theorem 1 that $\phi_1(i, j) > 0$ for all i and j is nevertheless reasonable since it is equivalent to the condition that the Cramer-Rao lower bound be of the order of $\frac{1}{p}$ as $p \to \infty$.

REFERENCES

1. O. Barndorff-Nielsen (1978). Information and Exponential Families in Statistical Theory. John Wiley and Sons, New York.

2. J. Bryant (1979). On the clustering of multidimensional pictorial data, Pattern Recognition, 11(2), 115-125.

3. K. C. Chanda (1954). A note on the consistency and maxima of the roots of likelihood equations, Biometrika, 41, 56-61.

4. K. L. Chung (1974). A Course in Probability Theory, Second Edition, Academic Press, New York.

5. H. Cramér (1946). Mathematical Methods of Statistics, Princeton University Press, Princeton, N.J.

6. R. C. Dahiya and R. M. Korwar (1980). Maximum likelihood estimates for a bivariate normal distribution with missing data, Annals of Statistics, 8(3), 687-692.

7. A. D. Dempster, N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Statist. Soc. B, 39, 1-38.

8. R. Dubes and A. K. Jain (1979). Validity Studies in clustering methodologies, Pattern Recognition, 11(2), 235-254.

9. R. O. Duda and P. E. Hart (1973). Pattern Classification and Scene Analysis, John Wiley and Sons, New York.

10. V. Foutz (1977). On the unique consistent solution to the likelihood equations, JASA, 72, 357, 147-148.

11. V. S. Huzurbazar (1948). The likelihood equation, consistency and the maxima of the likelihood function, Annals of Eugenics, 14, 3, 185-200.

12. L. LeCam (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, Univ. of California Publ. in Statist., 1, 277-330.

13. C. Peters and F. Kampe (1980). Numerical trials of HISSE, Report no. 75, Department of Mathematics, Univ. of Houston.

14. B. C. Peter, Jr. and H. F. Walker (1978). An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions, SIAM J. Appl. Math. B, 35, 2, 362-378.

15. S. D. Silvey (1961). A note on the maximum likelihood in the case of dependent random variables, J. Roy. Statist. Soc. B, 23, 444-452.

16. R. D. Tarone and G. Grueuhage (1975). A note on the uniqueness of roots of the likelihood equations for vector valued parameters, JASA 70, 903-904.

17. A Wald (1949). A note on the consistency of the consistency of the maximum likelihood estemate, Ann. Math. Statist. 20, 595-601.

18. J. H. Wolfe (1970). Pattern clustering by multivariate mixture analysis, Multivariate Behaviorial Research 5, 329-350.

19. J. Wolfowitz (1949). On Wald's proof of the consistency of the maximum likelihood estimate, Ann. Math. Statist. 20, 601-602.

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

4 Title and Subtitle
An Iterative Procedure for Obtaining Maximum Likelihood Estimates in a Mixture Model

5 Report Date
September, 1980

6 Performing Organization Code

7 Author(s)
Richard A. Redner
The University of Tulsa

8 Performing Organization Report No
1

10 Work Unit No

9 Performing Organization Name and Address
Division of Mathematical Sciences
The University of Tulsa
Tulsa, Oklahoma 74104

11. Contract or Grant No
NAS-9-14689

13 Type of Report and Period Covered
Unscheduled Technical

12 Sponsoring Agency Name and Address
Earth Ovservations Division
Johnson Space Center
Houston, Texas 77058

14 Sponsoring Agency Code

15 Supplementary Notes

Principal Investigator: L. F. Guseman, Jr.

16 Abstract

In this paper we investigate the problem of estimating the parameters for a mixture of densities from, possibly distinct, exponential families. The likelihood equations used by Hasselblad (1969) are necessary conditions for a local maximum of the likelihood function. We show that a particular repeated substitution scheme, determined by the likelihood equations, converges locally to the strongly consistent maximum likelihood estimate. This generalizes the results of Peters and Walker (1978).

17 Key Words (Suggested by Author(s))	18 Distribution Statement
Exponential families, maximum likelihood estimate, mixture densities.	

19 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

AN ITERATIVE PROCEDURE FOR OBTAINING MAXIMUM LIKELIHOOD ESTIMATES IN A MIXTURE MODEL

Richard A. Redner

NASA/National Research Council Associate
Johnson Space Center
Houston, Texas 77058
and
Division of Mathematical Sciences
The University of Tulsa
Tulsa, Oklahoma 74104

Report #1

Prepared For

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

September, 1980

SUMMARY

In this paper we investigate the problem of estimating the parameters for a mixture of densities from, possibly distinct, exponential families. The likelihood equations used by Hasselblad (1969) are necessary conditions for a local maximum of the likelihood function. We show that a particular repeated substitution scheme, determined by the likelihood equations, converges locally to the strongly consistent maximum likelihood estimate. This generalizes the results of Peters and Walker (1978).

Some key words: exponential families, maximum likelihood estimate, mixture densities.

1. *Introduction*

Let X be an n-dimensional random variable whose density p (with respect to some σ-finite measure) is a convex combination of densities p_i, where <u>each</u> p_i belongs to <u>some</u> exponential family, i.e.,

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

$$\alpha_i^o > 0 \qquad \sum_{i=1}^{m} \alpha_i^o = 1$$

$$p_i(x) = r_i(q_i^o)\, h_i(x) \exp \langle q_i^o, f_i(x) \rangle_i$$

and where $\langle \cdot, \cdot \rangle_i$ is an inner product on R^{n_i} defined by $\langle x, y \rangle_i = x^t \Sigma_i^{-1} y$.

If $\{x_k\}_{k=1}^{N}$ is an independent sample on R^n then a maximum likelihood estimate of $\{\alpha_i^o, q_i^o\}$ is a choice of parameters $\{\alpha_i, q_i\}_{i=1}^{m}$ which locally maximizes

$$L = \frac{1}{N} \sum_{k=1}^{N} \log p(x_k)$$

with $\{\alpha_i, q_i\}_{i=1}^{m}$ replacing $\{\alpha_i^o, q_i^o\}_{i=1}^{m}$ in the evaluation of p.

If we assume that this choice is to be made from some open neighborhood Ω_i of the true parameters q_i^o and that for each i and j, $E_{q_1}|f_{1j}| < \infty$ then a necessary condition for a local maximum is that

$$\alpha_i = \frac{1}{N} \sum_{k=1}^{N} \frac{\alpha_i p_i(x_k)}{p(x_k)}$$

$$\theta_i = \frac{1}{N} \sum_{k=1}^{N} \frac{f_i(x_k) p_i(x_k)}{p(x_k)} \Big/ \frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}$$

where $\theta_i = E_{q_1}(f_1)$.

Equations of this type will be referred to as likelihood equations and these were introduced by Hasselblad (1969) for the case that each p_1 belonged to the same exponential family. We will see that this restriction is not essential. The case that each p_1 is a multivariate normal density has a longer history and has been considered by Day (1969), Duda and Hart (1973), Peters and Walker (1978), Wolfe (1970), and others. All of these authors considered a particular repeated substitution scheme to iteratively solve the likelihood equations.

2. Assumptions and a change of parameters.

At this time it is necessary to change the way each family is parameterized. The following lemma will provide some insight into this change. The lemma is essentially a rearrangement of some ideas presented in Berk (1972) and Barndorff-Nielsen (1978) and is outlined below. Throughout this paper "∇" will denote the Fréchet derivative of

a vector valued function of a vector variable. For questions concerning Fréchet derivatives, see Luenberger (1969).

Lemma 1 Let $p_o(x,q) = \hat{r}(q)h(x) \exp \langle q,f \rangle_o$ for $q \epsilon \Omega_o$ an open subset of R^{n_o}. If $p_o(x,q) = p_o(x,\hat{q})$ a.s. implies that $\hat{q} = q$, then $\theta(q) \equiv E_q(f)$ is a 1-1 function. We also have that $\theta(\Omega_o)$ is an open subset of R^{n_o} and $q(\theta)$ is a continuously differentiable function with $\nabla_\theta q$ nonsingular.

Proof In Chapter 8 of Barndorff-Nielsen (1978) we have that $\xi(q)$ is 1-1 and infinitely differentiable Since $\theta(q)$ is continuous, it follows from the Brouwer invariance of domain theorem see Dugunji page 358 (1966)) that $\theta(\Omega_o)$ is open. We also have that

$$\nabla_q \theta = \nabla_q E_\theta(f) = \{\int (f-\theta)(f-\theta)^t p_\theta\} \Sigma^{-1}.$$

Since $\theta(\Omega)$ is open and $E_q(f) = \theta$ it follows that $\nabla_q \theta$ is nonsingular. The final conclusion of the lemma follows from the inverse function theorem.

Throughout the rest of this paper we will make the following assumptions.

1) $p_1(x,q_1)$ is defined for each $q_1 \in \Omega_1$ an open subset of R^{n_i} containing q_1^o and q_1 is uniquely determined by $p_i(x, q_1)$.

2) If S is a proper subspace of R^t, $t = m + \sum_{i=1}^{m} n_1$, then

$$\text{Prob}\left\{\left(\begin{array}{c} p_1(x) \\ \vdots \\ p_m(x) \\ p_1(x)\{f_1(x) - \theta_1\} \\ \vdots \\ p_m(x)\{f_m(x) - \theta_m\} \end{array}\right) \in S\right\} < 1$$

where the probability and functional evaluation are taken with respect to $\{\alpha_i^o, \theta_i^o\}_{i=1}^m$.

We note that this assumption is a generalization of identifiability (see Yakowitz and Spragins (1968) and Teicher (1963)). That this is a nontrivial change can be seen in the following example.

Example Let $p_1(x) = \tau e^{-x\tau}$ and $p_2(x) = \tau^2 x e^{-x\tau}$. Clearly p_1 and p_2 are identifiable. We now observe that

$$p_1(f_1 - \theta_1) = p_1(x - \tfrac{1}{\tau})$$

and so

$$p_1(f_1 - \theta_1) + \frac{1}{\tau} p_2 - \frac{1}{\tau} p_1 = 0.$$

By defining $\theta_1 = E_{q_1}(f_1)$ and using lemma 1 we can proceed to the new parameterization of p_1, i.e ,

$$p_1(x,\theta_1) = r_1(\theta_1) h_1(x) \exp \langle q_1(\theta_i), f_1(x) \rangle_i.$$

This change in parameters does not change the necessary conditions for a local maximum of L.

We now consider a statistical property of solutions to the likelihood equations. The following lemma is a consequence of the fact that the conditions of Chanda (1954) are satisfied by $p(x)$ and is offered without proof. The reader is referred to Peters and Walker (1978) for further discussion.

Lemma 2 Given any sufficiently small neighborhood of the true parameters, with probability one as N approaches infinity, there is a unique solution to the likelihood equations in that neighborhood and this solution is a maximum likelihood estimate.

This solution is called the strongly consistent maximum likelihood estimate.

3. THE GENERAL ITERATIVE PROCEDURE

A natural iterative procedure for solving the likelihood equations is suggested by their fixed point form. We generate a sequence of estimates by repeatedly substituting the last estimate into the right hand side of the likelihood equations. This generates a new estimate. Hasselblad (1969) and Day (1969) have shown many examples where this work. Peters and Walker (1978) have proven that if each p_i is a multivariate normal density, then this procedure converges locally to the strongly consistent maximum likelihood estimate. Our proof of the local convergence for exponential families generalizes this result and the proof is patterned after their argument. Before we proceed further it will be helpful to introduce some notation

Since θ_i ranges over $\theta_i(\Omega_i)$ an open subset of R^{n_1}, the natural parameter space is a subset of

$$R^t = R^m \oplus R^{n_1} \oplus \ldots \oplus R^{n_m}$$

where $t = m + \sum_{i=1}^{m} n_1$. We then have that

$$\gamma \equiv \begin{pmatrix} \alpha_m \\ \vdots \\ \alpha_m \\ \theta_1 \\ \vdots \\ \theta_m \end{pmatrix}$$

is an element of R^t. If for $i=1,\ldots, m$ we let

$$A_1(\gamma) = \frac{1}{N} \sum_{k=1}^{N} \frac{\alpha_1 p_1}{p}$$

$$M_1(\gamma) = \frac{1}{N} \sum_{k=1}^{N} \frac{f_i p_1}{p} \Big/ \frac{1}{N} \sum_{k=1}^{N} \frac{p_1}{p} \quad ,$$

then the likelihood equations become

$$2) \qquad \gamma = \begin{pmatrix} A(\gamma) \\ M(\gamma) \end{pmatrix}$$

where $A = \begin{pmatrix} A_1 \\ \cdot \\ \cdot \\ A_m \end{pmatrix}$ and $M = \begin{pmatrix} M_1 \\ \cdot \\ \cdot \\ M_m \end{pmatrix}$.

Equivalent to equation 2 is

$$\gamma = \Phi_\varepsilon(\gamma) \equiv (1-\varepsilon)\, I + \varepsilon \begin{pmatrix} A(\gamma) \\ M(\gamma) \end{pmatrix} .$$

We define the repeated substitutions scheme by

$$\gamma^{i+1} = \Phi_\varepsilon (\gamma^i) .$$

The operator Φ_ε is said to be locally contractive near a point γ if for some norm $|| \cdot ||$ on R^t there is a number $0 \leq \lambda < 1$ such that

$$|| \Phi_\varepsilon (\gamma') - \gamma|| \leq \lambda \, || \gamma' - \lambda \, ||$$

whenever γ' is sufficiently close to γ.

4. LOCAL CONTRACTABILITY

We will now establish the following theorem.

Theorem 1. With probability one as N approaches infinity, Φ_ε is a locally contractive mapping (in some norm) about the strongly consistent maximum likelihood estimate whever $o < \varepsilon < 2$.

Proof. For any norm on R^t one can write

$$\Phi_\varepsilon (\gamma') - \gamma = \nabla\Phi (\gamma) [\gamma' - \gamma] + o \left(||\gamma - \gamma' ||^2 \right)$$

where γ is a solution to the likelihood equations. We can see that the theorem will be proved if one can show that with probability one, $\nabla\Phi_\varepsilon$ converges to an operator which has norm less than one.

We can write $\nabla\Phi_\varepsilon(\gamma)$ as a matrix of Frechet derivatives

$$\nabla\Phi_\varepsilon(\gamma) = (1-\varepsilon)\ I + \varepsilon \begin{pmatrix} \nabla_\alpha A & \nabla_\theta A \\ \nabla_\alpha M & \nabla_\theta M \end{pmatrix}.$$

We recall that $\nabla_{q_1}\theta_i$ is nonsingular and since

$$\nabla_{q_1}\theta_1\Sigma_i = \int (f_i-\theta_i)\ (f_i-\theta_i)^T\ p_i(\theta_i),$$

we have that $\Sigma_i^{-1}\nabla_{\theta_1} q_i$ is positive definite with respect to the usual inner product on R^{n_1}. So we define $\langle\cdot,\cdot\rangle'_i$ for i=1, ..., m by

$$\langle x, y\rangle'_i = \alpha_i x^T\ \Sigma_i^{-1}\nabla_{\theta_i} q_1\ y$$

and let $b_i = p_1/p$.

By direct calculation, using the likelihood equations, we see that if γ is the strongly consistent maximum likelihood estimate then

$$\nabla_\alpha A(\gamma) = I - (\text{diag } \alpha_i)\, \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} b_1(x_k) \\ \vdots \\ b_m(x_k) \end{pmatrix} \begin{pmatrix} b_1(x_k) \\ \vdots \\ b_m(x_k) \end{pmatrix}^T$$

$$\nabla_\theta A(\gamma) = -\frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} b_1(x_k) \\ \vdots \\ b_m(x_k) \end{pmatrix} \begin{pmatrix} \langle b_1(x_k)\{f_1(x_k) - \theta_1\}, \cdot \rangle_1 \\ \\ \langle b_m(x_k)\{f_m(x_k) - \theta_m\}, \cdot \rangle_m \end{pmatrix}^T$$

$$\nabla_\alpha M(\gamma) = -\text{diag } \alpha_i \, \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} b_1(x_k)\{f_1(x_k) - \theta_1\} \\ \vdots \\ b_m(x_k)\{f_m(x_k) - \theta_m\} \end{pmatrix} \begin{pmatrix} b_1(x_k) \\ \vdots \\ b_m(x_k) \end{pmatrix}^T$$

$$\nabla_\theta M(\gamma) = \left(\text{diag } \frac{1}{N} \sum_{K=1}^{N} \frac{f_i(x_k)}{p(x_k)} \nabla_{\theta_1} p_1(x_k) \right)$$

$$- \frac{1}{N} \sum_{k=1}^{N} \begin{pmatrix} b_1(x_k)\{f_1(x_k) - \theta_1\} \\ \vdots \\ b_m(x_k)\{f_m(x_k) - \theta_m\} \end{pmatrix} \begin{pmatrix} \langle b_1(x_k)\{f_1(x_k) - \theta_1\}, \cdot \rangle_1 \\ \\ \langle b_m(x_k)\{f_m(x_k) - \theta_m\}, \cdot \rangle_m \end{pmatrix}^T$$

We observe that $\nabla \Phi_\varepsilon(\gamma)$ can be written as

$$\nabla \Phi_\varepsilon(\gamma) = \frac{1}{N} \sum_{k=1}^{N} F(x_k, \gamma)$$

where $\nabla_\gamma F(x, \gamma)$ exists and has the property that for any norm $||\cdot||$ on $\nabla_\gamma F(x, \gamma)$ there exists a real valued function g such that

$$|| \nabla_\gamma F(x, \gamma) || \leq g(x)$$

and $\int g(x)\, p(x, \gamma^o) < \infty$

for every γ in some neighborhood of γ^o. It follows from this that $\nabla_\varepsilon \Phi$ evaluated at the maximum likelihood estimate converges to $E\{\nabla_\gamma \Phi_\varepsilon(\gamma^o)\}$. Hence it will suffice to show that in some norm $||\cdot||$, $E\{\nabla_\gamma \Phi_\varepsilon(\gamma^o)\}$ has norm less than one.

Let

$$V(x) = \begin{pmatrix} b_1(x) \\ \vdots \\ b_m(x) \\ b_1(x)\{f_1(x) - \theta_1\} \\ \vdots \\ b_m(x)\{f_m(x) - \theta_m\} \end{pmatrix}$$

and let $\langle \cdot\,, \cdot \rangle$ denote the inner product induced on R^t by scalar multiplication and $\langle \cdot\,, \cdot \rangle_i$ $i=1, \ldots, m$.

Since

$$E\left\{\frac{1}{N}\sum_{k=1}^{N} \frac{f_i(x_k)}{p(x_k)} \nabla_{\theta_i} p_i(x_k)\right\}(\gamma^o)$$

$$= \nabla_{\theta_i} \theta_i = I$$

have that

$$E\{\nabla \Phi_\varepsilon(\gamma^o)\} = I - \varepsilon \begin{pmatrix} \operatorname{diag} \alpha_i & 0 \\ 0 & I \end{pmatrix} \int V \langle V, \cdot \rangle \, p \cdot$$

We can denote this as $I - \varepsilon QR$ where

$$Q = \begin{pmatrix} \text{diag } \alpha_i & 0 \\ 0 & I \end{pmatrix}$$

and $R = \int V \langle V, \cdot \rangle p$. By assumption 2 we have that QR is positive definite with respect to $\langle \cdot , Q^{-1} \cdot \rangle$. The theorem will be proved if it can be shown that for

$$W = \begin{pmatrix} y_1 \\ \vdots \\ y_m \\ z_1 \\ \vdots \\ z_m \end{pmatrix} \in R^t,$$

that $\langle W, Q^{-1}[QR]W \rangle = \langle W, RW \rangle \leq \langle W, Q^{-1}W \rangle$.

By an application of Swartzes inequality and the fact that

$$(\nabla_q \theta)\ \Sigma = \int (f-\theta)(f-\theta)^T p_o$$

we have the following.

$$\langle W, RW \rangle = \int \langle W, V \rangle \langle V, W \rangle\ p$$

$$= \int \left[\sum_{i=1}^{m} \left\{ \frac{y_i p_i}{p} + \left\langle z_1, \ (f_1 - \theta_1^o) \frac{p_1}{p} \right\rangle_i' \right\} \right]^2 p$$

$$\leq \int \sum_{i=1}^{m} \left\{ \frac{y_1}{\alpha_1^o} + \frac{1}{\alpha_i^o} \left\langle z_i, \ (f_1 - \theta_i^o) \right\rangle_1' \right\}^2 \frac{\alpha_1^o p_1}{p} \ p$$

$$= \sum_{i=1}^{m} \left[\frac{y_i^2}{\alpha_i^o} + \frac{1}{\alpha_1^o} z^T \alpha_i^o \Sigma_i^{-1} \nabla_{\theta_i} q_1^o \left\{ E(ff^T) - \theta^o \theta^{oT} \right\} \Sigma_1^{-1} \nabla_{\theta_i} q_i^o \ \alpha_i^o z_i \right]$$

$$= \sum_{i=1}^{m} \left[\frac{y_i^2}{\alpha_i^o} + z_i^T \alpha_1^o \Sigma_i^{-1} \nabla_{\theta_1} q_i^o z_i \right]$$

$$= \langle W, Q^{-1} W \rangle \ .$$

This completes the proof.

We now consider a useful generalization of this theorem. Consider the case that the random variable X is a mixture of densities p_i, i=1, ..., m+k for k>0, where each p_i is from some exponential family for i=1,...m and where p_1 is an arbitrary but completely determined density for i = m+1,..., m+k. The appropriate likelihood equations are

$$\alpha_i = \frac{1}{N} \sum_{k=1}^{N} \frac{\alpha_i p_i(x_k)}{p(x_k)} \qquad i = 1,\ldots, m+k$$

$$\theta_1 = \frac{1}{N} \sum_{k=1}^{N} \frac{f_i(x_k) p_i(x_k)}{p(x_k)} \Big/ \frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)} \qquad i=1,\ldots,m.$$

Let $\hat{\Phi}_\varepsilon$ be the appropriate operator determined by these likelihood equations. It can be seen that the proof of Theorem 1 can be easily extended to prove the following theorem.

Theorem 2 Let assumption 1 be satisfied for $i=1,\ldots,m$ and suppose that whenever S is a proper subspace of R^t, $t = m+k+\sum_{i=1}^{m} n_i$, then

$$\text{Prob}\left\{\begin{pmatrix} p_1(x) \\ \vdots \\ p_{m+k}(x) \\ p_1(x)\{f_1(x) - \theta_1\} \\ \vdots \\ p_m(x)\{f_m(x) - \theta_m\} \end{pmatrix} \in S\right\} < 1.$$

It follows that with probability one as N approaches infinity, $\hat{\zeta}_\varepsilon$ is a locally contractive mapping (in some norm) about the strongly consistent maximum likelihood estimate whenver $0 < \varepsilon < 2$.

5. DISCUSSION

We observe that Theorem 1 is sufficiently general to include most exponential families and almost arbitrary mixtures between such families. In fact, it covers mixtures between families where the associated measures are not equivalent. Theorem 1 also applies to many situations where some subset of the usual parameters are known or where the parameters are constrained.

It should also be pointed out that although Theorem 1 applies to mixtures of multivariate normals, it is not based on the traditional likelihood equations. Instead of iterating on the covariances, the procedure updates the non-central second moment. This results in a different iterative procedure, whose difference is more than cosmetic. The difference in the updated covariances is given by $(\hat{\mu}_i - \mu_i)(\hat{\mu} - \hat{\mu}_i)^T$ where $\hat{\mu}_1$ is the new estimate for the mean given μ_1. However, there seems to be no practical difference between the two schemes and one has to favor the Peters and Walker scheme since it involves the covariances directly

Finally, we observe that the remarks made by Peters and Walker (1978) concerning the optimal choice of ε are applicable to this paper and the reader is referred to their paper for a discussion of this.

BIBLIOGRAPHY

1. Barndorff-Nielson, O. (1978). Information and Exponential Families in Statistical Theory. John Wiley and Sons. New York.

2. Berk, Robert H. (1972). Consistency and assymptotic normality of MLE's for exponential models. Ann. Math. Stat. 43, 193-204.

3. Chanda, K.C. (1954). A note on the consistency and maxima of the roots of the likelihood equations. Biometrika 41, 56-61.

4. Day, N.E. (1969).Estimating the components of a mixture of normal distributions 56, 463-474.

5. Duda, R.O. and Hart, P.E. (1973). Pattern Classification and Scene Analysis. John Wiley and Sons, New York.

6. Dugunji, James (1966). Topology. Allyn and Bacon, Inc. Boston, Massachusetts.

7. Hasselblad, V.A. (1966). Estimation of parameters for a mixture of normal distributions. Technometrics 8, 431-446.

8. Hasselblad, V.A. (1969). Estimation of finite mixtures from the exponential family. J. Amer. Stat. Assoc. Dec, 1459-1471.

9. Loéve, M. (1977). Probability Theory. D. Van Nostrand Co., New York.

10. Luenberger, D.G. (1969). Optimization by Vector Space Methods. John Wiley and Sons, Inc., New York.

11. Peters, B.C. and Walker, H.F. (1978) An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions. Siam J. Appl. Math. 35, 362-378.

12. Teicher, Henry (1963). Identifiability of finite mixtures, Ann. of Math. Stat. 34, 1265-1269.

13. Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analysis. Mult. Behav. Res. 5, 329-350.

14. Yakowitz, Sidney J. and Spragins, John D. (1968). On the identifiability of finite mixtures. Ann. of Math. Stat. 39, 209-214.

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

4 Title and Subtitle	5 Report Date
Spatial Correlation in LANDSAT: An Empirical Study	November, 1980
	6 Performing Organization Code

7. Author(s)	8 Performing Organization Report No
William A. Coberly The University of Tulsa	2
9 Performing Organization Name and Address	10 Work Unit No
Division of Mathematical Sciences The University of Tulsa Tulsa, Oklahoma 74104	11 Contract or Grant No NAS-9-14689
12 Sponsoring Agency Name and Address	13 Type of Report and Period Covered
Earth Observations Division Johnson Space Center Houston, Texas 77058	Unscheduled Technical
	14 Sponsoring Agency Code

15 Supplementary Notes

Principal Investigator: L. F. Guseman, Jr.

16 Abstract

Neighboring pixels in a LANDSAT image are not statistically independent observations as is sometimes assumed in many analysis procedures. This study investigates empirically the characteristics of spatial correlation in four 1976 LACIE sample segments for which ground truth is available.

17 Key Words (Suggested by Author(s))	18 Distribution Statement
Spatial Correlation, digital image modeling, LANDSAT image analysis	

19 Security Classif (of this report)	20 Security Classif (of this page)	21 No of Pages	22 Price*
		23	

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

SPATIAL CORRELATION IN LANDSAT

AN EMPIRICAL STUDY

William A. Coberly

Division of Mathematical Sciences
The University of Tulsa
Tulsa, Oklahoma 74104

Report #2

Prepared For

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

November, 1980

SPATIAL CORRELATION IN LANDSAT

AN EMPIRICAL STUDY

William A. Coberly
The University of Tulsa

1. INTRODUCTION

Data analysts who have worked with LANDSAT data have observed that neighboring pixels are not independent measurements on disjoint areas of the target scene. This spatial correlation or dependency is induced by a number of factors - overlap of the instantaneous field of view (IFOV), atmospheric scattering, optical and electro-mechanical components of the sensor system. These factors are are in addition to any intrinsic spatial correlation which might exist in the target scene. This spatial correlation violates a number of assumptions usually made in the digital processing and analysis of LANDSAT data, especially the statistical analysis. A few studies (1, 2) have investigated its effects on the accuracy of various statistical procedures. However, a more fundamental analysis of spatial correlation is required in order to enhance our understanding of LANDSAT image representation and modelling. In particular, a better understanding of the boundary or mixed pixel

phenomenon requires the incorporation of spatial correlation into the model.

Two approaches should be undertaken. First an analytical determination of the spatial correlation induced by the atmosphere and the sensor system, based on a linear system representation of these factors should be made. The second approach is an empirical determination of the spatial correlation structure. This is the purpose of this exploratory study.

2. SPATIAL CORRELATION

A complete study should consider the two dimensional properties of spatial correlation. However, in this study only the one dimensional characteristics, in the direction of the scan line, will be studied. This is a reasonable start since a number of the factors, such as detector response and electronic amplification and recording, are one dimensional.

Define $X_1, X_2, \cdots, X_L$ to be the random digital measurements along one scan line for a single channel of the multispectral scanner. Let $m_1 \triangleq E(X_1)$ be the mean value of X_1 for $1 = 1, \cdots, L$. Then the autocovariance function is given by

$$\gamma(i,i+k) = E((X_i - m_i)(X_{i+k} - m_{i+k})).$$

We now impose the assumption of covariance stationarity, which may not hold for large scan angles, but should be a reasonable assumption for small scan angles. Now γ depends only on the lag k, and is independent of scan line position i. That is,

$$\gamma(k) \triangleq \gamma(i,i+k).$$

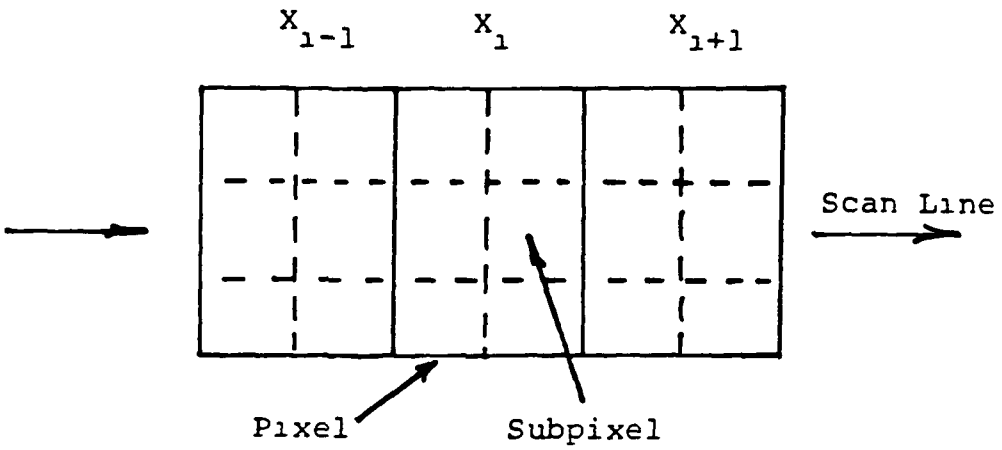
That is, we are assuming that the distribution of the pixels along a scan line is covariance stationary, changing only in mean. Note that $\gamma(0)$ is the variance and the autocorrelation (spatial correlation) is given by

$$\rho(k) = \gamma(k)/\gamma(0)$$

for $k = 0,1, \cdots$.

3. ESTIMATION OF THE MEAN

The mean function m_i is, of course, in general not known. However, for the segments used in this study, digital ground truth was available and this suggests a way to estimate the mean for each of the pixels. The digital ground truth is tabulated at the subpixel level, six subpixels per pixel according to the following scheme.



If the pixel has the same ground truth label assigned to each of the six subpixels, then it is said to be "pure". A "field" is an interval along a scan line of pure pixels with the same ground truth label. A "field" may be one pixel in width or many. Pixels which are not "pure", that is, those containing conflicting subpixel ground truth labels, will be called "boundary" pixels.

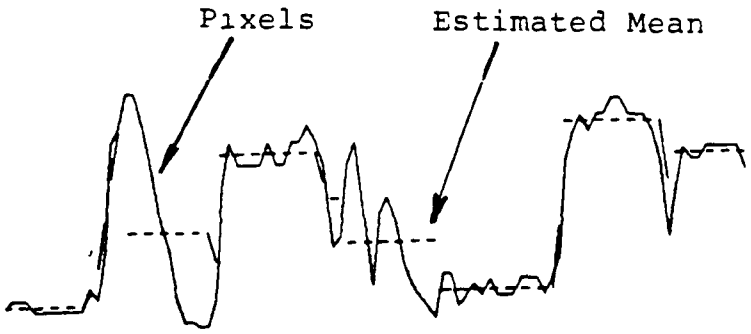
The estimate of the mean function for a scan line is defined as follows:

$$\hat{m}_i = \begin{cases} \text{field mean of } X_i \text{ if contained in a field} \\ \text{a moving average if } X_i \text{ is a boundary pixel} \end{cases}$$

The moving average used is

$$(X_{i-2} + 2X_{i-1} + 2X_i + 2X_{i+1} + X_{i+2})/8.$$

In Figures 1 - 8, the pixels X_i are plotted (solid lines) superimposed on the estimated mean function m_i (dotted line) for the four LANDSAT channels and the two tassel-cap coordinates "brightness" and "greenness". One scan line for two acquisitions of each of four segments is presented.



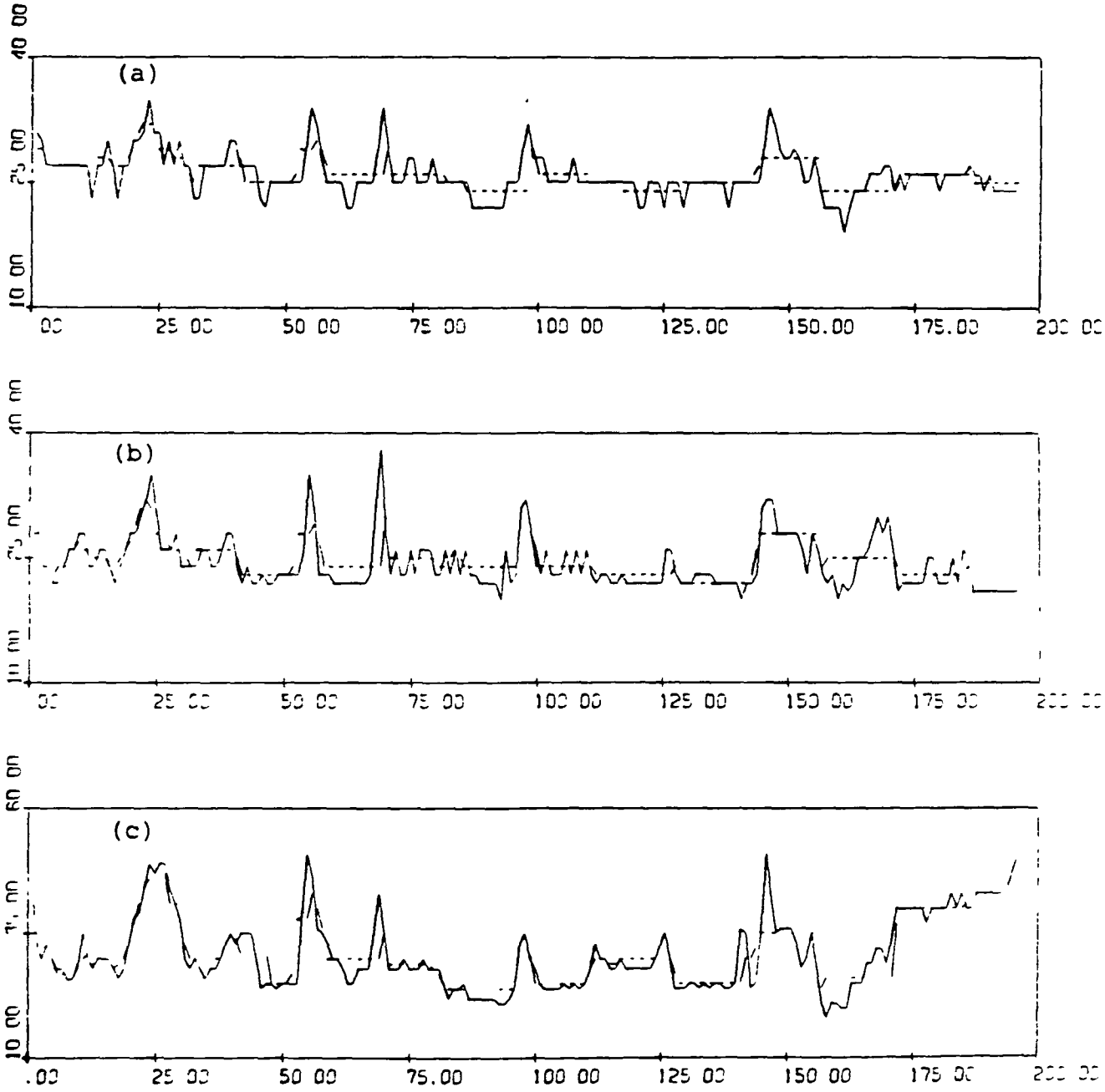


Figure 1. Pixel radiance and estimated mean plot for segment 1618/145, line 62. (a)-(d) channels 1-4, (e) brightness, (f) green coordinate.

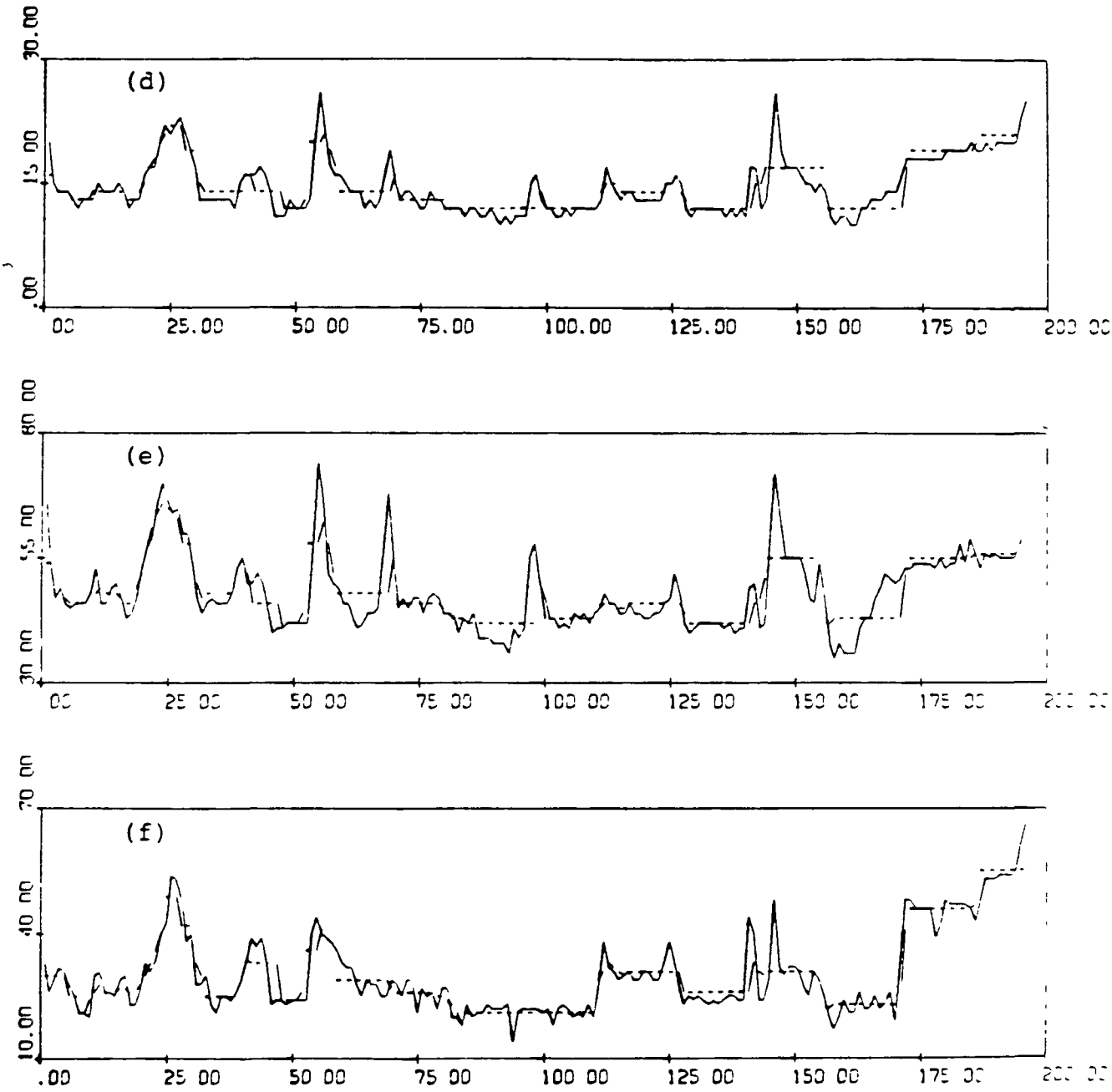


Figure 1. Continued.

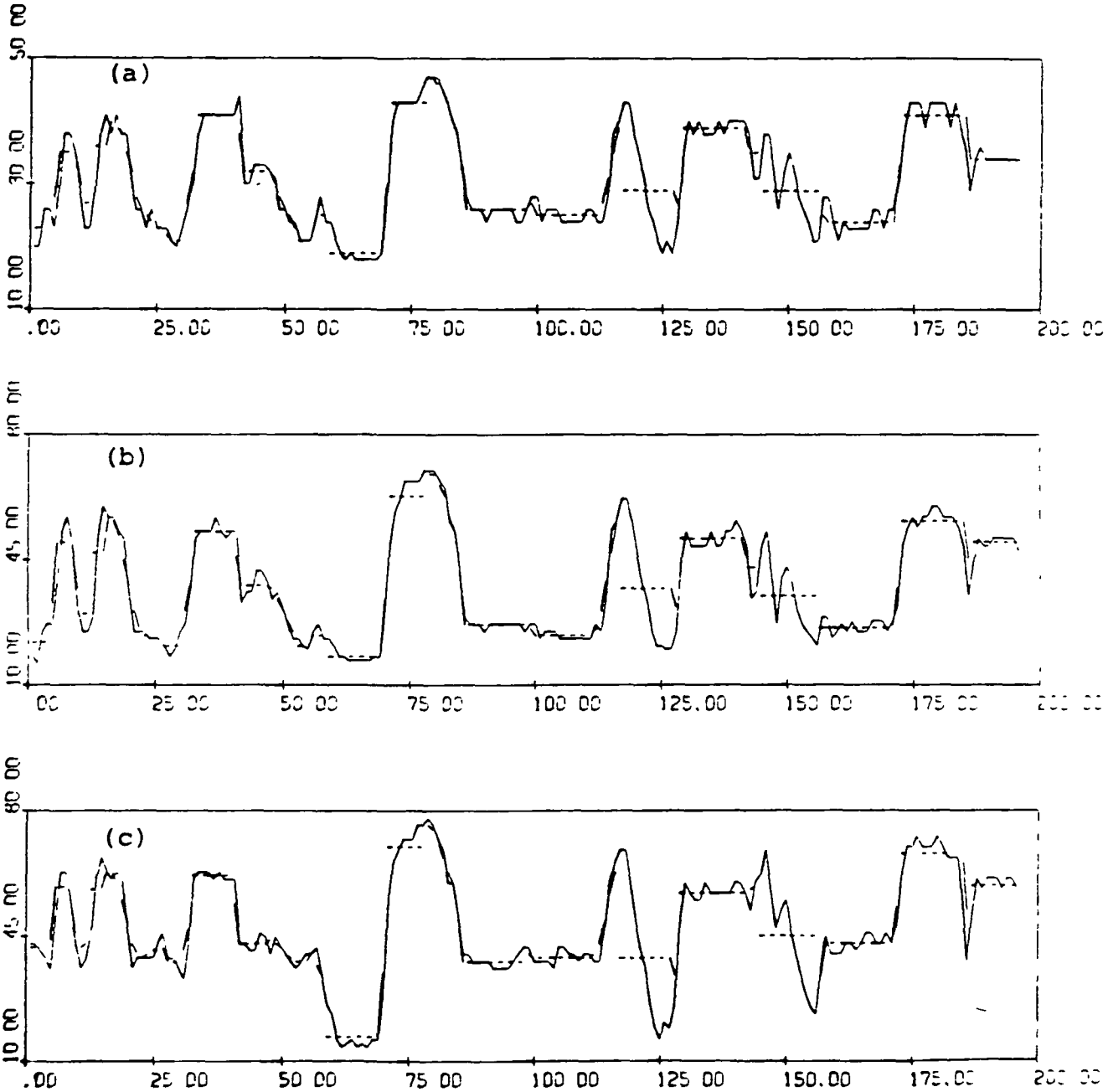


Figure 2. Pixel radiance and estimated mean plot for segment 1618/235, line 62. (a)-(d) channels 1-4, (e) brightness, (f) green coordinate.

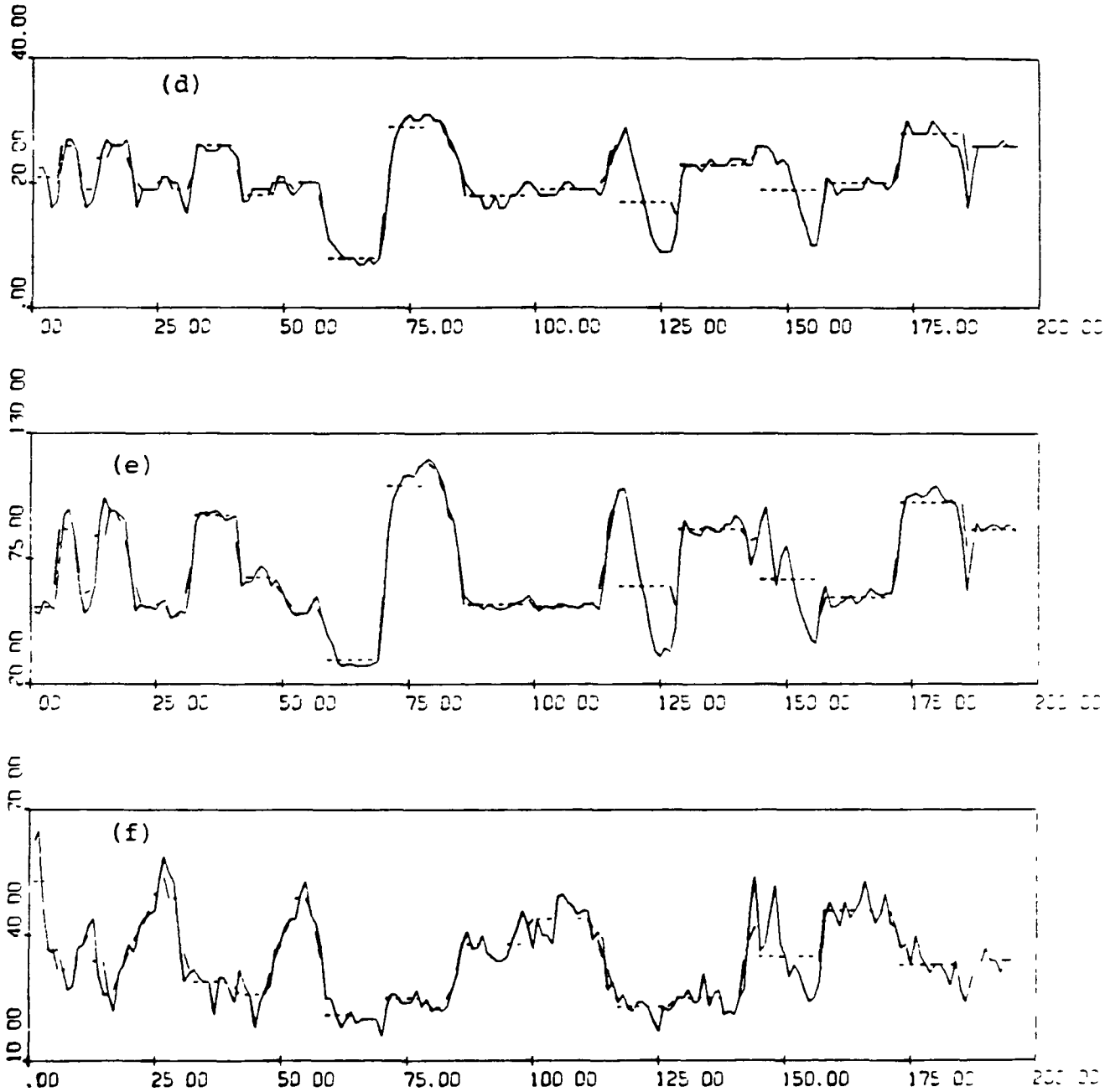


Figure 2. Continued.

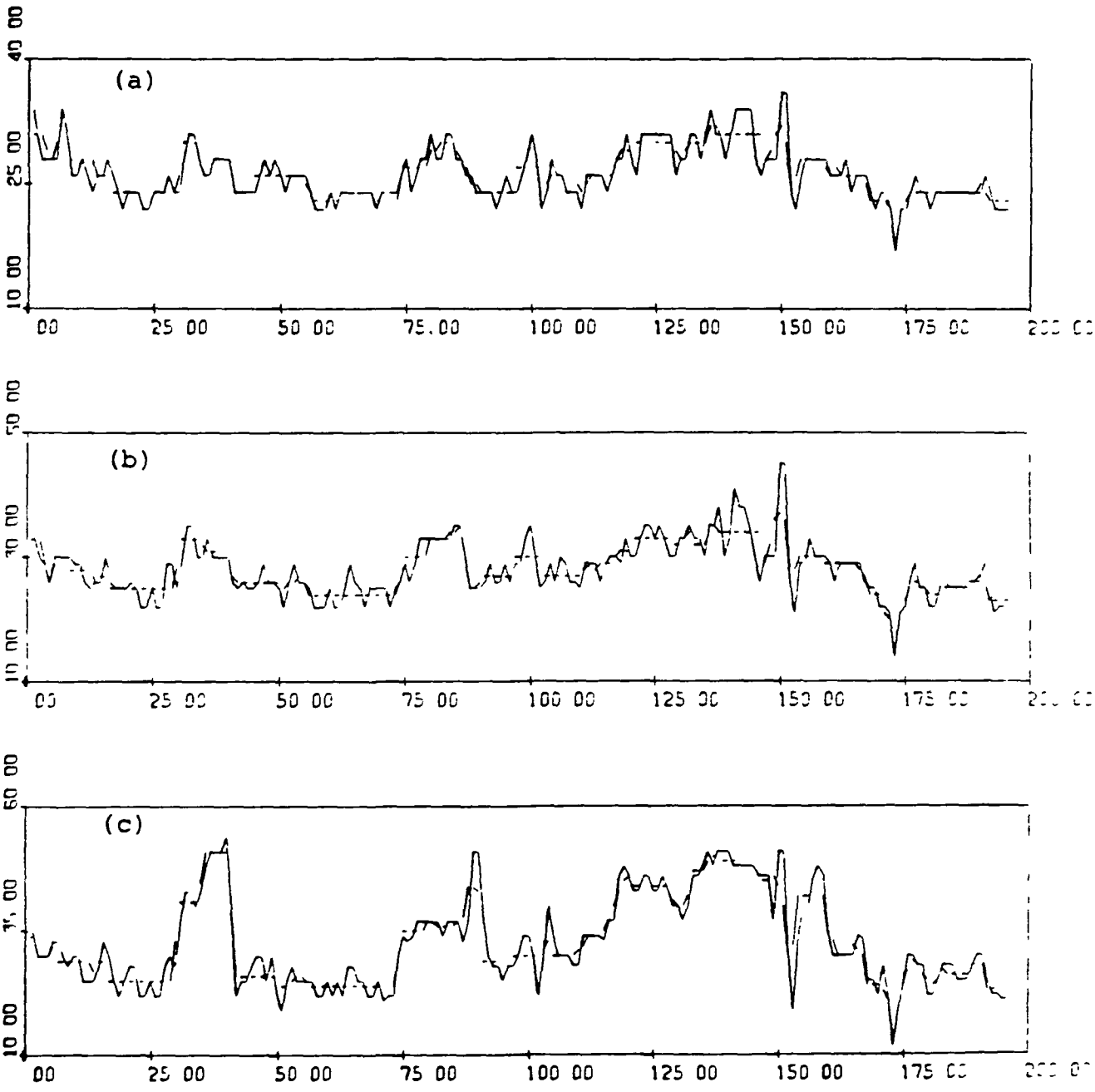


Figure 3. Pixel radiance and estimated mean plot for segment 1633/129, line 62. (a)-(d) channels 1-4, (e) brightness, (f) green coordinate.

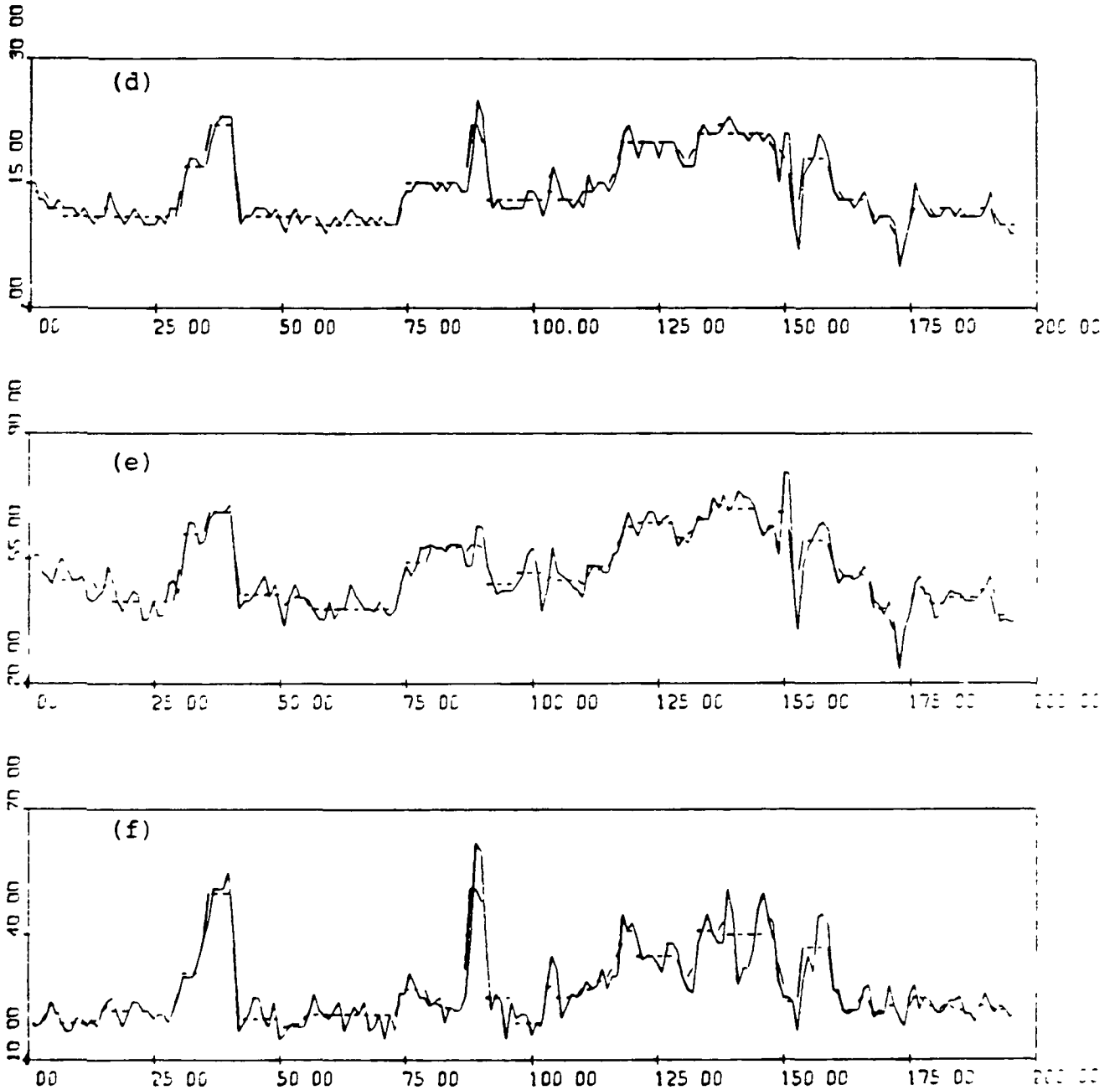


Figure 3. Continued.

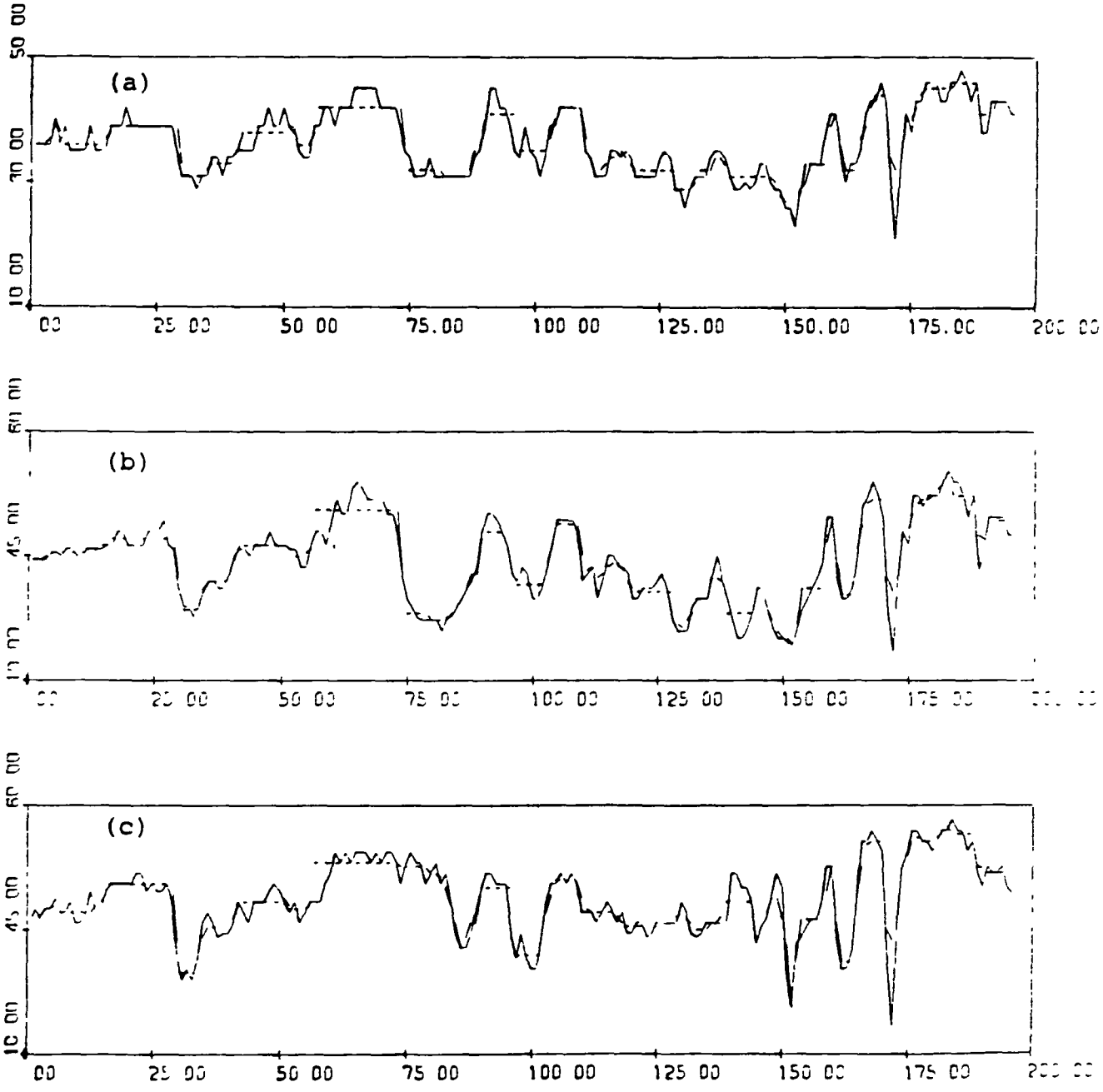


Figure 4. Pixel radiance and estimated mean plot for segment 1633/236, line 62. (a)-(d) channels 1-4, (e) brightness, (f) green coordinate.

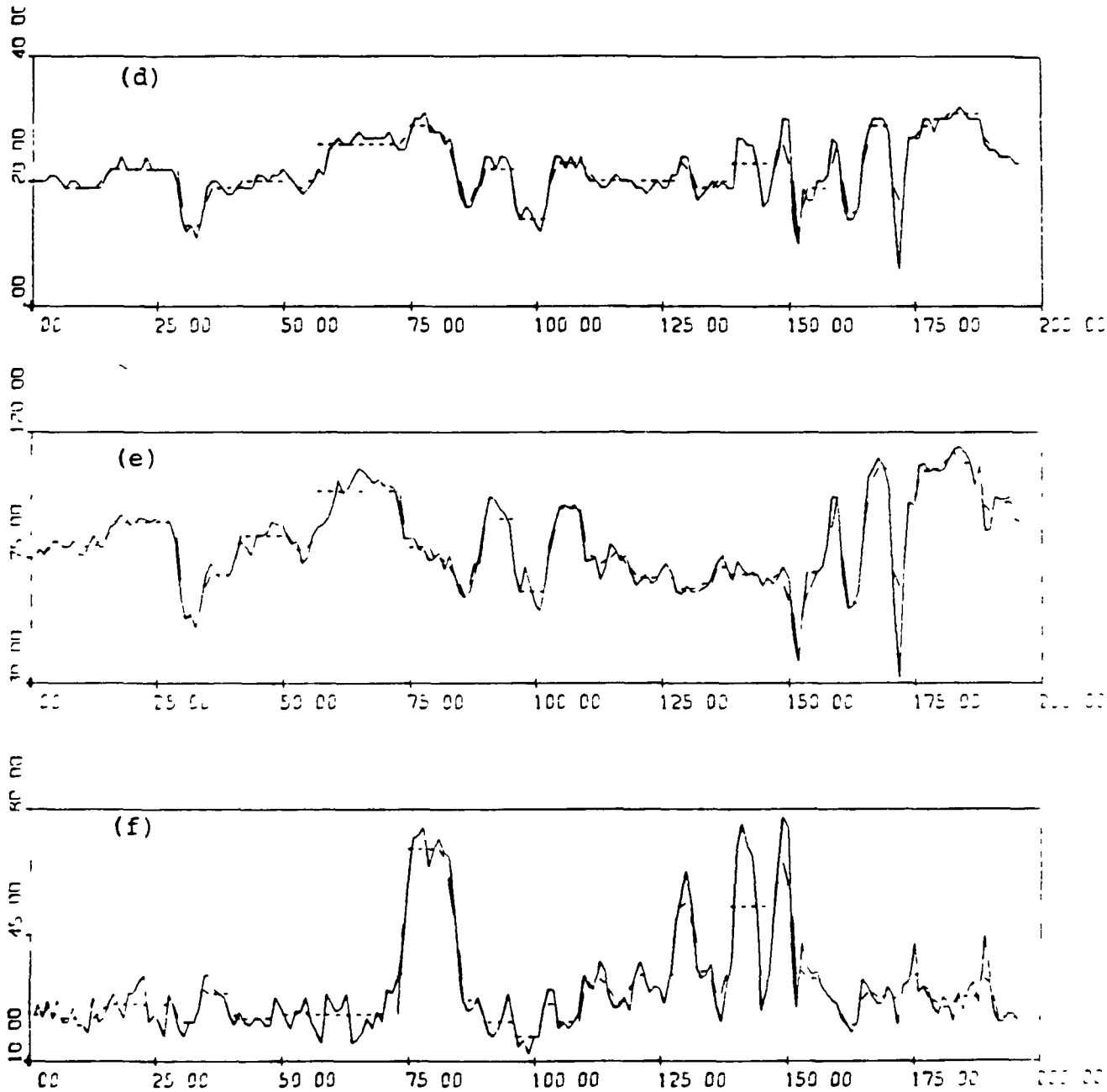


Figure 4. Continued.

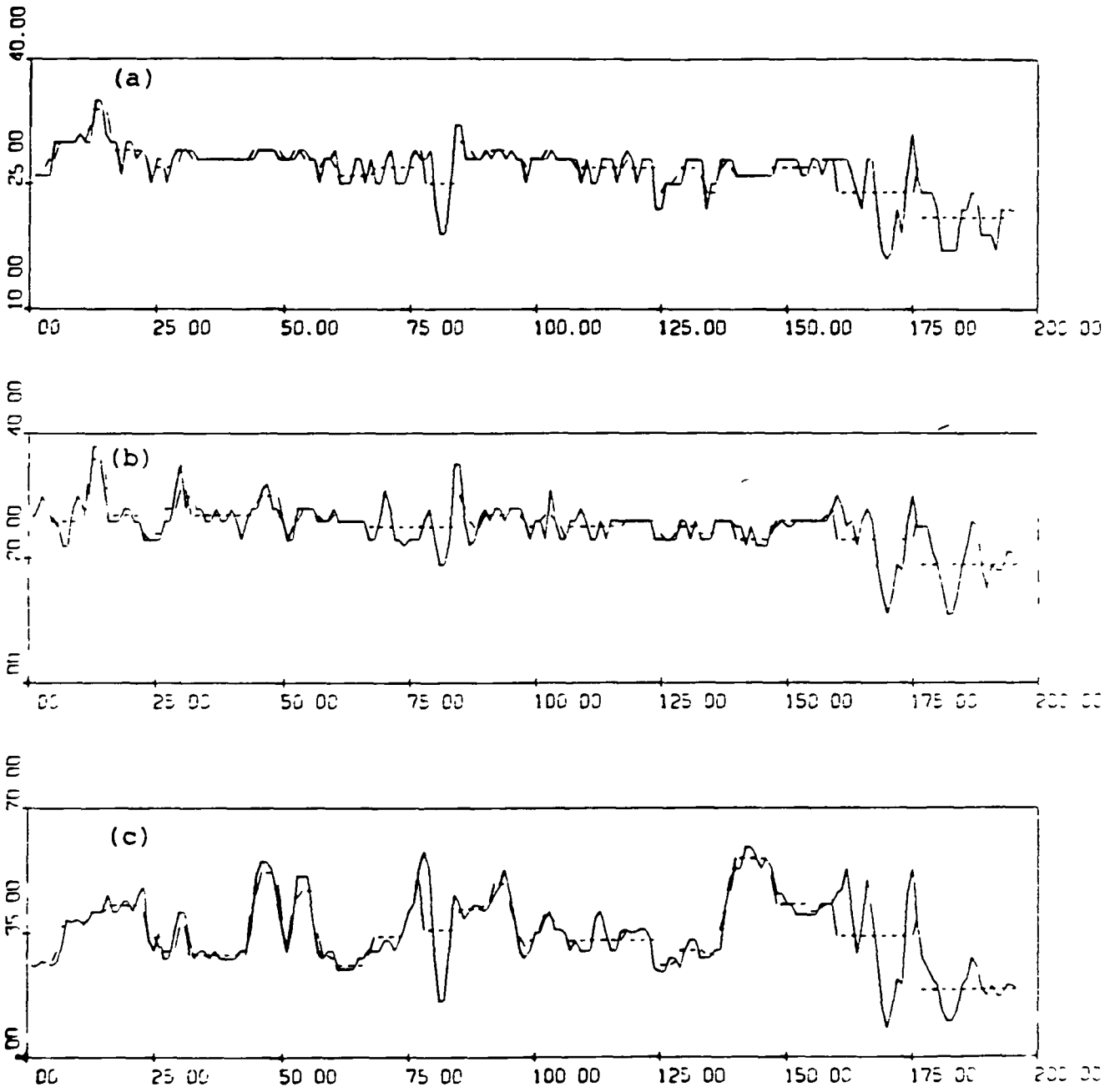


Figure 5. Pixel radiance and estimated mean plot for segment 1642/145, line 11. (a)-(d) channels 1-4, (e) brightness, (f) green coordinate.

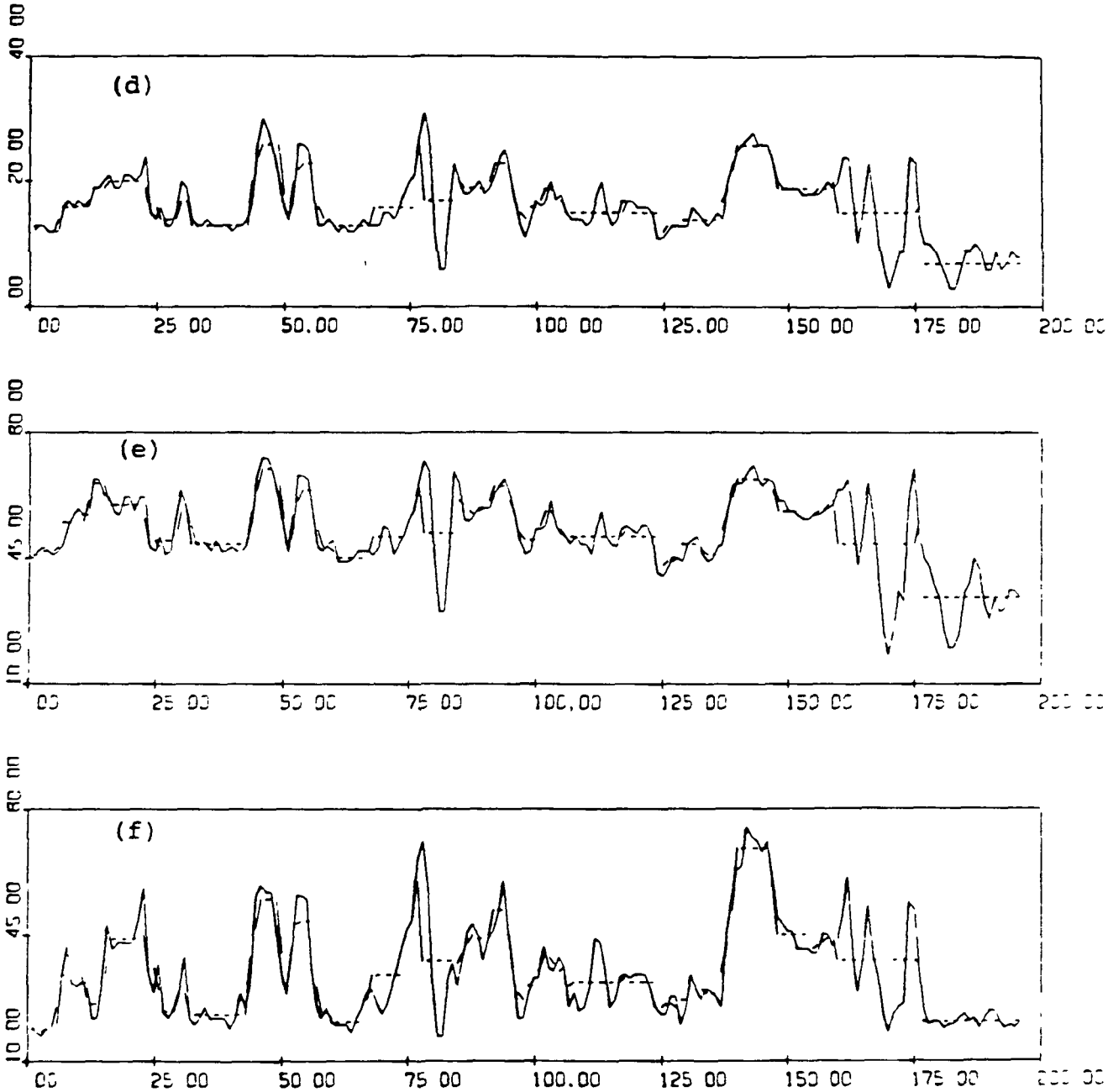


Figure 5. Continued.

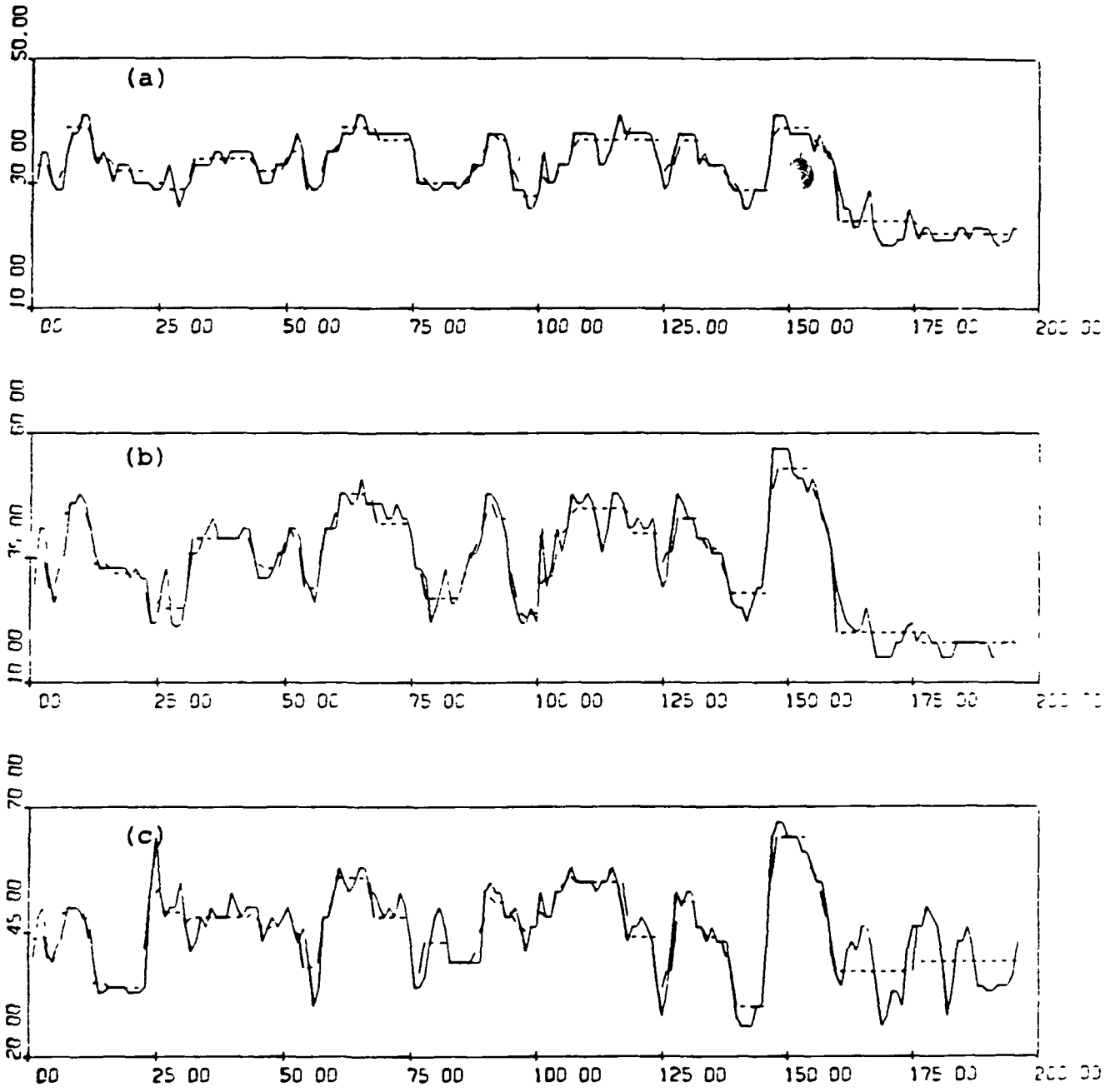


Figure 6. Pixel radiance and estimated mean plot for segment 1642/236, line 11. (a)-(d) channels 1-4, (e) brightness, (f) green coordinate.

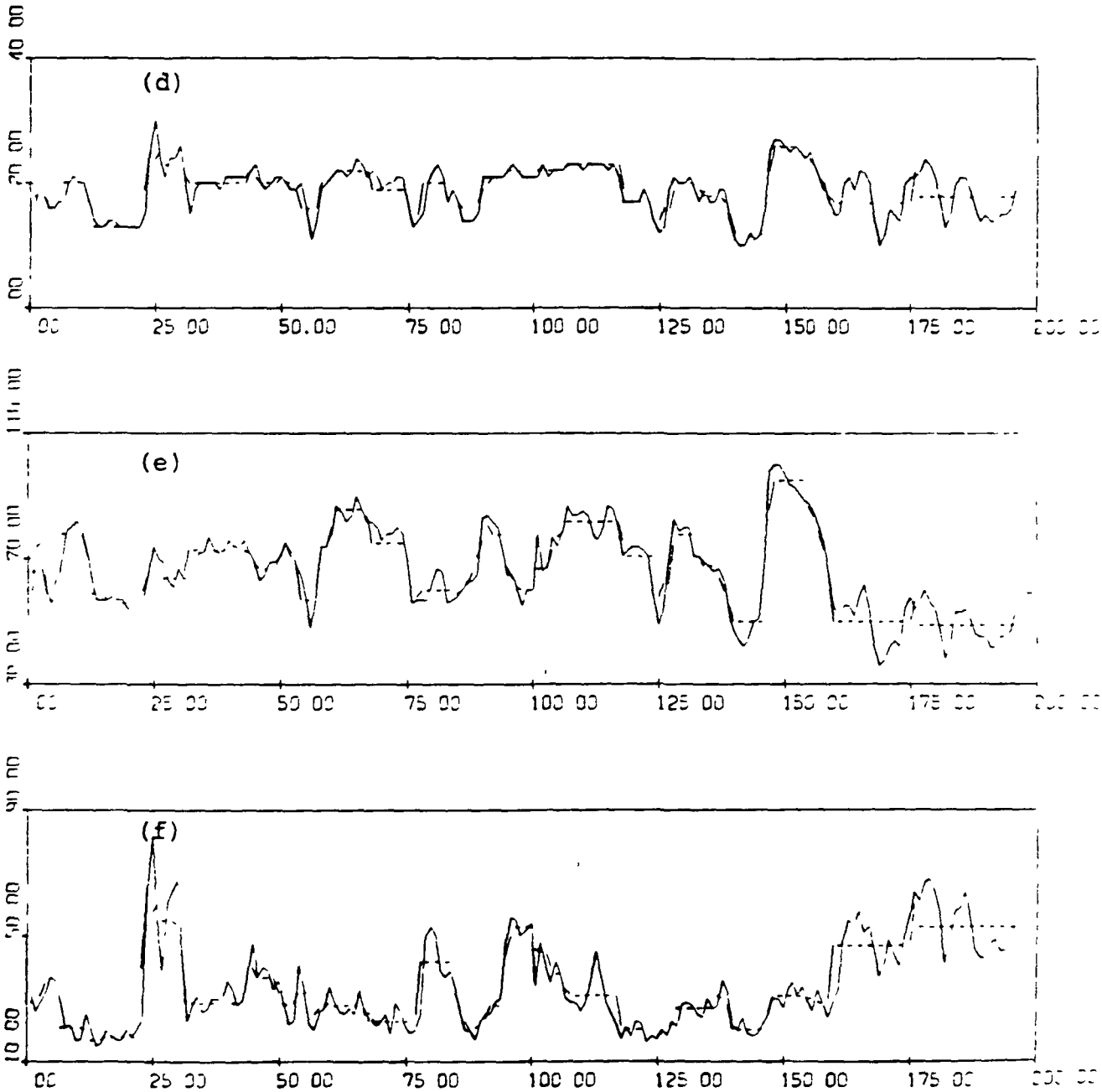


Figure 6. Continued.

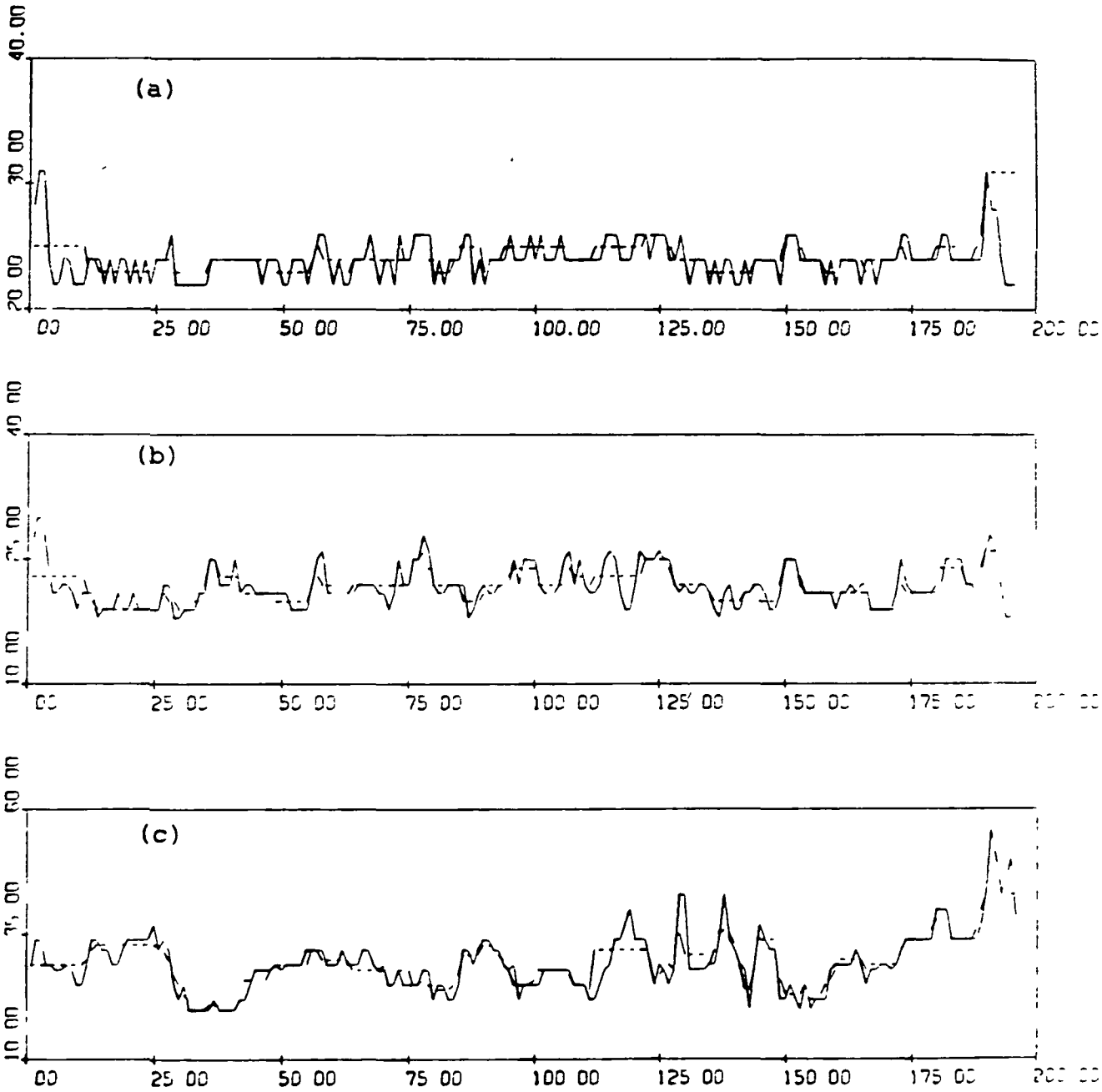


Figure 7. Pixel radiance and estimated mean plot for segment 1645/145, line 62. (a)-(d) channels 1-4, (e) brightness, (f) green coordinate.

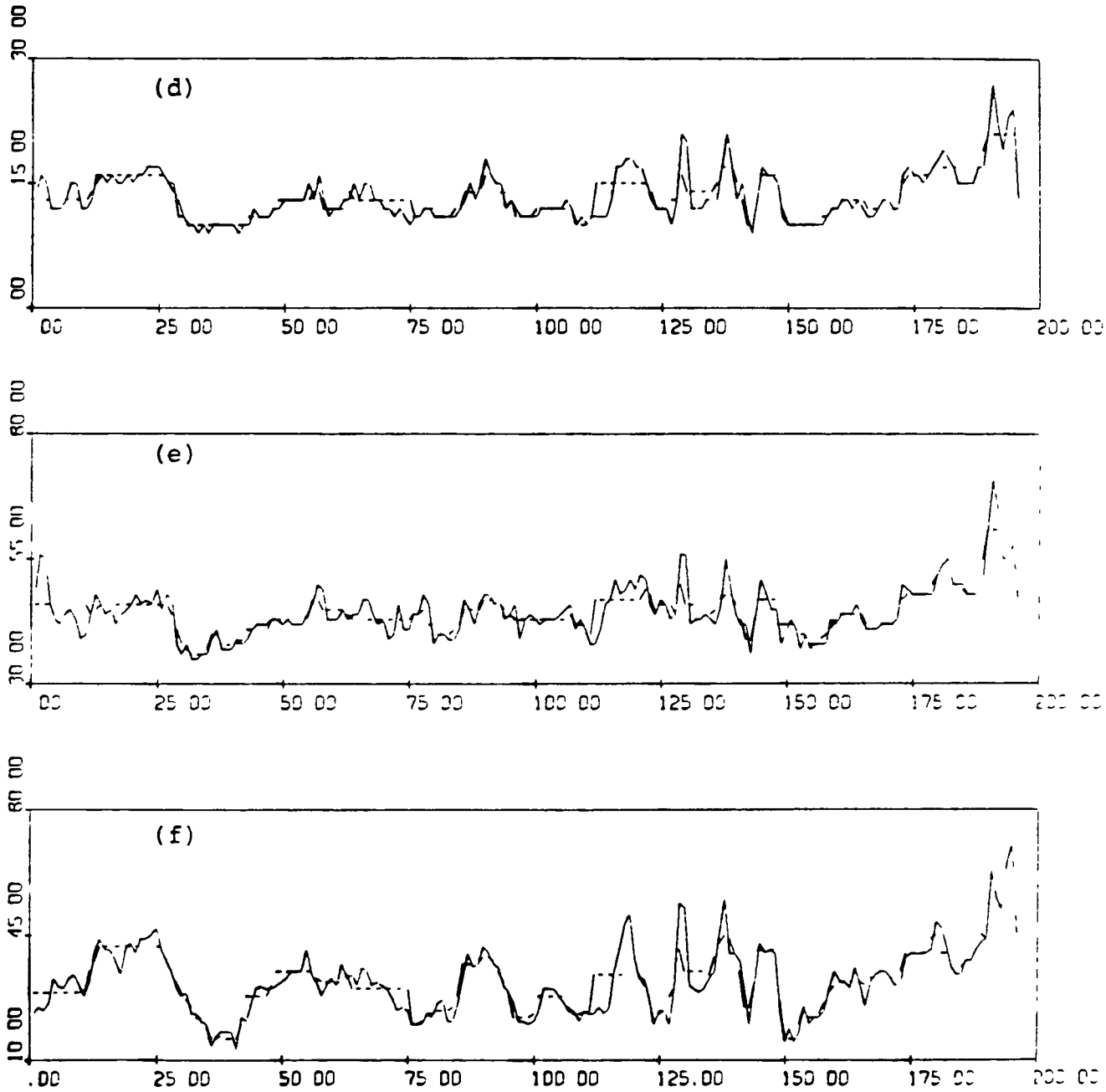


Figure 7. Continued.

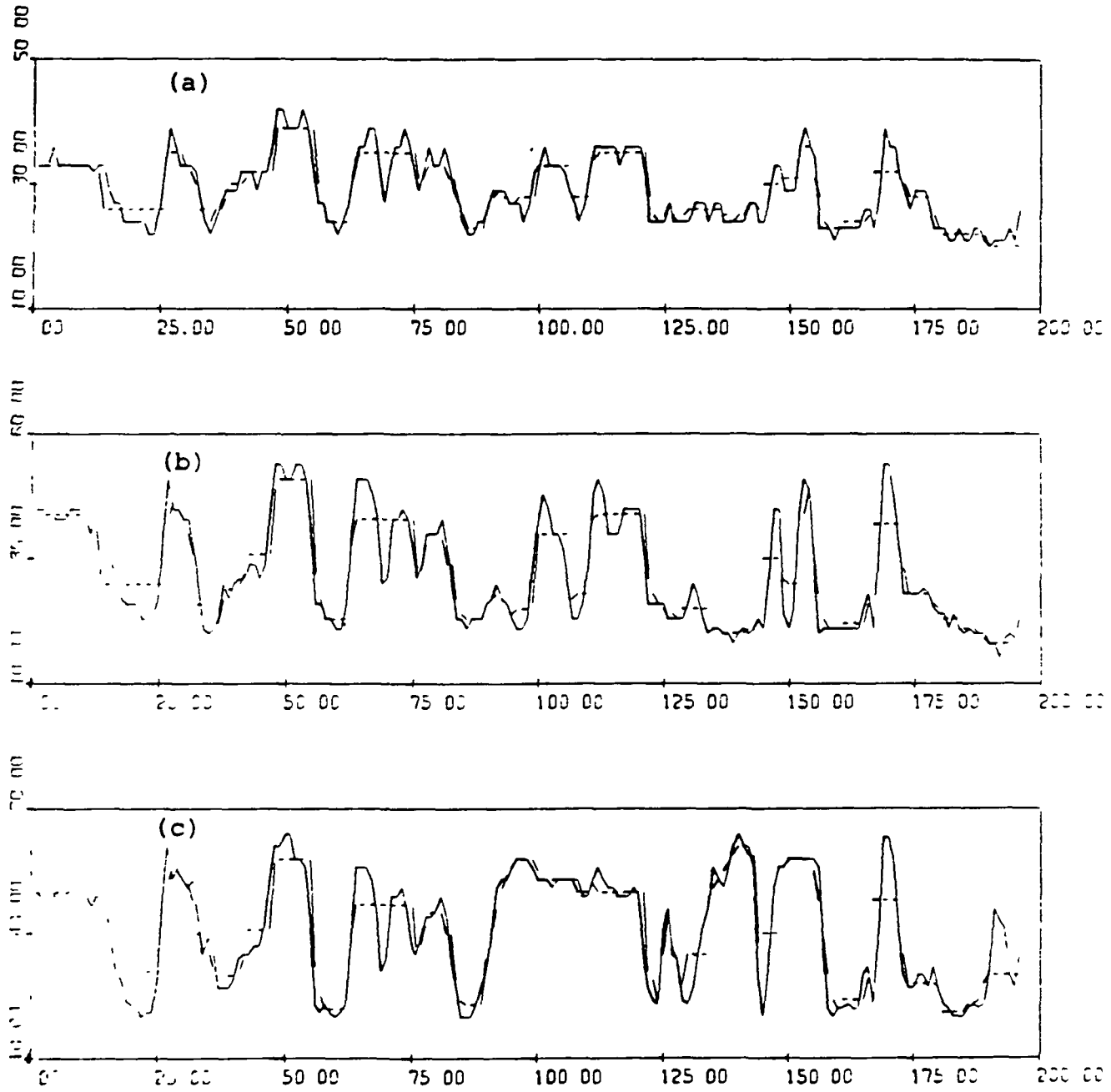


Figure 8. Pixel radiance and estimated mean plot for segment 1645/236, line 62. (a)-(d) channels 1-4, (e) brightness, (f) green coordinate.

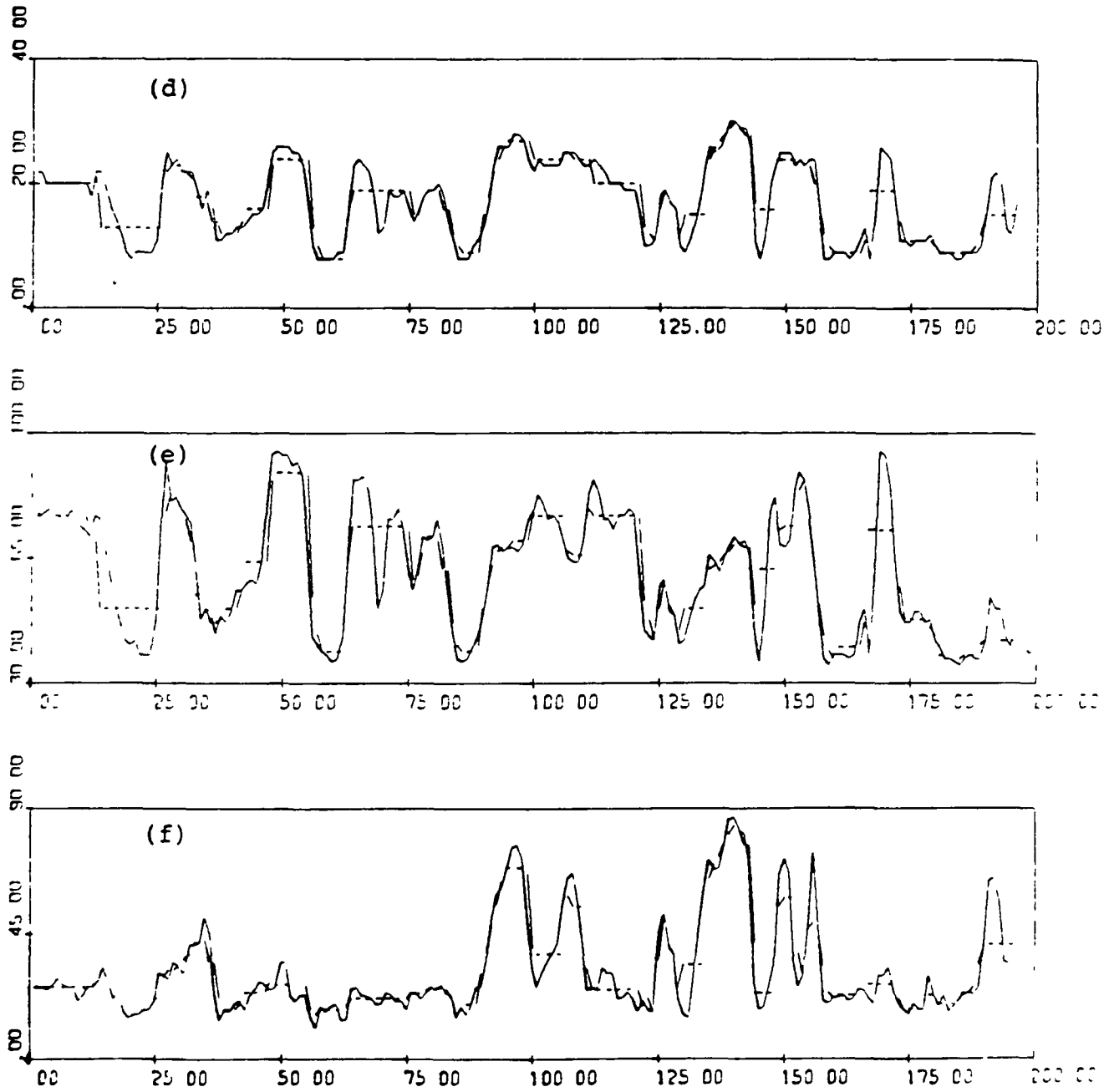


Figure 8. Continued.

4. ESTIMATING THE SPATIAL CORRELATION

For a given scan line and channel, the sample spatial correlation is calculated by

$$\hat{\gamma}(k) = \frac{1}{L} \sum_{i=1}^{L-k} (X_i - \hat{m}_1)(X_{1+k} - \hat{m}_{1+k})$$

and

$$\hat{\rho}(k) = \hat{\gamma}(k)/\hat{\gamma}(0)$$

for $k = 0,1, \cdots$. In this study the sample spatial correlation was calculated for every third scan line for each of the four channels on each segment acquisition. In Table 1 the average spatial correlation function over all scan lines used in the calculations is tabulated for two acquisitions for each of four segments. Although the coefficients are not the same from segment to segment, the pattern is very consistent. The lag 1 correlation is distinctly non-zero over all segments and channels and the lag 3 and larger order correlations are essentially zero. The lag 2 correlation is zero for some cases and non-zero for others.

In Figures 9-16, the histograms of the estimates for $\hat{\rho}(1)$ and $\hat{\rho}(2)$ and the scatter plots of $\hat{\rho}(1)$ versus $\hat{\rho}(2)$ are presented for all scan lines processed in the study.

5. BOUNDARY PIXELS AND SPATIAL CORRELATION

The spatial correlation observed has considerable implications in the characterization of boundary or mixed pixels. The usual notion of mixed pixel is one in which the instantaneous field of view intersects at least two real label classes in the target scene. In fact, spatial correlation may induce the mixed pixel effect even when the IFOV target is composed of a single class, due to the mixing of neighboring pixels by the correlating mechanism. By understanding this spatial correlation phenomenon, better automatic boundary finding or field finding algorithms, specifically developed for LANDSAT data applications, should result.

REFERENCES

(1) Tubbs, J. D. and Coberly, W. A., "Spatial correlation and its effect upon classification results in LANDSAT", Proceedings of the 12th International Symposium on Remote Sensing of Environment, Manilla, The Phillipines, April 1978.

(2) Basu, J. P. and Odell, P. L., "Effect of intraclass corelation among training samples on the misclassification probabilities of Bayes' procedure", Pattern Recognition Vol. 16, pp. 13-16. (1974)

(3) Arsts,I., Goldstein,B., Hayden,L., Kidd,R., and Miller,L. "A linear model approach for solving the sensor problem", MSC Internal Note No. 72-FM-276, NASA/Johnson Space Center, Houston, Texas (1972)

(4) Advanced Scanners and imaging systems for earth Observations, NASA SP-335 (1973)

TABLE 1. Estimated spatial correlation functions.

Segment	Chan	Lag 1	2	3	4	5	6
1618/145	1	.229	-.036	-.030	-.039	-.051	-.056
	2	.302	-.051	-.048	-.058	-.080	-.072
	3	.352	-.060	-.100	-.087	-.078	-.081
	4	.286	-.053	-.099	-.087	-.089	-.081
1618/235	1	.445	.114	-.004	-.075	-.088	-.089
	2	.501	.129	-.008	-.077	-.114	-.118
	3	.486	.097	-.034	-.081	-.096	-.090
	4	.486	.091	-.029	-.064	-.079	-.089
1633/129	1	.309	-.015	-.029	-.044	-.046	-.065
	2	.378	-.005	-.040	-.039	-.054	-.065
	3	.387	.017	-.019	-.048	-.078	-.094
	4	.421	.046	-.007	-.035	-.071	-.083
1633/236	1	.335	.032	-.023	-.042	-.048	-.061
	2	.446	.058	-.042	-.072	-.090	-.083
	3	.396	.035	-.034	-.068	-.090	-.087
	4	.432	.051	-.031	-.072	-.083	-.074

TABLE 1. Continued.

Segment		Lag 1	2	3	4	5	6
1642/145	1	.213	-.057	-.060	-.050	-.064	-.045
	2	.337	-.024	-.055	-.066	-.070	-.072
	3	.365	.003	-.023	-.054	-.077	-.079
	4	.393	.007	-.029	-.046	-.062	-.069
1642/236	1	.219	-.033	-.031	-.044	-.059	-.042
	2	.310	-.019	-.053	-.070	-.088	-.091
	3	.354	.014	-.050	-.079	-.110	-.108
	4	.406	.015	-.046	-.076	-.116	-.119
1645/145	1	.109	-.066	-.011	-.011	-.021	-.015
	2	.178	-.108	-.047	-.023	-.015	-.009
	3	.260	-.048	-.034	-.035	-.039	-.051
	4	.293	-.027	-.029	-.024	-.021	-.046
1645/236	1	.343	-.005	-.045	-.060	-.070	-.073
	2	.424	.013	-.071	-.087	-.083	-.097
	3	.426	.023	-.049	-.061	-.076	-.091
	4	.441	.033	-.043	-.058	-.068	-.081

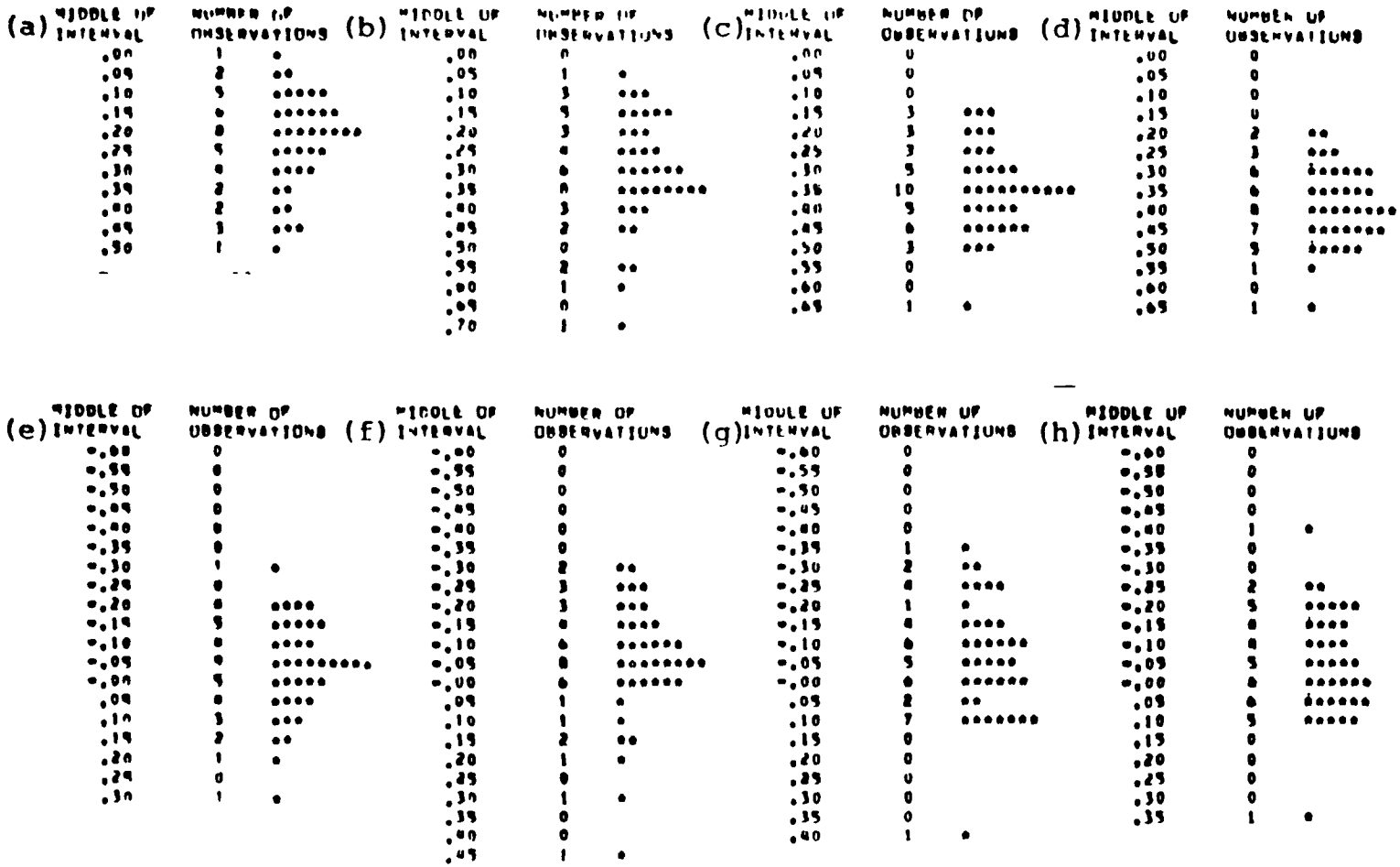


Figure 9. Histograms for 1618/145. (a)-(d) Lag 1 spatial correlations for channels 1-4. (e)-(h) Lag 2 spatial correlations for channels 1-4. Computed for every third scan line.

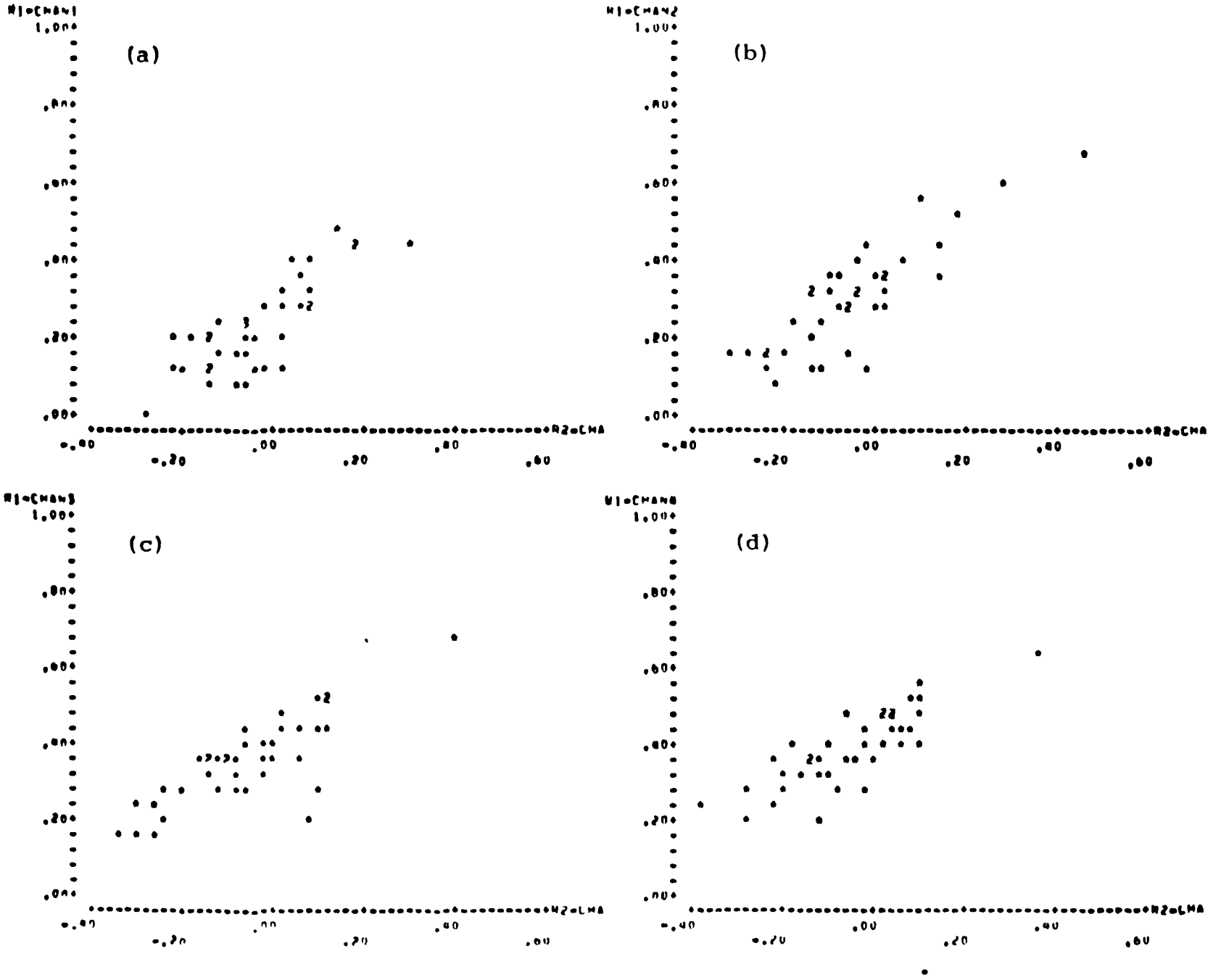


Figure 10. Scatter plots of Lag 1 spatial correlation versus Lag 2 spatial correlation for 1618/145. (a)-(d) channels 1-4. Computed for every third scan line.

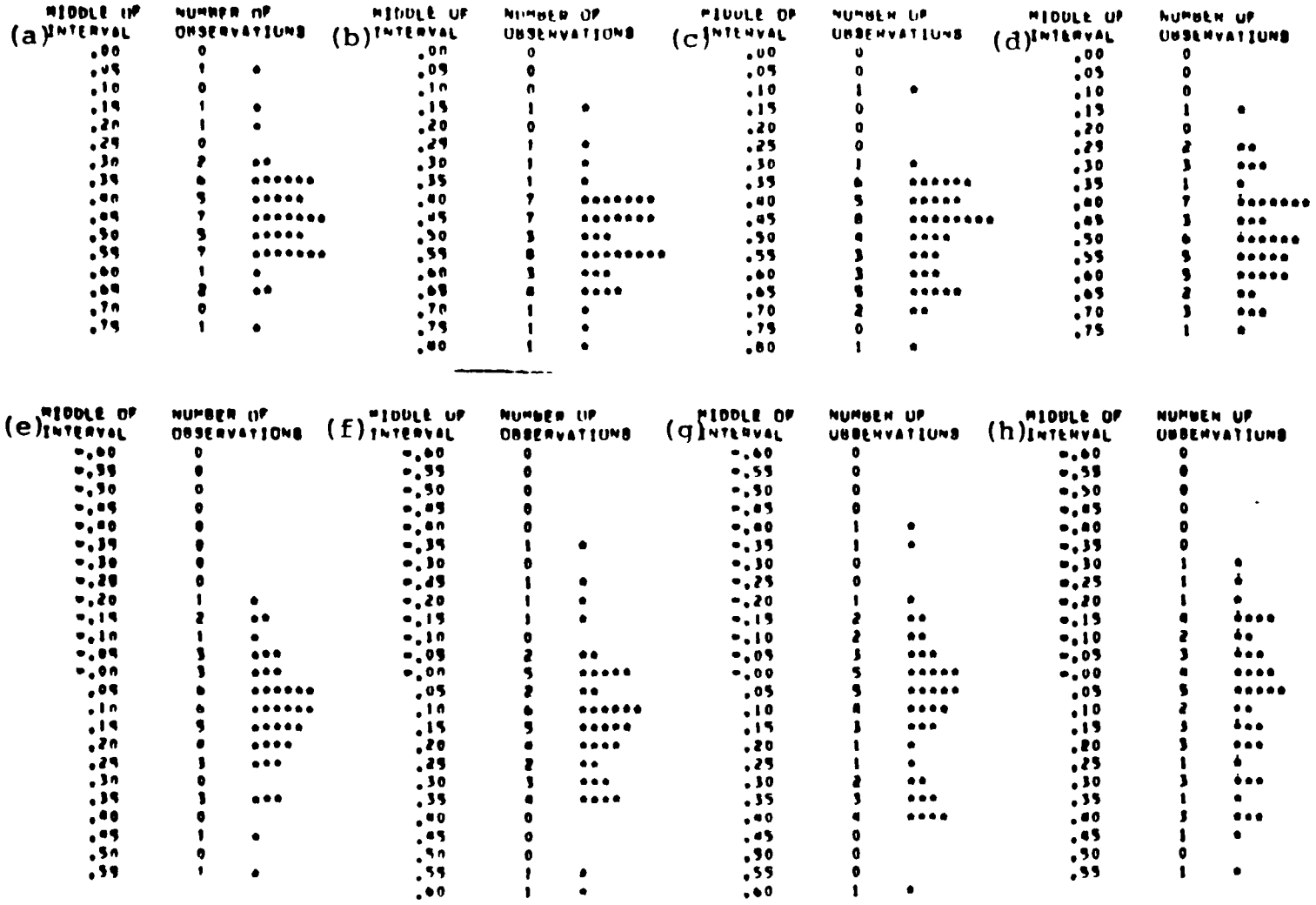


Figure 11. Histograms for 1618/235. (a)-(d) Lag 1 spatial correlations for channels 1-4. (e)-(h) Lag 2 spatial correlations for channels 1-4. Computed for every third scan line.

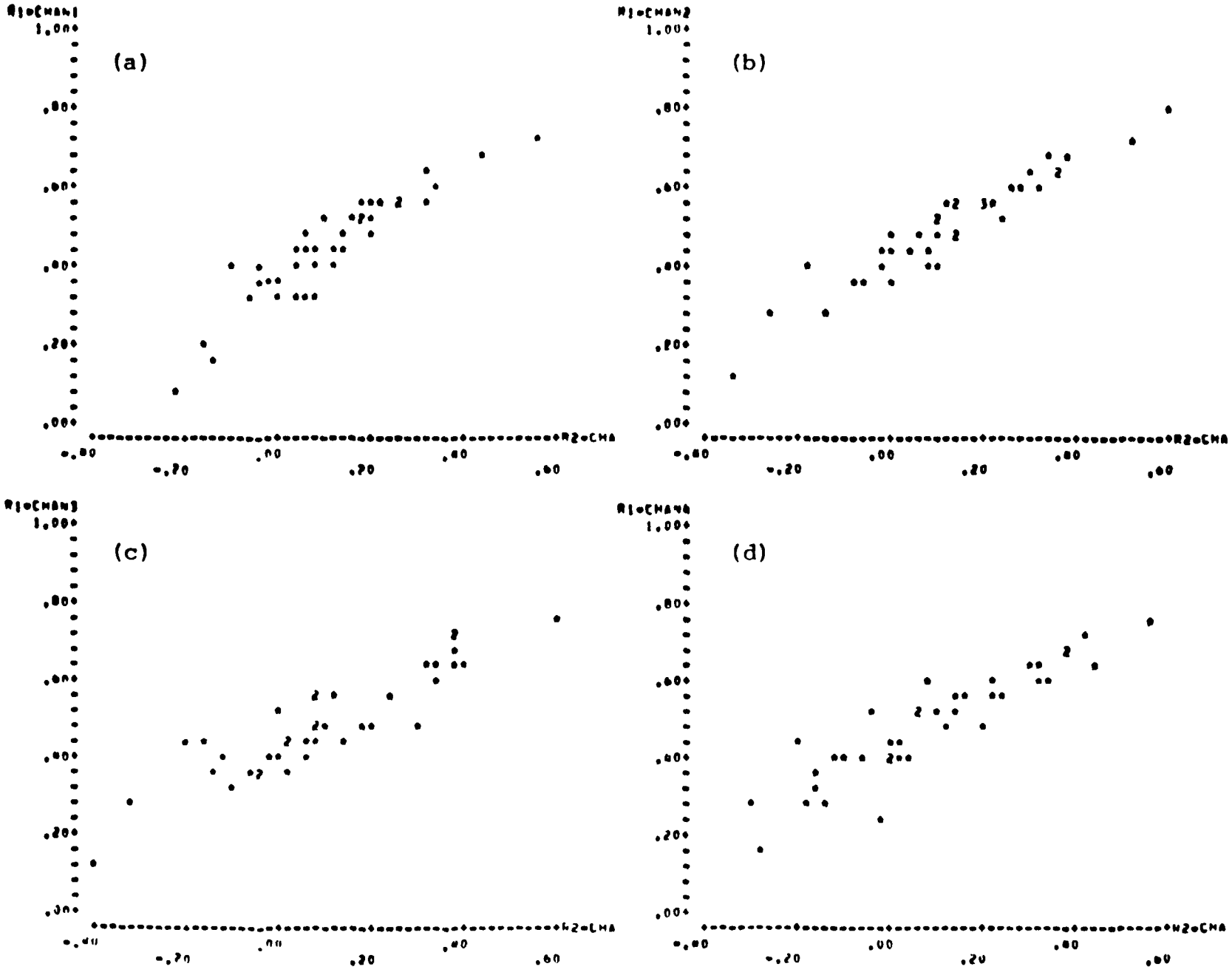


Figure 12. Scatter plots of Lag 1 spatial correlation versus Lag 2 spatial correlation for 1618/235. (a)-(d) channels 1-4. Computed for every third scan line.

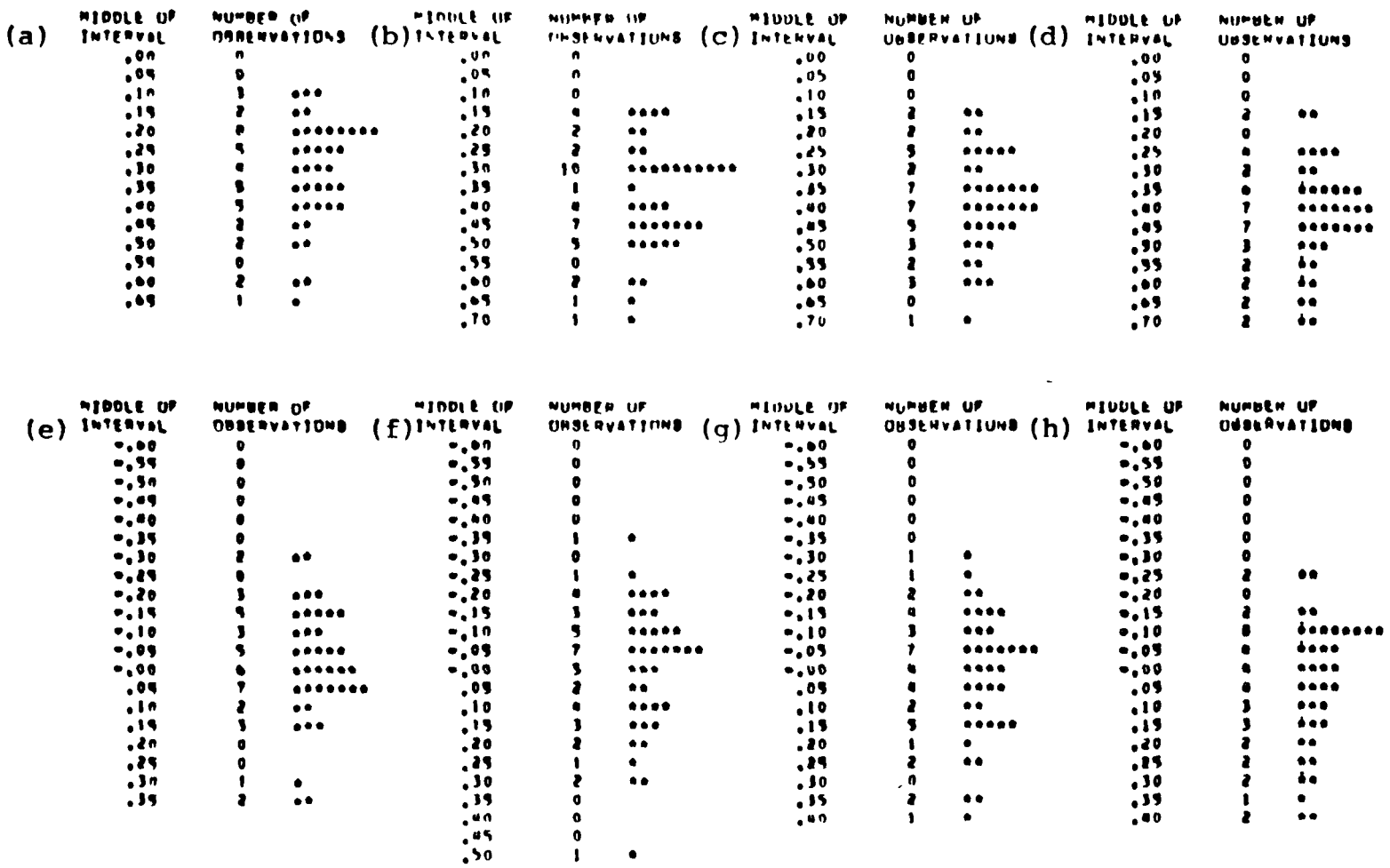


Figure 13. Histograms for 1633/129. (a)-(d) Lag 1 spatial correlations for channels 1-4. (e)-(h) Lag 2 spatial correlations for channels 1-4. Computed for every third scan line.

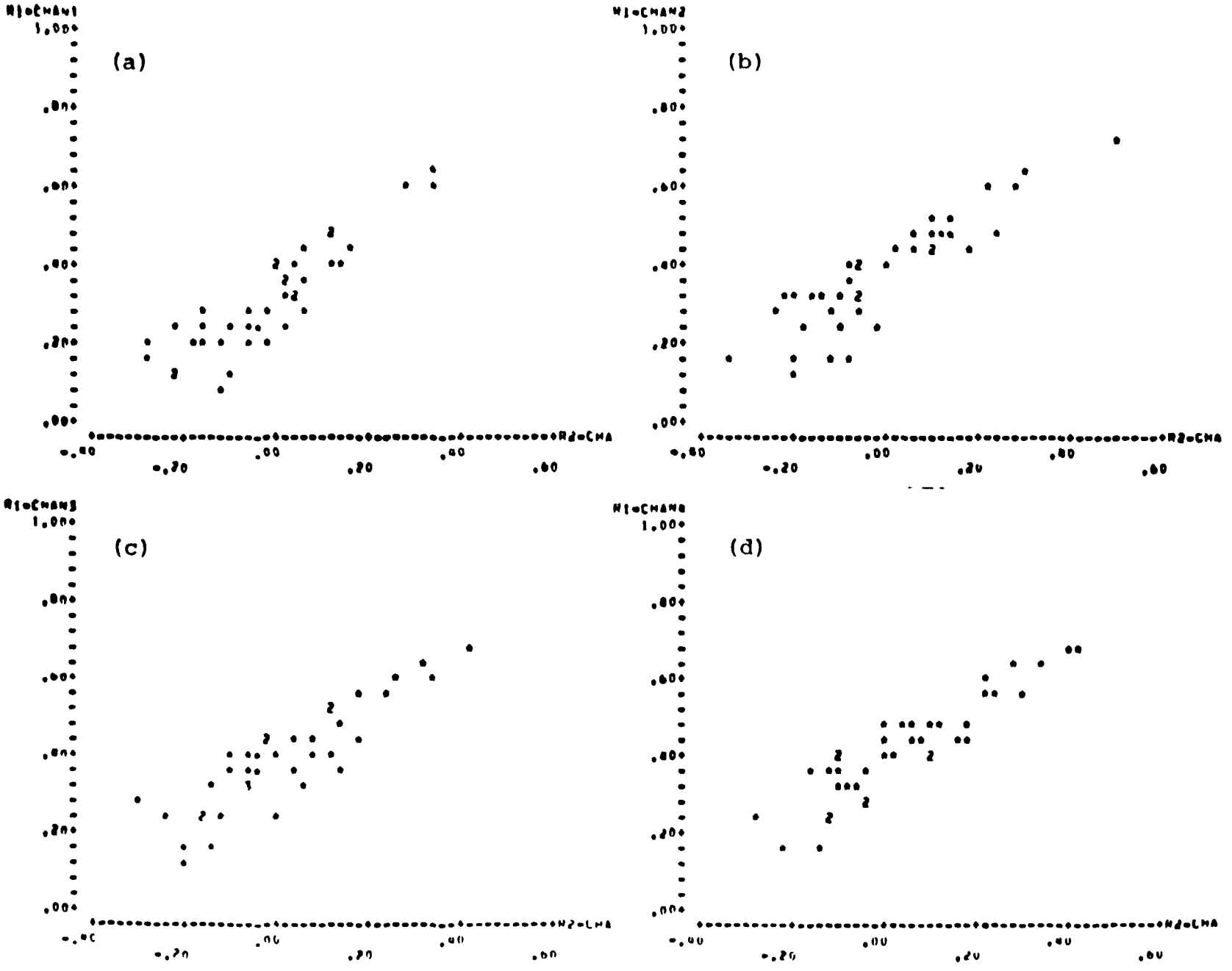


Figure 14. Scatter plots of Lag 1 spatial correlation versus Lag 2 spatial correlation for 1633/129. (a)-(d) channels 1-4. Computed for every third scan line.

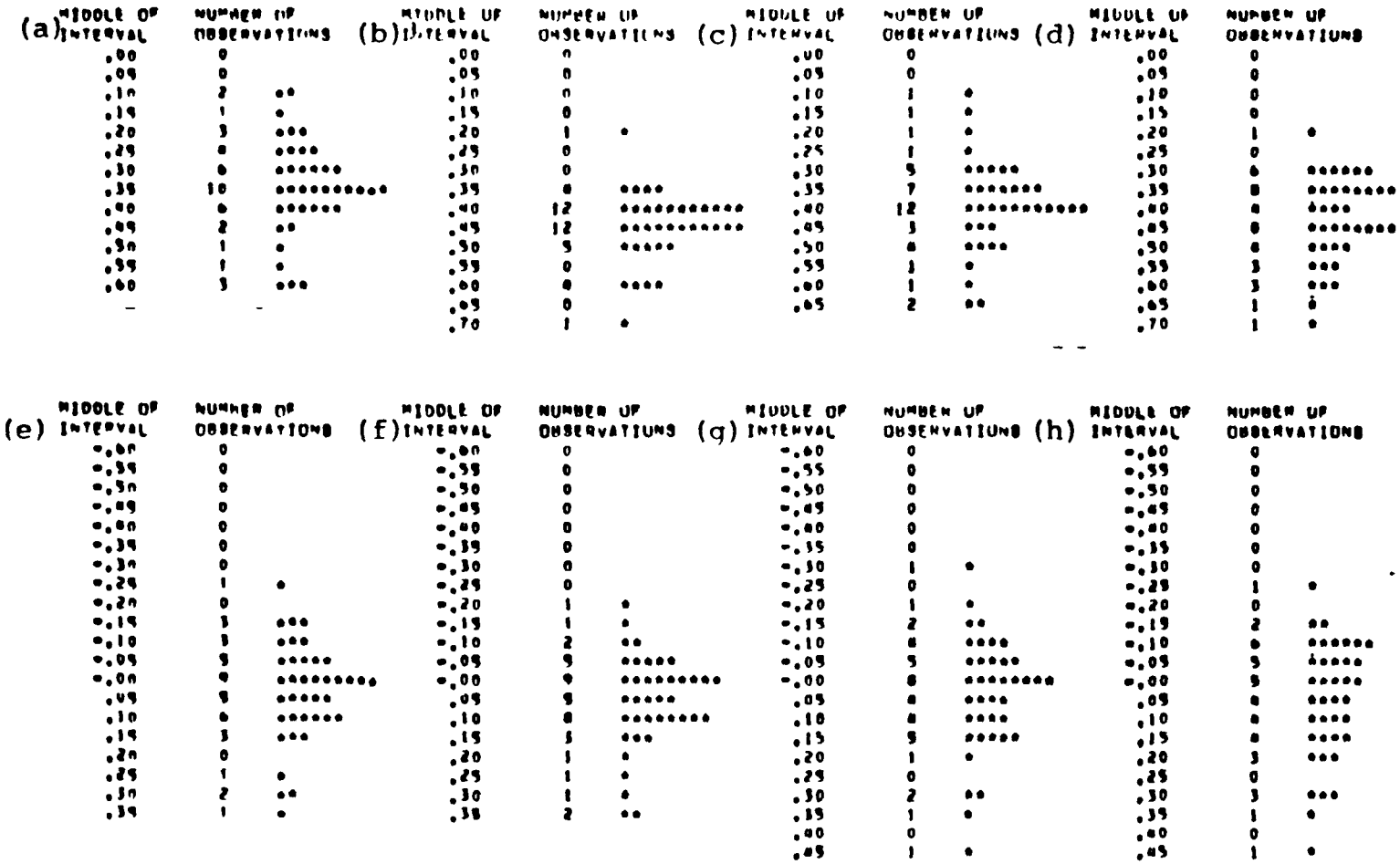


Figure 15. Histograms for 1633/236. (a)-(d) Lag 1 spatial correlations for channels 1-4. (e)-(h) Lag 2 spatial correlations for channels 1-4. Computed for every third scan line.

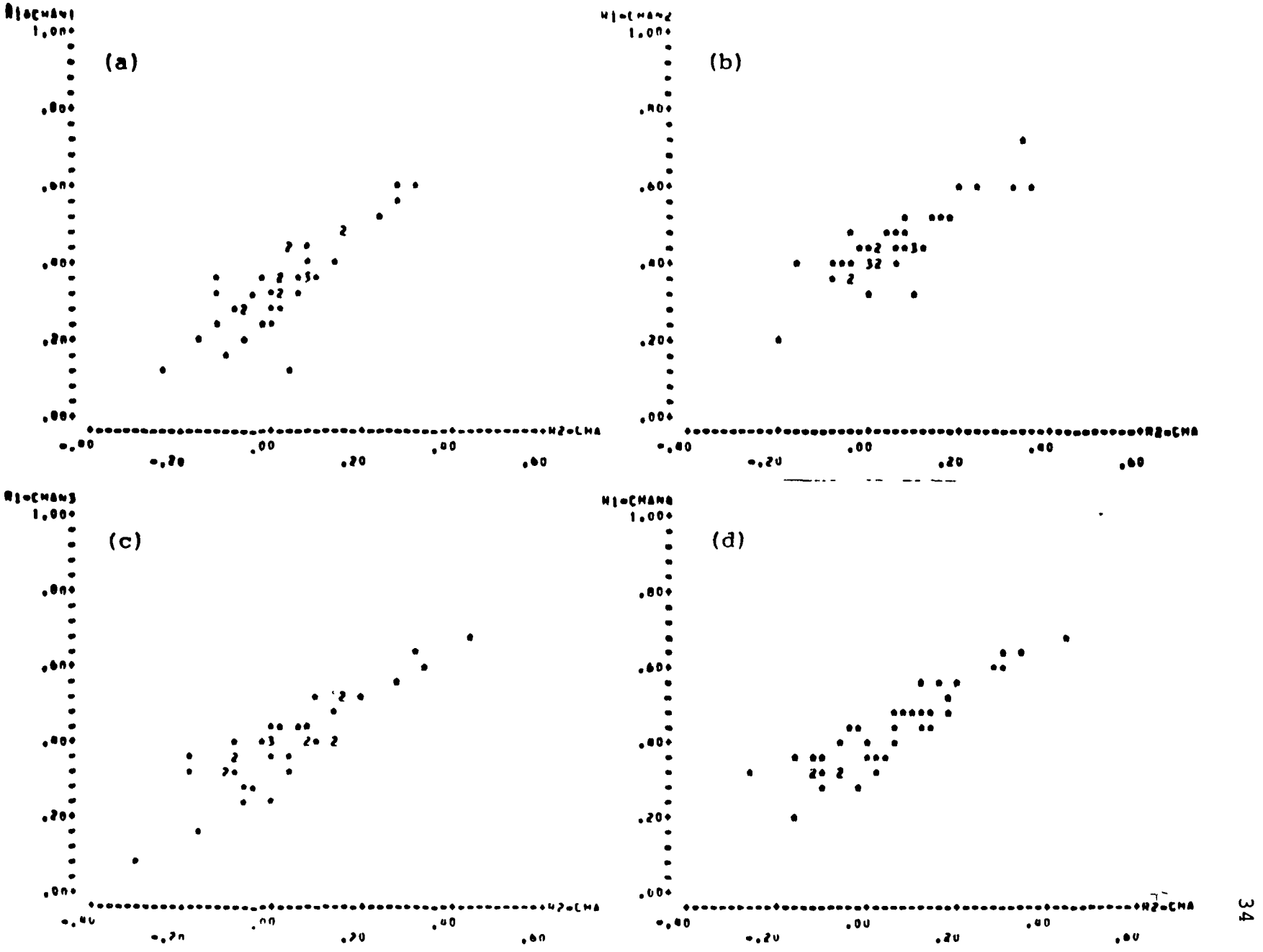


Figure 16. Scatter plots of Lag 1 spatial correlation versus Lag 2 spatial correlation for 1633/236. (a)-(d) channels 1-4. Computed for every third scan line.

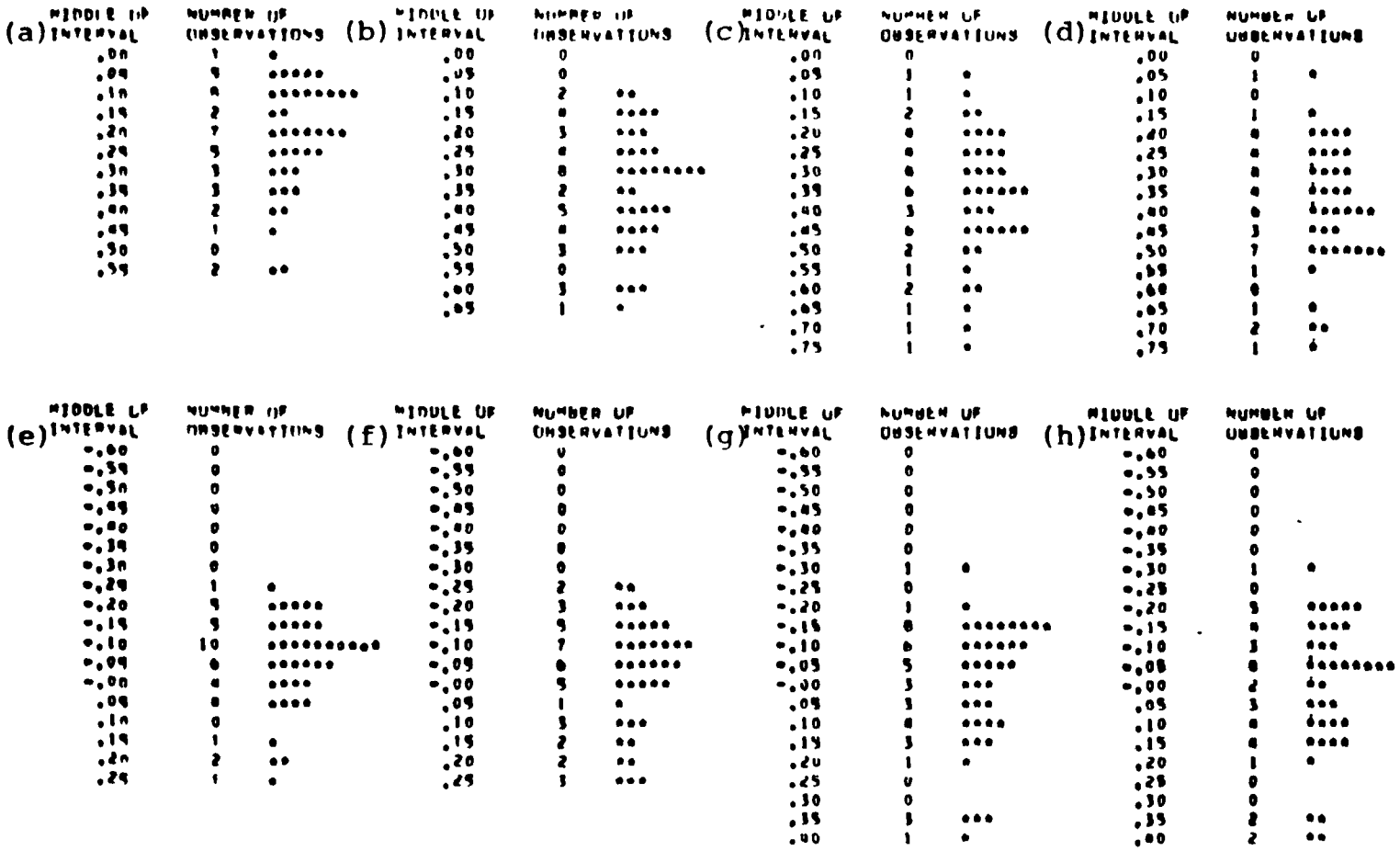


Figure 17. Histograms for 1642/145. (a)-(d) Lag 1 spatial correlations for channels 1-4. (e)-(h) Lag 2 spatial correlations for channels 1-4. Computed for every third scan line.

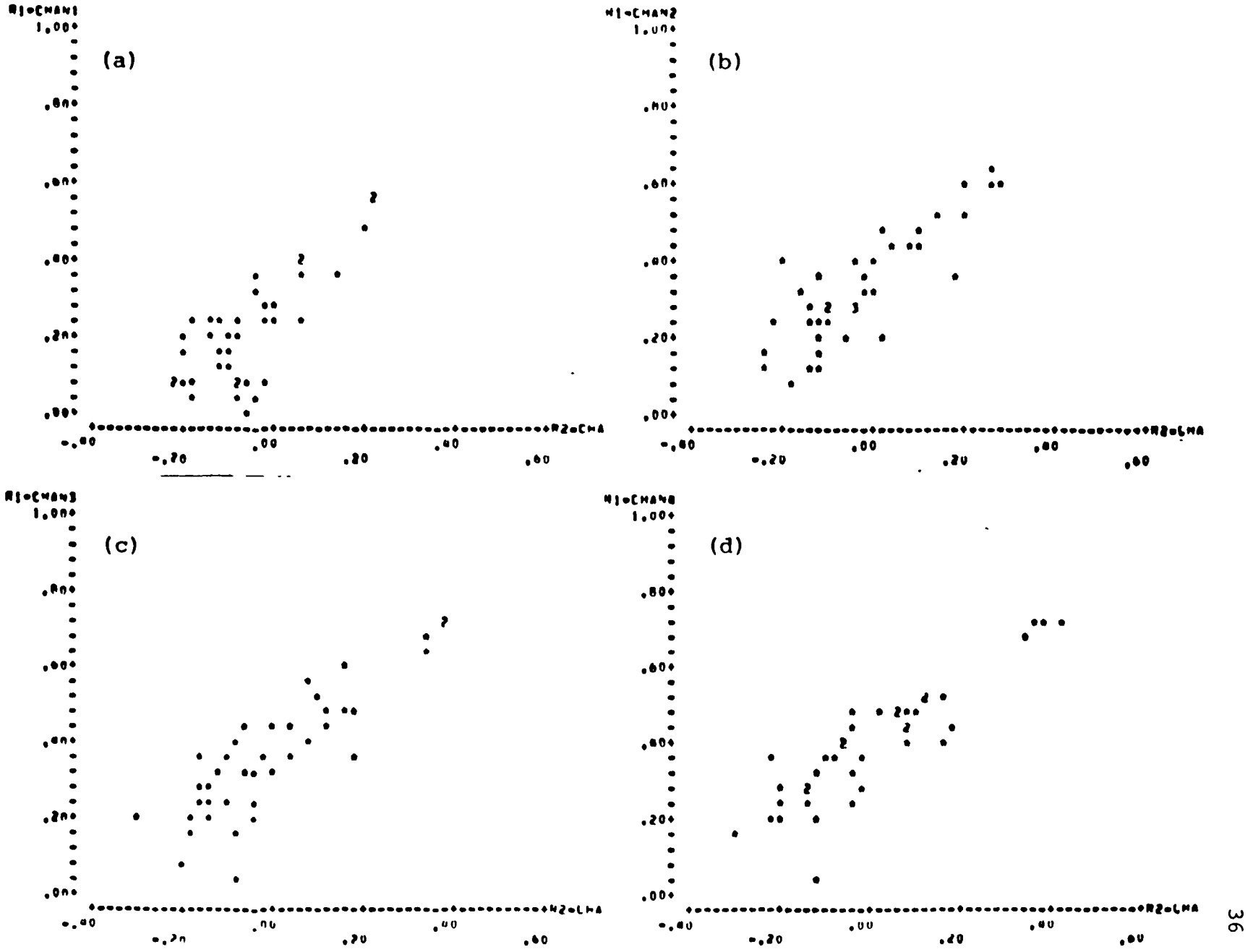


Figure 18. Scatter plots of Lag 1 spatial correlation versus Lag 2 spatial correlation for 1642/145. (a)-(d) channels 1-4. Computed for every third scan line.

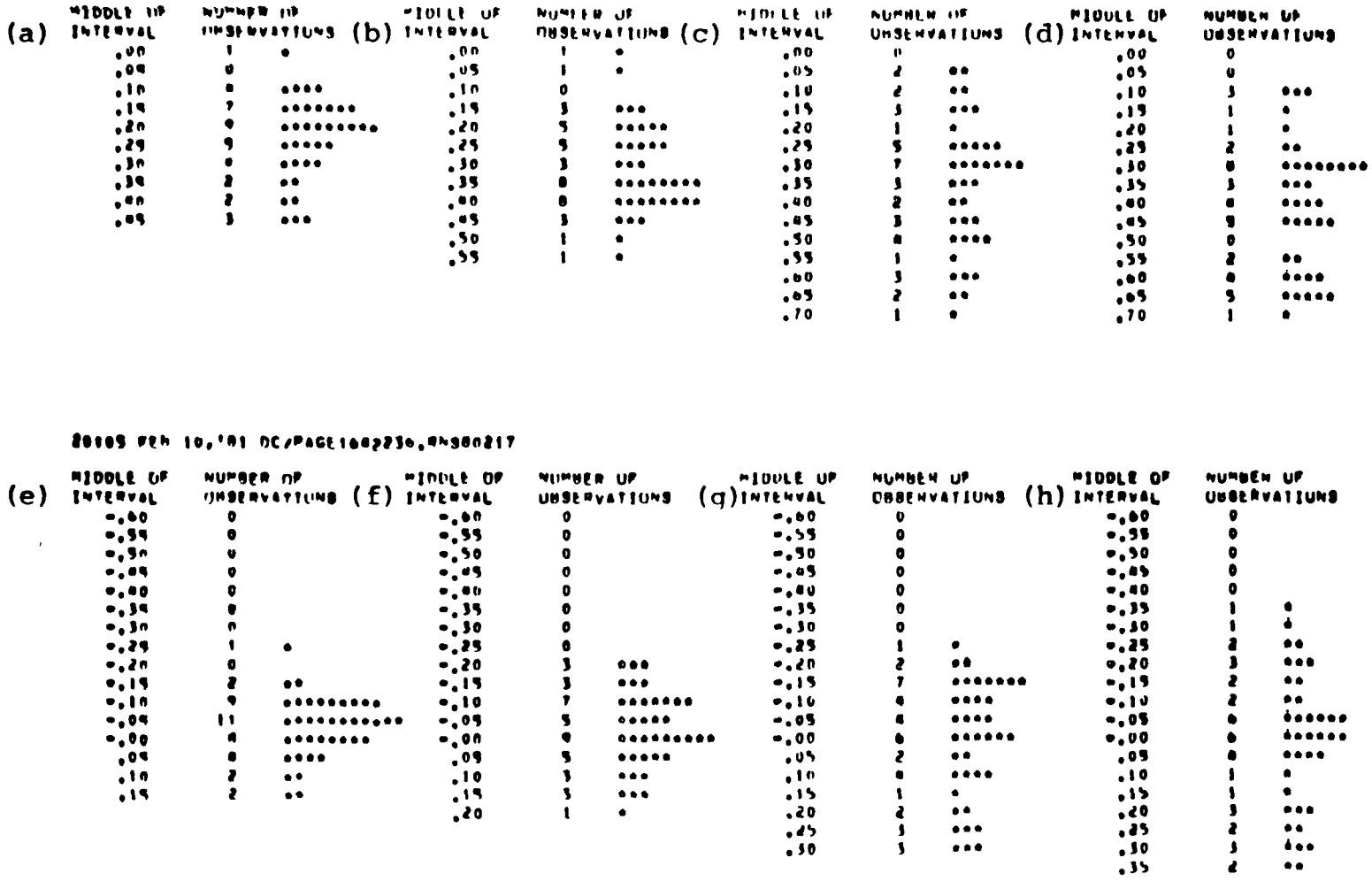


Figure 19. Histograms for 1642/236. (a)-(d) Lag 1 spatial correlations for channels 1-4. (e)-(h) Laq 2 spatial correlations for channels 1-4. Computed for every third scan line.

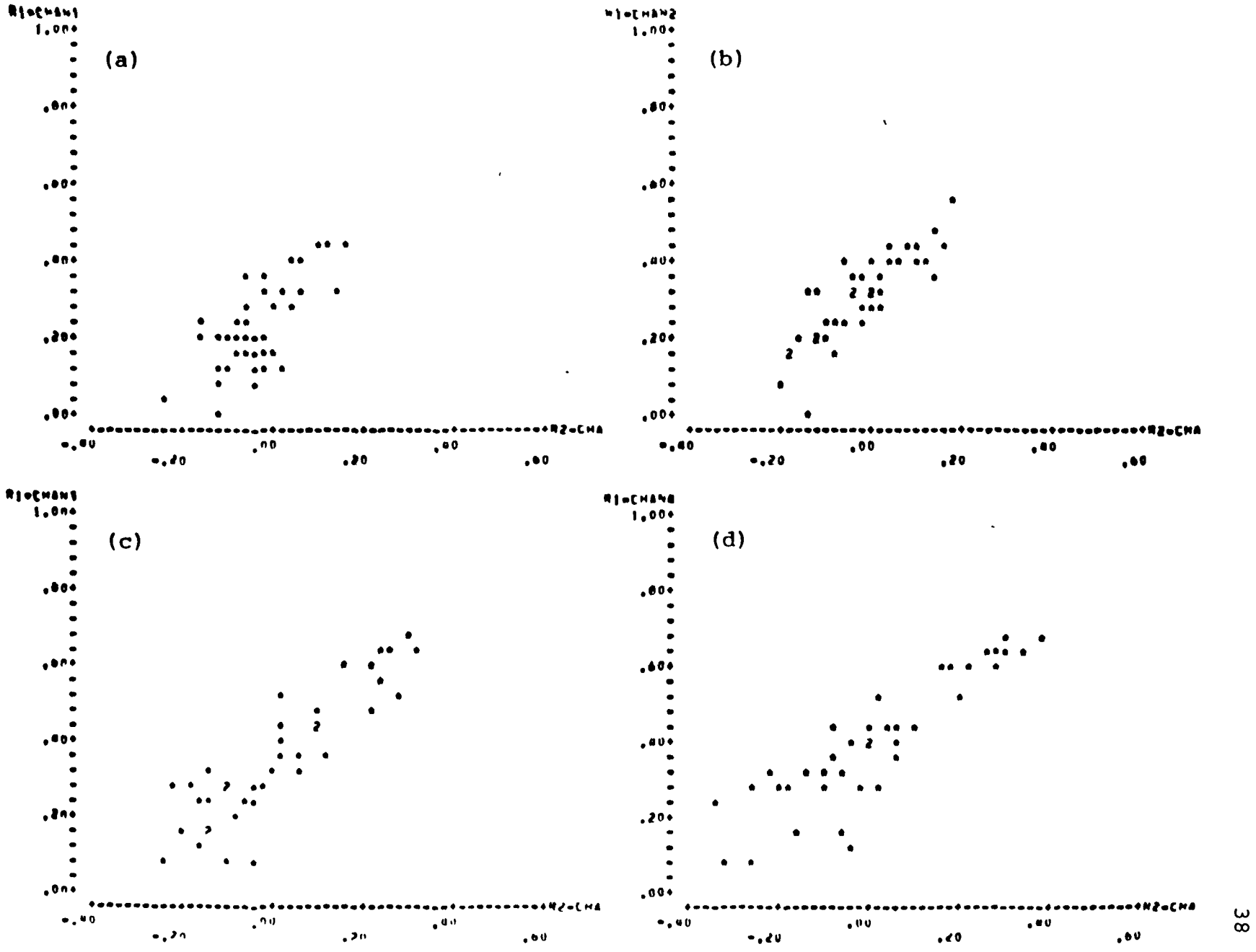


Figure 20. Scatter plots of Lag 1 spatial correlation versus Lag 2 spatial correlation for 1642/236. (a)-(d) channels 1-4. Computed for every third scan line.

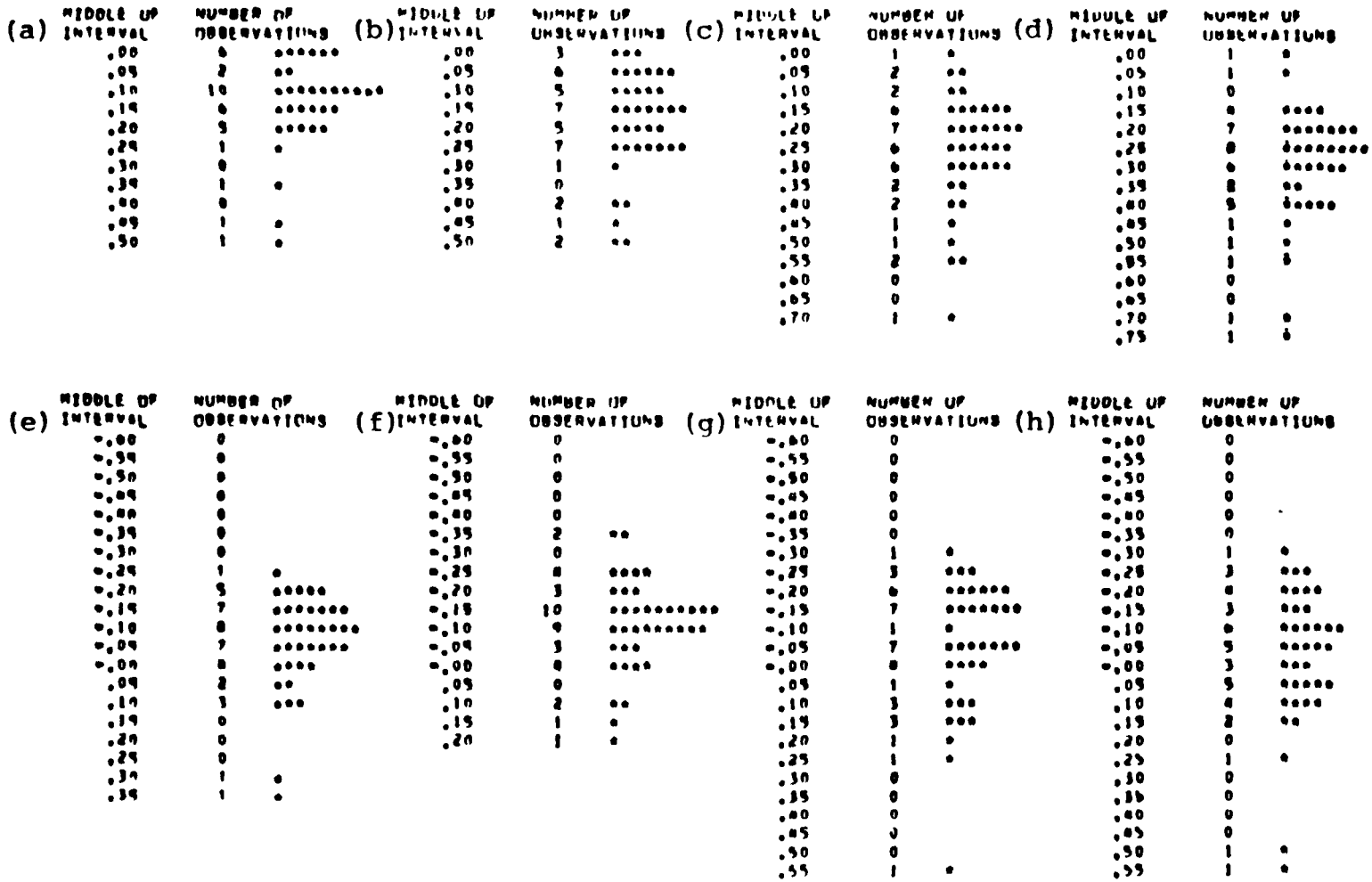


Figure 21. Histograms for 1645/145. (a)-(d) Lag 1 spatial correlations for channels 1-4. (e)-(h) Lag 2 spatial correlations for channels 1-4. Computed for every third scan line.

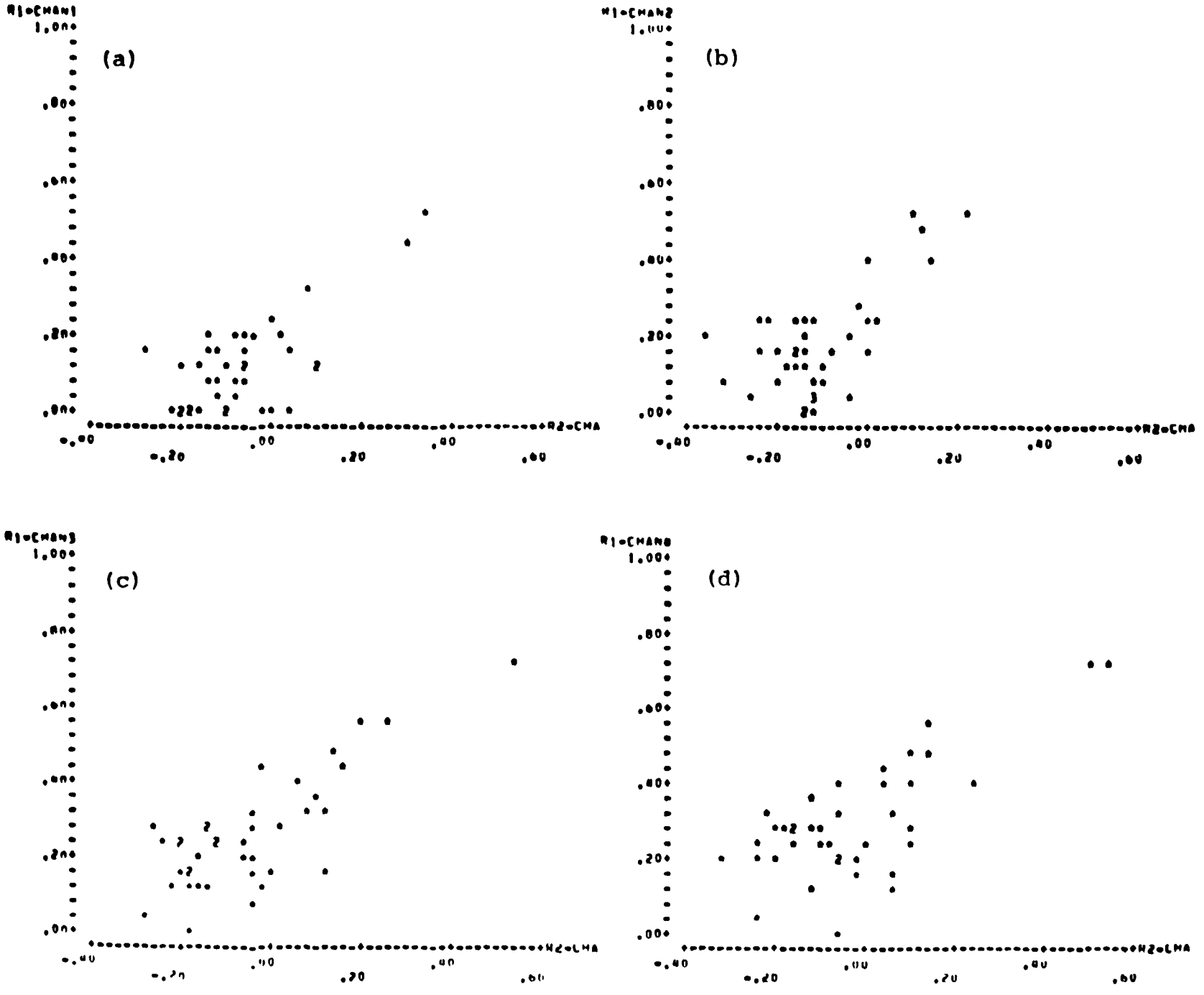


Figure 22. Scatter plots of Lag 1 spatial correlation versus Lag 2 spatial correlation for 1645/145. (a)-(d) channels 1-4. Computed for every third scan line.

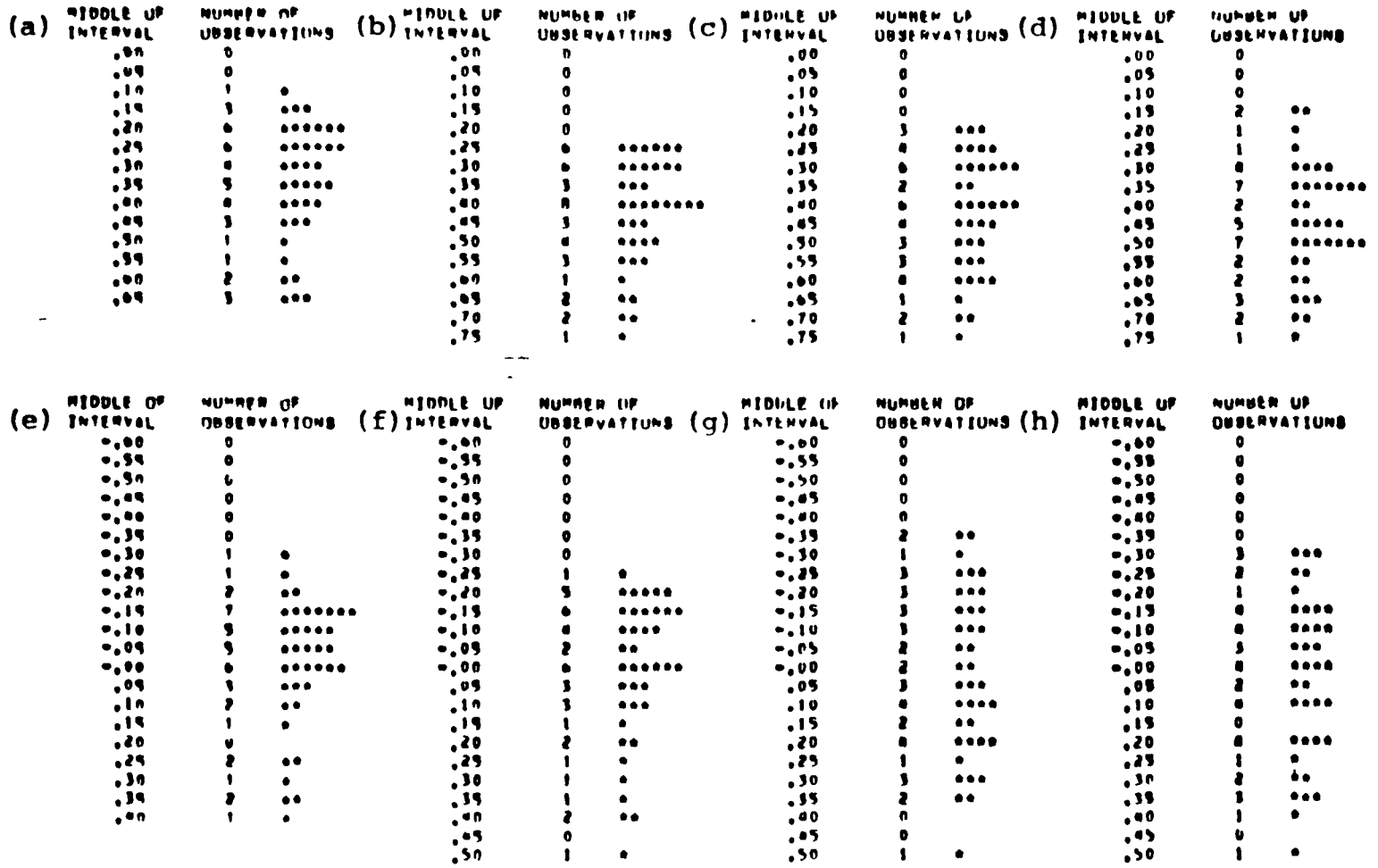


Figure 23. Histograms for 1645/236. (a)-(d) Lag 1 spatial correlations for channels 1-4. (e)-(h) Lag 2 spatial correlations for channels 1-4. Computed for every third scan line.

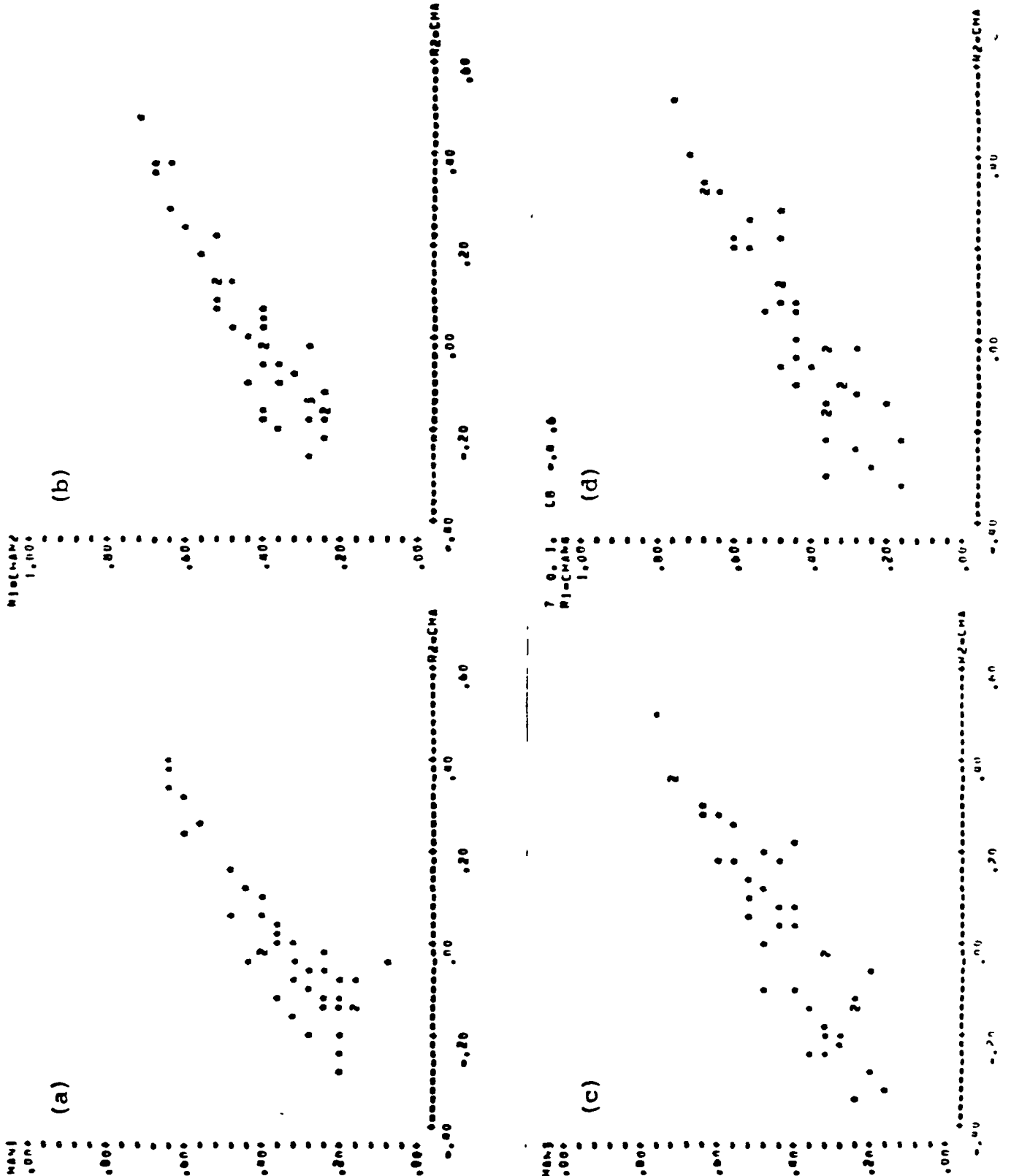


Figure 24. Scatter plots of Lag 1 spatial correlation versus Lag 2 spatial correlation for 1645/236. (a)-(d) channels 1-4. Computed for every third scan line.

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

4 Title and Subtitle

Information in Remotely Sensed Data for Estimating Proportion in Mixture Densities

5 Report Date: November 1980

6 Performing Organization Code

7 Author(s)

Virgil R. Marco, Jr. and Patrick L. Odell

8 Performing Organization Report No: 23

10 Work Unit No

9 Performing Organization Name and Address

Program in Mathematical Science
University of Texas at Dallas
P. O. Box 688
Richardson, Texas 75080

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

13. Type of Report and Period Covered: Unscheduled Technical

2 Sponsoring Agency Name and Address

Earth Observations Division
Johnson Space Center
Houston, Texas 77058

14. Sponsoring Agency Code

5. Supplementary Notes

Principal Investigator: L. F. Guseman, Jr.

5 Abstract

Data taken remotely by multichannel sensors on a near earth satellite can be modeled as a collection of multivariate data points. In the application that motivates this paper each $p \times 1$ data vector represents a measure of reflectance from acre location on the surface of the earth. Each of the p elements of the data vector is a reflectance measure at a preassigned wave length of light. Conceptually, each crop class defines a set of reflectance measures that can be modeled by a multivariate unimodel probability density function unique for each crop class.

7 Key Words (Suggested by Author(s))

proportion estimation, mixture density
multivariate density, classification theory,
remote sensing

18 Distribution Statement

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

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

INFORMATION IN REMOTELY SENSED DATA FOR ESTIMATING PROPORTION IN MIXTURE DENSITIES

by

Virgil R. Marco, Jr.
Program in Mathematical Science
University of Texas at Dallas
Richardson, Texas 75080

Report #23

Prepared For

Earth Observations Division
NASA/Johnson Space Center
Houston, Texas
Contract NAS-9-14680-11S

November 1980

INFORMATION IN REMOTELY SENSED DATA FOR ESTIMATING PROPORTION IN MIXTURE DENSITIES[1]

Virgil R. Marco, Jr., and Patrick L. Odell
University of Texas at Dallas
Box 688, Richardson, Texas, 75080

I. INTRODUCTION

Data taken remotely by multichannel sensors on a near earth satellite can be modeled as a collection of multivariate data points. In the application [1] that motivates this paper each $p\times 1$ data vector represents a measure of reflectance from (1.1) acre location on the surface of the earth. Each of the p elements of the data vector is a reflectance measure at a preassigned wave length of light. Conceptually, each crop class defines a set of reflectance measures that can be modeled by a multivariate unimodel probability density function unique for each crop class.

Let there be m-crop classes and let the p.d.f.

$$P_i(x) = p_i(x;\mu_i,\Sigma_i) \quad i = 1,\ldots,m \tag{1.1}$$

denote the distribution of the random data vector X given that the measurements were made on the i^{th} crop class, Π_i, $i = 1,\ldots,m$. Also let the multivariate mixture p.d.f.

---

[1]This research was supported in part by the National Aeronautics and Space Agency, Johnson Space Center under Contract NAS 9-14689-95.

$$p(x) = \sum_{i=1}^{m} \alpha_i p_i(x) \qquad (1.2)$$

such that $\alpha_i \quad 0 \quad i = 1,2,\ldots,m$ and $\sum_{i=1}^{m} \alpha_i = 1$ denote the distribution of the multivariate observations given that the data is unlabeleld, that is modeled by $p(x)$ in (1.2).

Definition 1. A random sample is said to be unlabeled if the random vectors are selected from a population defined by (1.2).

Definition 2. A random sample of unlabeled data is said to be classified data if, according to some classification rule $R = (R_1, R_2, \ldots, R_m)$, each vector in the sample is assigned to one of the (crop) classes $\Pi_1, \Pi_2, \ldots, \Pi_m$.

Definition 3. A random sample of unlabeled data is said to be verified data if each vector is classified as being from the true subclass Π_i for some $i = 1,2,\ldots,$ or with probability one.

Verified data is classified data in which there is zero probability of misclassification.

Definition 4. A random sample is said to be labeled if it is selected from a single class Π_i and the identity of i^{th} population is known.

The difference between verified and labeled data is that the verified data must be labeled a posteriori while the labeled data is labeled prior to taking the sample. In both types of samples, one knows with certainty the label of the population from which the samples came.

The purpose is to estimate the vector or proportions $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ which defines the function $p(x)$ in (1.2). If α_i

denotes the proportion of vectors in the mixture from class Π_i and N the total number of vectors in the region, then

$$\hat{A}_i = (1.1)\, N\, \hat{\alpha}_i \tag{1.3}$$

is an estimate of acreage of crop class Π_i, as a function of an estimate of the proportion $\hat{\alpha}_i$ and α_i. Hence, our interest is to estimate α well.

Three different types of data are available for estimating the elements of α arise naturally in the application involving remote sensing from space. They all are maximum likelihood estimators for α using

(a) unlabeled data,

(b) classified data, or

(c) verified data, respectively.

The cost of acquiring unlabeled data is less than the cost of acquiring classified data which is in turn less than the cost of acquiring verified data. The computation of sample size allocations when samples from more than one type of data are available arises naturally. In the case of sample design one can control the type of data to be selected and the optimal mix of sampling can be accomplished. It is important to note that one <u>always</u> has available a random sample of unlabeled data; hence if C_u denotes the cost per unit of taking unlabeled data then

$$C_v = C_u + c_v = K_v C_u$$

$$C_c = C_u + c_c = K_c C_u$$

are the per unit cost where C_V and C_u are the costs of classifying and verifying in unlabeled data point respectively. The values K_V and K_C are multiplicative constants that give in addition to an additive model a second multiplicative representation of the costs.

One would expect $C_u < C_c < C_v$ in most space science applications. It is important to note that in the space application unlabeled data is available as basic for two of the three methodologies for estimating α, and except for missing data that the totality of unlabeled data is also available. The cost of machine processing every vector is a realistic limiting factor for unlabeled and classified data while the cost of resources to visit each location for verification is the major limiting factor for obtaining verified data.

However, it is not intuitively clear which type of data contains greatest amount of information for estimating α for a fixed sample size. The purpose of this paper is to compute and order with respect to magnitude the information content of the three types of data, and discuss the implications of that ordering for the space application.

The term information content of the data is defined as the inverse of the Cramer-Rao matrix lower bound for unbiased estimators for α. This is the matrix form of Fisher's Information.

II. INFORMATION CONTENT OF VARIOUS TYPES OF DATA

2.1 Fisher's Information: Let X denote a random observation from a multivariate (p-variate) population whose p.d.f. is defined by (1.2). If we denote the parameter vector by $\alpha = (\alpha_1,\ldots,\alpha_{m-1})^T$ then by the usual theory (Cramer [2], Rao [3]) the $(m-1 \times 1)$ random vector

$$S = \frac{\partial \ell np(x)}{\partial \alpha} \tag{2.1.1}$$

is such that

$$E[S] = \phi$$

and

$$E[S\,S^T] = -E\left[\frac{\partial^2 \ell np(x)}{\partial\alpha\ \partial\alpha}\right] = -E\left\{\frac{\partial^2 \ell np(x)}{\partial\alpha_i\ \partial\alpha_j}\right\} \overset{def}{=} \Lambda(\alpha) \tag{2.1.2}$$

where $\Lambda(\alpha)$ denotes Fisher's information for α contained in the sample X.

If $X_1,\ldots,X_n$ denote a random sample from a multivariate population whose p.d.f. is defined by (1.2), then the Fisher's information for α contained in this sample can be shown to be

$$E[S\,S^T] = n\,\Lambda(\alpha)\ . \tag{2.1.3}$$

Furthermore, $\Lambda^{-1}(\alpha)$ is the Cramer-Rao lower covariance matrix bound for unbiased estimators of the vector α. That is, if $\hat{\alpha}$ is any <u>unbiased</u> estimator for α, then the covariance matrix $\Lambda(\hat{\alpha})$ will never be less than $\Lambda^{-1}(\alpha)$. 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 or equal to (when $A - B = \Phi$) than A.

From (1.2) it follows that

$$p(x) = \sum_{j=1}^{m-1} \alpha_j p_j(x) + \left(1 - \sum_{j=1}^{m-1} \alpha_j\right) p_m(x) \qquad (2.1.4a)$$

$$= \sum_{j=1}^{m-1} \alpha_j [p_j(x)-p_m(x)] + p_m(x) \; . \qquad (2.1.4b)$$

It follows from (2.1.1) that

$$S_j = \frac{p_j(x) - p_m(x)}{\sum_{j=1}^{m} \alpha_j p_j(x)}$$

$$= \frac{p_j(x)-p_m(x)}{p(x)} \qquad (2.1.5)$$

and

$$\frac{\partial S_j}{\partial \alpha_K} = - \frac{[p_j(x)-p_m(x)][p_k(x)-p_m(x)]}{[p(x)]^2} \; . \qquad (2.1.6)$$

Therefore, the information for α is given by

$$\Lambda(\alpha) \overset{\text{def}}{=} \left\{ -E\left[\frac{\partial S_j}{\partial \alpha_K} \right] \right\}_{(m-1)\times(m-1)} \; . \qquad (2.1.7)$$

Fisher's information can be seen as the information contained in a random variable X about the parameter α. This should be interpreted

as the extent to which, on the average, the accuracy of estimating the unknown parameter α can be increased as a result of the observed value x of the random variable X.

In the ensuing sections of this paper, information for α contained in unlabeled, classified and verified data, defined earlier will be ordered.

Above, information is defined in terms of unbiased estimators.

2.2 Likelihood Function. If $X_1,X_2,\ldots,X_n$ denotes a simple random sample from $p(x)$ defined by (1.2) then the likelihood function is

$$L_u(X_1,\ldots,X_n) = \prod_{i=1}^{n} p(X_i) \tag{2.2.1a}$$

$$= \prod_{\ell=1}^{n}\left[\sum_{j=1}^{m} \alpha_j\, p_j(X_i)\right] \tag{2.2.1b}$$

the likelihood function for unlabeled data.

Let $X_1,X_2,\ldots,X_n$ denote a simple random sample from $p(x)$ which has been classified according to a rule $R = (R_1,R_2,\ldots,R_m)$, then each data vector X_k, $k = 1,2,\ldots,n$ generates through classification new data defined by the random variable $Y_i(X_k)$, $i = 1,2,\ldots,m$, where

$$Y_i(X_k) = 1 \quad \text{if} \quad X_k \in R_i \tag{2.2.2}$$

$$= 0 \quad \text{if} \quad X_k \notin R_k$$

whose joint p.d.f. is for each X_k a multinomial

$$h_{Y_1 \ldots Y_m}(y_1(X_k),\ldots,y_m(X_k)) = \prod_{i=1}^{m} g_i^{y_i(X_k)} \tag{2.2.3}$$

where

$$\begin{aligned} g_j &= Pr[X_k \in R_j] \\ &= \int_{R_i} p(x)dx \\ &= \sum_{j=1}^{m} \alpha_j \int_{R_i} p_j(x)dx \\ &= \sum_{j=1}^{m} \alpha_j p(i\ j)\ , \end{aligned}$$

the probability of classifying $I(X_k)$ in Π_j .

The likelihood function for classified data follows from (2.2.3), and is

$$\begin{aligned} L_C &= L(Y_1(X_1),\ldots,Y_m(X_1);\ldots;Y_1(X_n),\ldots,Y_m(X_n)) \\ &= \prod_{k=1}^{n} \prod_{i=1}^{m} g_i^{Y_i(X_k)} \\ &= \prod_{k=1}^{n} \prod_{i=1}^{m} \left[\sum_{j=1}^{m} \alpha_j P(i|j) \right]^{Y_i(X_k)} \\ &= \prod_{i=1}^{m} \left[\sum_{j=1}^{m} \alpha_j P(i|j) \right]^{N_i} \end{aligned} \tag{2.2.4}$$

where

$$N_i = \sum_{k=1}^{n} Y_i(X_k) \tag{2.2.5}$$

the number of sample vectors in R_i .

Let $I_1(X_1), I_2(X_2), \ldots, I_n(X_n)$ denote a random sample whose labels are known with probability one, that is, the data has been verified, then

$$T_j(I_k) = 1 \quad \text{if} \quad I_k \in \Pi_j$$
$$= 0 \quad \text{if} \quad I_k \in \Pi_j \tag{2.2.6a}$$

then the p.d.f. of $T = (T_1, \ldots, T_m)^T$ for each I_k is

$$f_{T_1,\ldots,T_m}(t_1,\ldots,t_m) = \prod_{i=1}^{m} [\alpha_i]^{t_i(I_k)} \quad . \tag{2.2.6b}$$

The likelihood function of a verified sample is

$$L_v = L_v(T_1(I_1),\ldots,T_m(I_1);\ldots;T_1(I_m),\ldots,T_m(I_n))$$

$$= \prod_{k=1}^{n} \prod_{i=1}^{m} [\alpha_i]^{T_i(I_k)}$$

$$= \prod_{i=1}^{m} [\alpha_i]^{n_i} \tag{2.2.7}$$

where

$$n_i = \sum_{k=1}^{n} T_i(I_k) \ , \tag{2.2.8}$$

the number of individuals in the sample from Π_i .

2.3 Information for α Contained in Unlabeled Data.

Let the following denote the information for α contained in unlabeled data: $X_1, \ldots, X_n$:

$$\Lambda_u(\alpha) = n\left\{\Lambda_{ij}^u(\alpha)\right\}_{(m-1)\times(m-1)} .$$

Using (2.1.2), (2.2.1b) and synthetic division, it can be shown that for $i = j$

$$\Lambda_{ij}^u = \left(\frac{\alpha_i+\alpha_m}{\alpha_i\alpha_m}\right)\left[1 - (\alpha_i+\alpha_m)B_{im} - \frac{\alpha_m}{(\alpha_i+\alpha_m)} \sum_{\substack{k=1\\k\neq i}}^{m-1} \alpha_k B_{ij} - \frac{\alpha_i}{(\alpha_i+\alpha_m)} \sum_{\substack{j=1\\j\neq i}}^{m=1} \alpha_j B_{jm}\right] \tag{2.3.1a}$$

and for $i \neq j$

$$\Lambda_{ij}^u = \frac{1}{\alpha_m}\left[1 - (\alpha_i+\alpha_m)B_{im} - (\alpha_j+\alpha_m)B_{jm} - \sum_{\substack{k=1\\k\neq i,j}}^{m-1} \alpha_k B_m + \alpha_m B_{ij}\right] \tag{2.3.1b}$$

where

$$0 \leq B_{ij} = \int_{\mathbb{R}^p} \frac{p_i(x)p_j(x)}{p(x)}\,dx \leq 1 \tag{2.3.1c}$$

and $B_{jk} = B_{kj}$, for all $j \neq k$.

When $B_{ij} = B$,

$$\Lambda_u(\alpha) = n(1-B)\{\Lambda_{1j}^u\} \tag{2.3.2a}$$

where

$$\Lambda^u_{ij} = \frac{\alpha_i+\alpha_m}{\alpha_i\alpha_m} \quad \text{for} \quad i = j \tag{2.3.2b}$$

$$= \frac{1}{\alpha_m} \quad \text{for} \quad i \neq j \ . \tag{2.3.2c}$$

When $m = 3$, the p.d.f. of a random variable X from a mixture population (unlabeled data) is

$$p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x) + \alpha_3 p_3(x) \tag{2.3.3a}$$

where

$$\alpha_1 + \alpha_2 + \alpha_3 = 1 \tag{2.3.3b}$$

and

$$\alpha_1 \geq 0 \ , \ \alpha_2 \geq 0 \ , \ \alpha_3 \geq 0 \ . \tag{2.3.3c}$$

It follows from (2.3.1a) - (2.3.1c) that the information contained in unlabeled data is given by

$$\Lambda_u(\alpha) = \begin{bmatrix} \Lambda^u_{11} & \Lambda^u_{12} \\ \\ \Lambda^u_{21} & \Lambda^u_{22} \end{bmatrix}$$

where

$$\Lambda^u_{11} = \frac{(1-\alpha_2)}{\alpha_2\alpha_3}\left[1 - \frac{\alpha_2\alpha_3}{1-\alpha_2} B_{12} - (1-\alpha_2)B_{13} - \frac{\alpha_1\alpha_2}{1-\alpha_2} B_{23}\right] \tag{2.3.4a}$$

$$\Lambda^u_{22} = \frac{(1-\alpha_1)}{\alpha_2\alpha_3}\left[1 - \frac{\alpha_1\alpha_3}{1-\alpha_1} B_{12} - \frac{\alpha_1\alpha_2}{1-\alpha_1} B_{13} - (1-\alpha_1)B_{23}\right] \tag{2.3.4b}$$

$$\Lambda_{12}^{u} = \Lambda_{21}^{u} = \frac{1}{\alpha_3}\left[1+\alpha_3 B_{12} - (1-\alpha_2)B_{13} - (1-\alpha_1)B_{23}\right] \quad . \qquad (2.3.4c)$$

Note that one minus (2.3.1c) can be regarded as a distance measure. That is, when the i^{th} and j^{th} populations are "close together" or "far apart" then $(1-B_{ij})$ will be small or large, respectively. In fact, several investigators [3], [5], [6] , have employed a form of (2.3.1c) as a probabilistic distance measure for feature selection. While Cover and Hart [8] have shown that $2\alpha_i\alpha_j B_{ij}$ corresponds to the asymptotic nearest neighbor probability of error, this motivates a possible estimating procedure (see section 4.) using a nearest neighbor procedure.

It is of interest to consider the behavior of B_{ij} in terms of a popular distance measure as the distance between the i^{th} and j^{th} populations diverges. This behavior is described in Lemma 2.3.1.

Lemma 2.3.1: Let the distance measure between the i^{th} and j^{th} populations be given by

$$\Delta_{ij} = \int \left[p_i(x) - p_j(x)\right] \log\left[\frac{p_i(x)}{p_j(x)}\right] dx \quad . \qquad (2.3.5)$$

If $\Delta_{ij} \to \infty$ for all $i \neq j$, then $B_{ij} \to 0$.

Proof: Toussant [4] has shown that

$$0 \leq B_{ij} \leq \frac{1}{2}\left(\frac{\Delta_{ij}}{4}\right)^{-\frac{1}{4}} \quad .$$

Note that as $\Delta_{ij} \to \infty$ then

$$\left(\frac{\Delta_{ij}}{4}\right)^{-\frac{1}{4}} \to 0 \ .$$

Note that (2.3.5) is known as the divergence between two distributions. For normal distributions with equal covariances, (2.3.5) reduces to the well known Malhanabis distance.

The following example can clarify some of the concepts introduced above:

Example 2.3.1:

$$P_1(x) = \begin{cases} x & , 0<x<1 \\ 2-x, & 1<x<2 \\ 0 & , \text{o.w.} \end{cases} \quad , p_2(x) = \begin{cases} x-1, & 1<x<2 \\ 3-x, & 2<x<3 \\ 0 & , \text{o.w.} \end{cases} \quad , p_3(x) = \begin{cases} x-2, & 2<x<3 \\ 4-x, & 3<x<4 \\ 0 & , \text{o.w.} \end{cases}$$

$$p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x) + \alpha_3 p_3(x) \quad .$$

Let $\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}$ then

$$B_{12} = \int_0^4 \frac{p_1(x)p_2(x)}{p(x)}\,dx = \int \frac{(2-x)(x-1)}{\frac{1}{3}(2-x+x-1)}\,dx$$

$$= \int_1^2 3(2-x)(x-1)dx$$

$$= \int_1^2 (3x-2-x^2)dx$$

$$= \frac{1}{2}$$

$$B_{23} = 3\int_2^3 (3-x)(x-2)dx = \frac{1}{2}$$

$$B_{13} = 0 \ .$$

$$\therefore \Lambda(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}) = \begin{bmatrix} 5 & \frac{5}{2} \\ \frac{5}{2} & \frac{7}{2} \end{bmatrix}$$

To conclude this section, a result that follows from Lemma 2.3.1 is given.

Theorem 2.3.1: Let Δ_{ij} be a distance measure defined by (2.3.5). If $\Delta_{ij} \to \infty$ for all $i \neq j$ then ,

$$\Lambda_u(\alpha) \to \Lambda_V(\alpha) = n\{\Lambda^V_{ij}\}$$

where

$$\Lambda^V_{ij} = \begin{cases} \dfrac{\alpha_i+\alpha_m}{\alpha_i\alpha_m} & \text{for } i = j \\ \dfrac{1}{\alpha_m} & \text{for } i \neq j \end{cases} .$$

Proof: Using equations (2.3.1a) - (2.3.1c) and letting $\Delta_{ij} \to \infty$, the Theorem follows from Lemma 2.3.1.

Note that (2.3.2a) can be written as

$$\Lambda_u(\alpha) = n(1-B)\Lambda_V(\alpha) . \tag{2.3.6}$$

The information matrix $\Lambda_V(\alpha)$ is the information for α contained in verified data. This is a topic of the next section.

2.4 Information for α Contained in Verified Data

Let $T_i(I_k)$ be defined as in (2.2.6a). It follows from (2.2.7) that

$$\ell n\, L_V = \ell n\left[\prod_{i=1}^{m} \alpha_i^{n_i}\right]$$

$$= \sum_{i=1}^{m} n_i\, \ell n[\alpha_i] \tag{2.4.1}$$

$$= \sum_{i=1}^{m-1} n_i\, \ell n[\alpha_i] + n_m\, \ell n\left[1 - \sum_{j=1}^{m-1} \alpha_j\right] ,$$

since $\sum_{j=1}^{m} \alpha_j = 1$.

From (2.1.1) then $S_j = \frac{\partial \ell n\, L}{\partial \alpha_j}$ it follows that

$$S_V = \frac{\partial\, \ell n\, L_V}{\partial \alpha} = \{S_j^V\}$$

where

$$S_j^V = \frac{\partial}{\partial \alpha_j}\left[\sum_{i=1}^{m} n_i\, \ell n\, \alpha_i\right] \tag{2.4.2}$$

$$= \frac{n_j}{\alpha_j} - \frac{n_m}{\alpha_n} \quad , j = 1,\ldots,m-1 .$$

In matrix notation

$$S_V = \Delta_\alpha \bar{n} \tag{2.4.3}$$

where the $(m-1)\times m$ matrix Δ_α is given by

$$\Delta_\alpha = \begin{bmatrix} \frac{1}{\alpha_1} & 0 & 0 & \ldots & 0 & 0 & -\frac{1}{\alpha_m} \\ 1 & \frac{1}{\alpha_2} & 0 & \ldots & 0 & 0 & -\frac{1}{\alpha_m} \\ \cdot & & & & & \cdot & \cdot \\ \cdot & & & & & \cdot & \cdot \\ \cdot & & & & & 0 & \cdot \\ 0 & 0 & 0 & \ldots & 0 & \frac{1}{\alpha_{m-1}} & -\frac{1}{\alpha_m} \end{bmatrix} \tag{2.4.4}$$

and

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

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

$$E[S_V] = E[\Delta_\alpha \bar{n}]$$

$$= \Delta_\alpha E[\bar{n}]$$

$$= n\, \Delta_\alpha\, \alpha \quad \text{since} \quad n_j \sim \text{multinomial}\ (n,\alpha_j) \ \text{for} \ j = 1,\ldots,m\ .$$

Now,

$$\Delta_\alpha \alpha = \begin{bmatrix} \frac{1}{\alpha_1} & 0 & 0 \ldots 0 & -\frac{1}{\alpha_m} \\ 0 & \frac{1}{\alpha_1} & 0 \ldots 0 & -\frac{1}{\alpha_m} \\ \vdots & & \vdots & \vdots \\ 0 \ldots\ldots 0 & & \frac{1}{\alpha_{m-1}} & -\frac{1}{\alpha_m} \end{bmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \Phi \qquad (2.4.5)$$

Thus,

$$E[S_V] = \Phi \quad . \qquad (2.4.6)$$

The information matrix for α when sampling from verified data can now be computed by finding the covariance matrix $V(S_V)$ of S_V using (2.4.3) and (2.4.6), that is,

$$\Lambda_V(\alpha) = V(S)$$

$$= \Delta_\alpha\, V(\bar{n})\, \Delta_\alpha^T \qquad (2.4.7)$$

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

$$V(n) = n[\mathrm{Diag}(\alpha_1,\ldots,\alpha_m) - \alpha\alpha^T] \quad . \tag{2.4.8}$$

From (2.4.7), (2.4.8) and (2.4.5),

$$\begin{aligned} \Lambda_V(\alpha) &= \Delta_\alpha[\mathrm{Diag}(\alpha_1,\ldots,\alpha_m) - \alpha\alpha^T]\,\Delta_\alpha^T \\ &= \Delta_\alpha[\mathrm{Diag}(\alpha_1,\ldots,\alpha_m)]\,\Delta_\alpha^T \quad . \end{aligned} \tag{2.4.9}$$

For exemplary purposes consider the case when $m = 3$, then since

$$\Delta_\alpha = \begin{pmatrix} \frac{1}{\alpha_1} & 0 & -\frac{1}{\alpha_3} \\ 0 & \frac{1}{\alpha_2} & -\frac{1}{\alpha_3} \end{pmatrix} ,$$

$$\Lambda_V(\alpha) = \begin{pmatrix} \frac{1}{\alpha_1} + \frac{1}{\alpha_3} & \frac{1}{\alpha_3} \\ \frac{1}{\alpha_3} & \frac{1}{\alpha_2} + \frac{1}{\alpha_3} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\alpha_1+\alpha_3}{\alpha_1\alpha_3} & \frac{1}{\alpha_3} \\ \frac{1}{\alpha_3} & \frac{\alpha_2+\alpha_3}{\alpha_2\alpha_3} \end{pmatrix} , \tag{2.4.10}$$

Suppose we are given an unlabeled sample

$$X_1,\ldots,X_n \, .$$

Then we verify this sample generating the sample

$T_1,\ldots,T_n$, where $T_i = (T_{i1},\ldots,T_{im})^T$

For estimating α_1 should we disregard the unlabeled sample or consider the joint sample (X_i,T_i) , $i = 1,\ldots,n$? The joint p.d.f. of (X_i,T_i) , $i = 1,\ldots,n$ is

$$p(x_i,t_i) = p(x_i|t_i)p(t_i) \ , \ t_i = (t_{i1},\ldots,t_{im})$$

$$= \prod_{j=1}^{m} [p_j(x_i)]^{t_{ij}} \prod_{j=1}^{m} [\alpha_j]^{t_{ij}}$$

$$= \prod_{j=1}^{m} [\alpha_j \ p_j(x_i)]^{t_{ij}} \ . \qquad (2.4.11)$$

To answer the above question consider the following theorem.

Theorem 2.4.1: The amount of information for α contained in the observation (x_i,t_i) is equal to the information for α contained in the observation t_i alone.

Proof: Taking the logs of both sides of the equality in (2.4.11), we see that

$$\ell n \ p \ (x_i,t_i) = \sum_{j=1}^{m} t_{ij} \ \ell n \ p_j(x_i) + \sum_{j=1}^{m} t_{ij} \ \ell n \ \alpha_j \ .$$

Now taking derivative with respect to α_j we have

$$\frac{\partial \ \ell n \ p(x_i,t_i)}{\partial \ \alpha_j} = 0 + \frac{\partial \sum_{j=1}^{m} t_{ij} \ \ell n \ \alpha_j}{\partial \ \alpha_j} = \frac{\partial \ \ell n \ p(t_i)}{\partial \ \alpha_j} \ .$$

Therefore,

$$-E\left[\frac{\partial^2 \ell n p(x_i,t_i)}{\partial \ \alpha_j^2}\right] = - E\left[\frac{\partial^2 \ell n p(t_i)}{\partial \ \alpha_j}\right] \ .$$

Thus, it follows from Theorem 2.4.1 that for estimating α the joint sample (X_i,T_i), $i = 1,\ldots,n$ contains no more information than the sample $T_1,\ldots,T_n$ alone.

2.5 Information for α Contained in Classified Data.

Using the likelihood function given in (2.2.4) for a random sample defined in (2.2.2), it follows that

$$\ell n\, L_c = \sum_{i=1}^{m} N_i\, \ell n\, g_i$$

$$= \sum_{i=1}^{m-1} N_i\, \ell n\, g_i + \left(N - \sum_{i=1}^{m-1} N_i\right) \ell n \left[1 - \sum_{i=1}^{m-1} g_i\right]$$

since

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

Also, from (1.3.6) and $\sum_{i=1}^{m} \alpha_i = 1$ that

$$g_i = \sum_{j=1}^{m-1} \alpha_j \left[P(i|j) - P(i|m)\right] + P(i|m) \qquad (2.5.1)$$

and

$$\frac{\partial\, g_i}{\partial\, \alpha_j} = P(i|j) - P(i|m) . \qquad (2.5.2)$$

From (2.1.1) and $S_j^c = \dfrac{\partial\, \ell n\, L_c}{\partial\, \alpha_j}$ it follows that

$$S_j^c = \sum_{i=1}^{m} N_i\, \frac{1}{g_i} \left[P(i|j) - P(i|m)\right] \qquad (2.5.3)$$

or in matrix notation

$$S_c = [\Delta_{ij}]^T G^{-1} \bar{N} \tag{2.5.4}$$

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

$$\Delta^*_{ij} = P(i|j) - P(i|m) , \tag{2.5.5}$$

$$G = \begin{bmatrix} g_1 & 0 & 0 & \cdots & 0 \\ 0 & g_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & g_m \end{bmatrix} , \tag{2.5.6}$$

and

$$N = (N_1, N_2, \ldots, N_m)^T . \tag{2.5.7}$$

Note that by the Cramer-Rao theory the expected value of S_c is the zero vector which we will verify directly.

$$\begin{aligned} E[S_c] &= E[\Delta^*_{ij}]^T G^{-1} \bar{N} \\ &= [\Delta^*_{ij}]^T G^{-1} E[\bar{N}] \\ &= [\Delta^*_{ij}]^T G^{-1}(Ng) , \end{aligned} \tag{2.5.8}$$

where

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

or

$$g = GJ \tag{2.5.9}$$

where

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

It follows from

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

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

$$[\Delta^*_{ij}]J = \Phi \tag{2.5.10}$$

and in turn from (2.5.8) and (2.5.9) that

$$E[S_c] = N[\Delta^*_{ij}]\, G^{-1}\, GJ = \Phi \quad . \tag{2.5.11}$$

The covariance matrix $V(S_c)$ of S_c can now be computed using (2.5.4) and (2.5.11), that is

$$V(S_c) = [\Delta^*_{ij}]^T\, G^{-1}\, V(\bar{N})\, G^{-1}[\Delta^*_{ij}] \tag{2.5.12}$$

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

$$\begin{aligned} V(\bar{N}) &= N[G - GJJ^T G] \\ &= NG(I - JJ^T G) \\ &= N[G - P\alpha\alpha^T P] \end{aligned} \tag{2.5.13}$$

where

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

From (2.5.10), (2.5.12), and (2.5.13)

$$\Lambda_c(\alpha) = V(S_c) = N[\Delta^*_{ij}]^T G^{-1}[\Delta^*_{ij}] , \qquad (2.5.14)$$

the information for α contained in classified data.

For completeness we state the following theorem.

Theorem 2.5.1:

$$\Lambda_c(\alpha) \to \Lambda_v(\alpha) \quad \text{as} \quad P \to I$$

where

$$P = \{P(i|j)\} .$$

Proof: In matrix notation,

$$g = P\alpha .$$

Let $P \to I$, then $g \to \alpha$ and

$$\Delta^*_{ij} \to \Delta^v_{ij} = \begin{cases} 1 & \text{for } i = j \neq m \\ -1 & \text{for } i = m \\ 0 & \text{o.w.} \end{cases}$$

that is,

$$\Delta^*_{ij} \to \begin{pmatrix} I_{m-1} \\ \hline -J^T_{m-1} \end{pmatrix} \quad .$$

Note that (2.4.9) can be written as

$$\Lambda_v(\alpha) = [I_{m-1} \,\vdots\, -J_{m-1}][\mathrm{Diag}(\tfrac{1}{\alpha_1},\ldots,\tfrac{1}{\alpha_m})]\begin{pmatrix} I_{m-1} \\ \hline -J^T_{m-1} \end{pmatrix} \qquad (2.5.15)$$

where I_{m-1} is a $(m-1)\times(m-1)$ identity matrix and

$$-J_{m-1} = \underbrace{(-1,-1,\ldots,-1)}_{m-1}{}^T \quad .$$

Thus,

$$\Lambda_c(\alpha) = [\Delta^*_{ij}]^T\, G^{-1}[\Delta^*_{ij}] \quad [\Delta^v_{ij}]^T[\mathrm{diag}(\tfrac{1}{\alpha_1},\ldots,\tfrac{1}{\alpha_m})][\Delta^v_{ij}] = \Lambda_v(\alpha)$$

as $P \to I$.

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

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

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

$$g_1 = 1 - g_2\ ,$$

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

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

then

$$\Lambda_{11}^{c}(\alpha) = \frac{N[P(1|1)-P(1|2)]^2}{g_1 g_2} \quad . \tag{2.5.16}$$

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

$$\Lambda_c(\alpha) = \frac{N}{g_1 g_2} = \frac{N}{\alpha_1 \alpha_2} = \Lambda_v(\alpha) \ .$$

Note that for this case, $\Lambda_c^{-1}(\alpha)$ is the variance of a sufficient statistic $\hat{\alpha}_1 = \frac{N_1}{N}$ for α_1 in a binomial probability density function.

III. THE MAIN RESULT

3.1 The Ordering of the Information for α .

For the two population case (m=2) , the information for contained in unlabeled, verified and classified data are given respectively by

$$\Lambda_u(\alpha) = \frac{N(1-B)}{\alpha_1\alpha_2} \quad , \quad \text{where} \quad B = \int_{R^p} \frac{p_1(x)p_2(x)}{p(x)}\, dx \tag{3.1.1a}$$

$$\Lambda_v(\alpha) = \frac{N}{\alpha_1\alpha_2} \quad , \tag{3.1.1b}$$

and

$$\Lambda_c(\alpha) = \frac{N[P(1|1)-P(1|2)]^2}{g_1 g_2} \quad . \tag{3.1.1c}$$

The similarity of Λ_v , Λ_c and Λ_u is striking and one notes in this case an obvious ordering exists, that is

$$\Lambda_v(\alpha) \geq \Lambda_c(\alpha) \tag{3.1.2a}$$

and

$$\Lambda_v \geq \Lambda_u(\alpha) \quad . \tag{3.1.2b}$$

The inequality (3.1.2a) holds since

$$\Lambda_c(\alpha) = \frac{N}{g_1 g_2}\,[P(1|1) - P(1|2)]^2$$

$$= \frac{N[P(1|1) - P(1|2)]^2}{[\alpha_1 P(1|1) + \alpha_2 P(1|2)][1-g_1]} \quad .$$

However,

$$g_1 = \alpha_1 P(1\,1) + (1-\alpha_1)P(1|2)$$

$$g_2 = 1 - g_1$$

implies

$$g_1 g_2 = \alpha_1(1-\alpha_2)[P(1|1)-P(1|2))^2 + \frac{1}{\alpha_2} P(1|1)[1-P(1|1)]$$

$$+ \frac{1}{\alpha_1} P(1|2)[1-P(1|2)]] \quad .$$

Let

$$R_c = \frac{[P(1|1)-P(1|2)]^2}{\frac{g_1 g_2}{\alpha_1(1-\alpha_1)}} \quad . \tag{3.1.3}$$

Since $0 \leq R_c \leq 1$, one can conclude for $m = 2$, that

$$\Lambda_c(\alpha) = \frac{N}{\alpha_1(1-\alpha_1)}\ R_c$$

or

$$\Lambda_c(\alpha) \leq \frac{N}{\alpha_1(1-\alpha_1)} = \Lambda_v(\alpha) \quad .$$

From (2.6.1a) and the fact that

$$0 \leq R_u \overset{\text{def}}{=} 1 - B \leq 1 \tag{3.1.4}$$

implies that (3.1.2b) holds, that is, for $m=2$

$$\Lambda_u(\alpha) \leq \Lambda_v(\alpha) \ .$$

In this section, we will establish the following ordering of the information for α :

$$\Lambda_c(\alpha) \leq \Lambda_u(\alpha) \leq \Lambda_v(\alpha) \ . \tag{3.1.5}$$

(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".) This result will be given in a corollary to a Theorem proved by Rao [3].

Note that classified data defined in (2.2.2) is a explicit transformation of the unlabeled data. Knowing this, it follows directly from the following Theorem due to Rao [3] that

$$\Lambda_c(\alpha) \leq \Lambda_u(\alpha) \ .$$

Theorem 3.1.1 (Rao): The matrix $\Lambda_X - \Lambda_T$ is semi-positive definite, where Λ_T is the information matrix in a measureable function T of X.

The ordering between Λ_v with Λ_u and Λ_c is not as straightforward. The ordering (3.1.5) is proved in corollary 3.1.1 which will be proved very similarily to the proof of Theorem 3.1.1 once the following three lemmas are proved.

Suppose one takes an unlabeled sample and then classifies it, then let

$$Z = (X^T, Y(X)) \quad , \quad Y(X) = (Y_1(X), \ldots, Y_m(X))$$

when $Y_j(X) = 1,0$ if $x \in R_j$, $x \notin R_j$ respectively.

Lemma 3.1.1: The p.d.f. for Z is given by

$$p_Z(z) = \begin{cases} p_X(x) , & \text{if } X \in R_j \text{ and } y_j = 1 \text{ for some } j = 1,\ldots,m \\ 0 , & \text{o.w.} \end{cases} \tag{3.1.6}$$

Proof:

$$p_Z(z) = p(x,y)$$

$$= Pr(Y(x) = y | X=x)\ p_X(x)$$

Now (3.1.6) follows from

$$P_r(Y(X) = y \mid X=x) = \begin{cases} 1 & \text{if } X \in R_j \text{ and } y_j = 1 \text{ for some } j = 1,\ldots,m \\ 0 & \text{o.w.} \end{cases}$$

since $P_r(Y_j(x) = 1 \text{ and } Y_k(x) = 1) = 0$ for $j \neq k$.

Recall from Sections 2.3 - 2.5 that

$$S_u = \{S_j^u\} , \tag{3.1.7a}$$

$$S_v = \{S_j^v\} , \tag{3.1.7b}$$

$$S_c = \{S_j^c\} , \tag{3.1.7c}$$

for $j = 1,\ldots,m-1$

where

$$S_j^u = \frac{p_j(x) - p_m(x)}{p(x)} \tag{3.1.8a}$$

$$S_j^v = \frac{T_j}{\alpha_j} - \frac{T_m}{\alpha_m} \tag{3.1.8b}$$

$$S_j^c = \sum_{i=1}^{m} \frac{Y_i}{g_i} \Delta_{ij} \tag{3.1.8c}$$

for $j = 1,\ldots,m-1$.

Furthermore, we know that

$$E\, S_u = E\, S_v = E\, S_c = \Phi \tag{3.1.9}$$

Lemma 3.1.2:

$$\text{(i)} \quad E[S_u|Y=y] = S_c \tag{3.1.10a}$$

$$\text{(ii)} \quad E[S_v|X=x] = S_u \tag{3.1.10b}$$

$$\text{(iii)} \quad E[S_v|Y=y] = S_c \ . \tag{3.1.10c}$$

Proof:

(i) For each $j = 1,\ldots,m-1$, it follows from (3.1.8a) that

$$E[S_j^u|Y=y] = \int \frac{p_j(x)-p_m(x)}{p(x)} \cdot \frac{p(x,y)}{h(y)}\, dx \ .$$

Let

$$Y = y_{(k)} = (0,\ldots,0,1_k,0,\ldots,0)$$

where 1_k indicates that $y_k = 1$. Then it follows from Lemma 3.1.1 that

$$E[S_j^u|Y=y_{(k)}] = \int_{R_k} \frac{p_j(x)-p_m(x)}{p(x)} \frac{p(x)}{g_k} dx =$$

$$= \frac{1}{g_k} [P(k|j)-P(k|m)]$$

$$= \frac{\Delta_{kj}}{g_k} \quad .$$

(Note that $g_k = h(y_{(k)})$.)

Thus, in general we have

$$E[S_j^u|Y=y] = \sum_{k=1}^{m} \frac{y_k}{g_k} \Delta_{kj} = S_j^c \ , \ j = 1,2,\ldots,m-1 \ .$$

(ii) For each $j = 1,\ldots,m-1$, it follows from (3.1.8b) that

$$E[S_j^v|X=x] = \sum_{\{t|p(t|x)>0\}} \frac{t_j}{\alpha_j} - \frac{t_m}{\alpha_m} f(t|x)$$

$$= \frac{f(t_{(j)}|x)}{\alpha_j} - \frac{f(t_{(m)}|x)}{\alpha_m}$$

where

$$t_{(k)} = (0,\ldots,0\ ,1_k,0,\ldots,0) \ .$$

Note that $f(t|x) = \frac{f(t,x)}{p(x)} = \frac{p(x|t)f(t)}{p(x)}$.

Hence, it follows that

$$E[S_j^v|X = x] = \frac{\alpha_j p_j(x)}{\alpha_j p(x)} - \frac{\alpha_m p_m(x)}{\alpha_m p(x)}$$

$$= \frac{p_j(x)-p_m(x)}{p(x)} = S_j^u \text{ , for } j = 1,\ldots,m-2 .$$

(iii) Suppose $y = y_{(k)}$ for $k = 1,\ldots,m$, then for $j = 1,\ldots,m-1$ it follows from (3.1.8b) that

$$E[S_j^v|Y=y_{(k)}] = \frac{f(t_{(j)}|y_{(k)})}{\alpha_j} - \frac{f(t_{(m)}|y_{(k)})}{\alpha_m} .$$

It can be easily shown as follows:

$$f(t_{(j)}|y_{(t)} = \frac{f(t_{(j)},y_{(k)})}{h(y_{(k)})}$$

$$= \frac{h(y_{(k)}|t_{(j)})f(t_{(j)})}{h(y_{(k)})}$$

$$= \frac{Pr(Y_{(k)}=1|t_{(j)}=1)_j}{g_k}$$

$$= \frac{P(k|j)\alpha_j}{g_k}$$

$$= Q(j|k)$$

Thus,

$$E[S_j^v | Y=y_{(k)}] = \frac{Q(j|k)}{\alpha_j} - \frac{Q(m|k)}{\alpha_m}$$

$$= \frac{\alpha_j P(k|j)}{\alpha_j g_k} - \frac{\alpha_m P(k|m)}{\alpha_m g_k}$$

$$= \frac{1}{g_k}[P(k|j) - P(k|m)]$$

$$= \frac{\Delta_{kj}}{g_k} \quad .$$

In general, we have

$$E(S_j^v | Y=y) = \sum_{i=1}^{m} \frac{y_i \Delta_{ij}}{g_i} = S_j^c \quad , \quad \text{for} \quad j = 1,\ldots,m-1.$$

Lemma 3.1.3: (i) $E(S_c S_u^T) = \Lambda_c$

(ii) $E(S_u S_v^T) = \Lambda_u$

(iii) $E(S_c S_v^T) = \Lambda_c$.

Proof:

(i) $E(S_c S_u^T) = E\{E(S_c S_u^T | Y=y)\}$

$= E\{S_c E(S_u^T | Y=y)\}$.

It follows from Lemma (3.1.2) that

$= E\{S_c S_c^T\} = \Lambda_c$.

(ii) and (iii) are similarly proved.

Corollary 3.1.1:

(i) $\Lambda_u - \Lambda_c = D_1$

(ii) $\Lambda_v - \Lambda_u = D_2$

(iii) $\Lambda_v - \Lambda_c = D_3$

where D_1, D_2 and D_3 are positive semi-definite matrices.

Proof:

(i) Since $ES_c = ES_u = \Phi$, then by definition, the covariance matrix of $S_u - S_c$ is given by

$$E(S_u - S_c)(S_u - S_c)^T . \tag{3.1.11}$$

Now (3.1.11) can be written as

$$E(S_u S_u^T - S_u S_c^T + S_c S_u^T + S_c S_c^T) = ES_u S_u^T - ES_u S_c^T - ES_c S_c^T + ES_c S_c^T .$$

It follows from Lemma (3.1.3) that

$$E(S_u - S_c)(S_u - S_c)^T = \Lambda_u - \Lambda_c^T - \Lambda_c + \Lambda_c$$

$$= \Lambda_u - \Lambda_c^T$$

$$= \Lambda_u - \Lambda_c \text{ , since } \Lambda_c \text{ is symmetric.}$$

Since by definition, (3.1.11) is positive semi-definite, then $\Lambda_u - \Lambda_c$ is positive semi-definite.

(ii) and (iii) are similarly proved.

4. APPLICATION AND CONCLUSIONS

The central questions now include the following: Should one spend resources to verify data to gain information? Should one spend the allocated amount on verifying a small amount of data or process a large amount of unlabeled data? Is there any advantage at all to processing classified data.

4.1 Concerning Classified Data

In the space application the total data set is made up of unlabeled data which can be processed directly to obtain the true value of α or more realistically due to the magnitude of the set he sampled to estimate α. Let $\hat{g}_i = \sum_{v=1} Y_{ij}/N = N_j/N$ be an estimator α_j $j = 1,2,\ldots,m$, then since in general

$$E[\hat{g}_i] = \sum_{j=1}^{m} P(i|j)\alpha_j \neq \alpha_i \qquad (4.1.1)$$

it follows that if $\hat{g} = (\hat{g}_1,\hat{g}_2,\ldots,\hat{g}_m)$, then $\hat{g}$ is a biased estimator for α. In matrix notation

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

where $g = [Pr(x \in R_i)]$, which implies

$$\alpha = P^{-1}g \quad .$$

Note that if one defines

$$\hat{\alpha} = P^{-1}\hat{g}$$

that $E[\hat{\alpha}] = P^{-1}E(\hat{g}) = P^{-1}P\alpha = \alpha$ and $\hat{\alpha}$ is an unbiased estimator for α, when P is known. Unfortunately, the matrix P^{-1} is unknown; hence must be estimated. The sample used to estimate P^{-1} is called test data. There is bias in the estimator $\hat{\alpha} = \hat{P}^{-1}\hat{g}$ when P^{-1} is replaced by $(P^{-1}) = (\hat{P})^{-1}$, hence $\hat{\alpha}$ will be biased.

Note also that in (4.1.1) it has been assumed that μ_i and Σ_i are known when in fact they are not known but must be estimated. The sample for estimating these parameters are called the training data (the data to "train" a classifier).

One must also select a classification rule. Two candidates naturally are candidates. The Bayes classification procedure and the maximum likelihood procedure. The Bayes classifier is optimal with respect to minimizes the expected costs of misclassification but unfortunately is a function of the elements of α hence in practice cannot be used. The analysis and results in this paper are not dependent on the type of classifier used.

In Table 4.1 the values of information for various values of α_1 when $m = 2$ and $n = 1$ as function of type of classifers and for various distance between the subpopulation $p_1(x)$ and $p_2(x)$ each assumed to be normal, hence $p(x) = \alpha_1 p_1(x) + (1-\alpha)p_2(x)$ is a mixture of two normals ($\Delta = \mu_1 - \mu_2$ and $\Sigma_1 = \Sigma_2$ = the identity matrix). The symbols Λ_B and Λ_{MLE} denote the information using a Bayes classifier and the maximum likelihood classifier, respectively; Λ_v is information using verified data.

Table 4.1. Approximate Values of Λ_u, Λ_B, Λ_{MLE} in a Mixture of Normals

Δ / α_1	$\Delta = 1/4$			$\Delta = 1/3$			$\Delta = 1/2$			$\Delta = 1$			$\Delta = 2$			$\Delta = 3$			$\Delta = \infty$
	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_u	Λ_B	Λ_{MLE}	Λ_v
.1	.063	.000	.040	.114	.000	.071	.263	.000	.160	1.15	.226	.647	4.56	3.36	2.66	7.98	7.21	5.78	11.1
.2	.063	.000	.040	.111	.000	.071	.250	.010	.158	.967	.435	.619	3.06	2.46	2.24	4.80	4.43	4.11	6.25
.3	.062	.000	.040	.110	.006	.070	.242	.060	.157	.875	.533	.601	2.51	2.08	2.01	3.76	3.50	3.41	4.76
.4	.062	.015	.040	.108	.042	.070	.237	.127	.156	.830	.575	.590	2.27	1.91	1.90	3.33	3.11	3.10	4.17
.5	.062	.040	.040	.108	.070	.070	.236	.156	.156	.816	.587	.587	2.20	1.86	1.85	3.21	3.00	3.00	4.00

$\Delta = (\mu_2 - \mu_1)/\sigma$, $(\sigma = 1)$

Λ_u = Unlabeled Information

Λ_B = Bayes Classified Information

Λ_{MLE} = MLE Classified Information

In Table 4.2 values of information are given for various values of Δ, k and α_1 when $\sigma_2^2 = k\sigma_1^2$ and p(x) is a mixture of two univariate normal p.d.f. The value selected for $\sigma_1^2 = 1$ and $n = 1$.

4.2 Conclusions

The surprising result that classified data has the least information is especially important since current practice in processing remote sensed data is to classify the unlabeled data. It is true that it may be easier to classify than compute the maximum likelihood estimates for α using unlabeled data. Hence classifying the data would be a necessary task. The information in classfied data is nearly equal to but always less than the information in unlabeled data.

Note also, if the expense to verify data is sufficiently small then the unlabeled data taken remotely from sapce is not needed. A random sample of locations on the earth's surface is sufficient to estimate α. The satellite data is of no value except to randomly select sites for verification.

If training data and test data are in reality classified data (that is, unlabeled data classified by photo interpreters) one can and should expect loss of information. However, training data and test data are in fact verified or labeled (ground truth with no classification error) one should expect better results in estimating α.

Table 4.2. Information Λ for Various Types of Data (v,u,c) Versus Values of the Parameters (k,Δ,α_1).

α_1	Type of Data	k = 1			k = 2		
		Δ = 1	2	3	1	2	3
0.1	v	11.11	11.11	11.11	11.11	11.11	11.11
	u	1.15	4.57	7.98	0.60	2.38	5.51
	c	0.65	2.66	5.78	0.47	1.68	3.79
0.3	v	4.76	4.76	4.76	4.76	4.76	4.76
	u	0.88	2.51	3.76	0.62	1.81	3.09
	c	0.60	2.01	3.41	0.48	1.48	2.69
0.5	v	4.00	4.00	4.00	4.00	4.00	4.00
	u	0.82	2.20	3.21	0.68	1.77	2.77
	c	0.59	1.86	3.00	0.61	1.47	2.50