BSR 2949

SIGNATURE DATA PROCESSING STUDY FINAL REPORT

Volume I
Summary

By:
Cheryl L. Crawford
Robert H. Dye
David S. Hanson

A-5 — A-6
missing in orig.



August 1970

Distribution of this report is provided in the interest of information exchange and should not be construed as endorsement by NASA of the material presented. Responsibility for the contents resides with the organization that prepared it.

Prepared Under Contract No. NAS 9-9848 by
THE BENDIX CORPORATION AEROSPACE SYSTEMS DIVISION
Ann Arbor, Michigan

Manned Spacecraft Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

FACILITY FORM 602

N71-12508 (ACCESSION NUMBER) — (THRU)
213 (PAGES) — 63 (CODE)
CR-108760 (NASA CR OR TMX OR AD NUMBER) — 08 (CATEGORY)

CONTENTS

		Page
1.	INTRODUCTION	1-1
2.	TECHNICAL APPROACH	2-1
	2.1 DEFINITION OF STUDY OBJECTIVES	2-1
	2.2 DECISION RULES	2-4
3.	STUDY PLAN	3-1
	3.1 MATHEMATICAL METHODS	3-2
	3.1.1 Principal Component Analysis (Factor Analysis)	3-4
	3.1.2 Bayesian Decision Processor	3-8
	3.1.3 Information Divergence	3-10
	3.1.4 Probability Density Functions	3-12
	3.1.5 Assembly of Methods	3-15
	3.2 HARDWARE CONCEPTS	3-15
	3.3 STUDY REVIEW	3-19
4.	PARAMETRIC STUDIES	4-1
	4.1 DESCRIPTION OF SIMULATED DATA SETS	4-3
	4.2 PRELIMINARY ANALYSIS OF SIMULATED DATA	4-8
	4.3 SAMPLE SMOOTHING	4-17
	4.3.1 Specialized Applications of Divergence	4-21
	4.3.2 Investigation of Smoothing Parameters	4-22
5.	REAL-WORLD STUDIES	5-1
	5.1 PILOT DATA SETS	5-2
	5.2 PRINCIPAL COMPONENT TRANSFORMATIONS (FACTOR ANALYSIS)	5-14
	5.2.1 Common Transformations	5-16
	5.2.2 Multiple Transformations	5-31
	5.3 CANONICAL TRANSFORMATIONS	5-86
6.	HARDWARE DEVELOPMENT	6-1
	6.1 HARDWARE DESCRIPTION	6-3

PRECEDING PAGE BLANK NOT FILMED

CONTENTS (CONT.)

	Page
6.2 SYSTEM PERFORMANCE	6-11
6.2.1 Imagery	6-12
6.2.2 Smoothing	6-16
7. CONCLUSIONS	7-1
APPENDIX A CLUSTER ANALYSIS	A-1

ILLUSTRATIONS

Figure	Title	Page
2-1	False Color Classification of Multispectral Data From Computer Analysis	2-9
2-2	Partition of a Two-Space Into Three Category Types by Theorems 1 and 2	2-12
3-1	Flow of Information Through Signature Data Processing	3-16
3-2	Functional Schematic of Analog Table Look-Up Device	3-18
4-1	Flow of Simulated Data	4-2
4-2	Scatter Diagrams, SDPS10	4-4
4-3	Training Set Groups of SDPS10	4-5/4-6
4-4	Simulated Data Scatter Diagrams	4-9/4-10
4-5	Signatures Misclassified as a Function of Within-Cluster Variance (Fixed Cluster Means)	4-11
4-6	Divergence as a Function of Within-Cluster Variance (Fixed Cluster Means)	4-12
4-7	Divergence and Number of Misclassifications as a Function of Number of Principal Components (Factor Scores) Used to Represent Spectral Signatures, SDPS10	4-16
4-8	Bayesian Decision Function Program Output	4-18
4-9	Percentage Error of Misclassification vs. Smoothing Parameters as Proportion of Within-Cluster Variance	4-24
4-10	Correlated Data Scatter Diagram	4-25
4-11	Factor Score Data Scatter Diagram	4-27
4-12	Sample Divergence as a Function of Sample Size	4-29
4-13	Density Divergence as a Function of Sample Size	4-31
5-1	EMSIDE Channel Locations vs. Color Band	5-3
5-2	Multispectral Scanner Pictorial Layout	5-4
5-3	Multispectral Scanner Airborne System	5-5
5-4	Multispectral Scanner, Cover Removed	5-6
5-5	Digital Sampling Imagery	5-10
5-6	Scatter Diagrams of Transformed Cotton	5-17
5-7	Scatter Diagrams of Transformed Sorghum	5-18
5-8	Scatter Diagrams of Transformed Citrus	5-19
5-9	Scatter Diagrams of Transformed Bare Soil	5-20
5-10	Scatter Diagrams of Training Sets	5-25/5-26
5-11	Covariance Matrices of Training Sets	5-30

ILLUSTRATIONS (CONT.)

Figure	Title	Page
5-12	Factor Analysis of Citrus	5-33
5-13	Factor Analysis of Cotton	5-34
5-14	Factor Analysis of Sorghum	5-35
5-15	Factor Analysis of Bare Soil	5-36
5-16	Scatter Diagrams of Crops and Their Backgrounds	5-39/5-46
5-17	Scatter Diagrams 2000 Factor Scores of Four Crops 1 vs. 2	5-47
5-18	Sample Divergence of Similar Categories	5-49
5-19	Marginal Density for Citrus	5-50
5-20	Marginal Density for Citrus	5-51
5-21	Summary of Two-Sample Divergence with Smoothing Parameters	5-53
5-22	Marginal Density for Sorghum	5-55
5-23	Marginal Density for Sorghum	5-56
5-24	Marginal Density for Bare Soil	5-57
5-25	Marginal Density for Cotton	5-58
5-26	Distribution of f (x Sorghum) for Samples of N = 600	5-60
5-27	Distribution of f (x Citrus) for Samples of N = 600	5-62
5-28	Distribution of f (x Bare Soil) for Samples of N = 600	5-63
5-29	Distribution of f (x Cotton) for Samples of N = 600	5-64
5-30	Distribution of Citrus Densities	5-72
5-31	Distribution of Cotton Densities	5-73
5-32	Distribution of Sorghum Densities	5-74
5-33	Distribution of Bare Soil Densities	5-75
5-34	Distribution of Citrus Densities; All Eight Factors	5-77
5-35	Distribution of Cotton Densities; All Eight Factors	5-78
5-36	Distribution of Sorghum Densities; All Eight Factors	5-79
5-37	Distribution of Bare Soil Densities; All Eight Factors	5-80
5-38	First Four Canonical Variables; Citrus	5-89/5-90
5-39	First Four Canonical Variables; Cotton	5-91/5-92
5-40	First Four Canonical Variables; Sorghum	5-93/5-94
5-41	First Four Canonical Variables; Bare Soil	5-95/5-96
5-42	Last Four Canonical Variables; Bare Soil	5-97/5-98
6-1	Block Diagram of Analog Table Look-Up Device	6-2
6-2	Three Views of Flying-Spot Scanner	6-4
6-3	Front View of Transformation Network Installed and Flying-Spot Scanner	6-5
6-4	Functional Schematic of Analog Table Look-Up Device	6-6
6-5	Transformation Network	6-7
6-6	Analog Processor and Calibration Electronics	6-9
6-7	Flying-Spot Scanner Electronics	6-10

ILLUSTRATIONS (CONT.)

Figure	Title	Page
6-8	Probability Density Imagery Produced from Analog Table Look-Up Device	6-13
6-9	Scatter Diagrams 2000 Factor Scores of Four Crops 1 vs. 2	6-14
6-10	Distribution of f (x|citrus) for Samples of N = 600	6-15
6-11	Histogram Prepared From Bivariate Density for Citrus	6-17
6-12	Smoothing Effect of Flying-Spot Scanner	6-18
A-1	Scatter Diagrams of Cluster Analyzed Data	A-6

TABLES

Table	Title	Page
4-1	Simulated Data Parameters, SDPS10	4-7
4-2	Expected Classification Results (Truth Tables)	4-7
4-3	Bayesian Decisions and Information Divergence, SDPS10	4-13
5-1	Performance Characteristics	5-7
5-2	Raw Data Classification, 8-Channel Specht Function	5-22
5-3	Eight-Factor Multivariate Classification Results	5-27
5-4	Four-Factor Multivariate Normal Classification and Divergence Results	5-28
5-5	Two-Sample Divergence for Sorghum	5-52
5-6	Classification Results	5-65
5-7	Expected Classification Results	5-65
5-8	Information Divergence—Samples of 600 Bits of Information	5-66
5-9	Information-to-Clutter Ratios	5-69
5-10	Computations With Last Four Principal Components Only	5-71
5-11	Computations with All Eight Principal Components	5-81
5-12	Information-to-Clutter Ratios—Last Four Principal Components	5-84
5-13	Information-to-Clutter Ratio—All Eight Principal Components	5-84
5-14	Eigenvalues Found in Canonical Analysis with Bare Soil as the Target Category	5-99
5-15	Canonical Analysis Eigenvalues	5-100

PRECEDING PAGE BLANK NOT FILMED

SECTION 1

INTRODUCTION

The scope of the studies described in this three-volume report is intended to serve as a contribution to the development of the hardware and software facilities necessary to accomplish automatically the extraction of the spectral information from line scanner multispectral image data. The applicability of the theoretical concepts included is not restricted to the study of data from multispectral line scanners, however. The data format produced by such instruments is unusually well suited to the development of techniques for extraction of spectral information. The use of such data at the existing level of capability in the general problem allows the investigator to work directly on the portions of the problem requiring the development of new concepts which can cope with high data rates and unprecedentedly large amounts of information. The favorable evaluation of concepts which can handle data from multispectral line scanners, and reduce to practice the extraction of spectral information therefrom, will extend the same practical benefits to the processing of multispectral data from most other image forming instruments as well. The only intermediate steps necessary will be to transform the data from other types of instruments into the format achieved directly with a multispectral line scanner.

The concepts developed and partially evaluated during the conduct of this study have been carried out with eight-channel multispectral line scanner video produced by the Bendix EMSIDE I scanner. The airborne equipment is illustrated in Section 5. 1. The end products of the study, contained in the three volumes of this report, consist of the following:

1. Volume I: a report on the study activities and findings

2. Volume II: a description of the generally applicable software used to conduct the study; included are analytical descriptions of equations evaluated by the various programs and flow diagrams which indicate how the computations are actually performed

3. Volume III: engineering specifications which describe the steps which, in our judgment, would result in the development of a Signature Data Processing System (SDPS).

The system described is not quite a stand-alone system. It has been devised to interface with the Ground Data Station (GDS) currently in process for use with data collected by a 24-channel airborne multispectral line scanner. The data station and the line scanner together comprise the Multispectral Data System (MSDS). Bendix is producing this system for NASA/MSC under Contract NAS 9-9089. The SDPS interface with the MSDS-GDS allows the former to possess the highly refined data handling, editing, and display capability of the latter without duplication of equipment and cost. The data input for the SDPS is taken from the GDS. Processing techniques desired for evaluation are applied in the SDPS; the result of applying these techniques will produce a magnetic tape as an output product. The magnetic tape will be suitable for use as input data to the GDS; when the tape is so used, the display obtained will be processed imagery which maps the spatial distribution of the desired categories based on the spectral information extracted from the raw data.

The SDPS for which specifications are submitted in Volume III should be regarded as a research tool. The problem, as defined in the next section, is not the theoretical one of how to extract information, but practical ones. The practical problems are how to store enough information to permit reliable recognition of information in the data, how to execute the necessary comparisons quickly, and how to determine whether a satisfactory job is performed.

The system specified would provide a research tool of sufficiently general flexibility and speed to permit accrual of practical user-oriented experience with a great many different techniques for evaluation. The specifications are functionally oriented. Rapidly improving state of the art in high-speed digital arithmetic electronics and low noise, wide bandwidth analog electronics make some of the alternative methods of implementing the necessary functions very difficult to select. Part of the development will require the conduct of trade-off analyses based on the then current state of the art.

The balance of Volume I is devoted to a description of the study activities conducted with the analytical tools implemented in the form of computer programs. The Problem Definition in Section 2 is followed by a description of the Study Plan in Section 3. This section is concluded with a review of the significant but unanticipated aspects of the study. Sections 4 and 5 describe the studies performed on simulated data, the Parametric Studies, and on Real World Data. Section 5.3, Canonical Transformations, concludes the study of Real World Data with a description of a promising approach to the problem of dimensionality reduction. Conclusions drawn from the study activities are presented in Section 7.

SECTION 2

TECHNICAL APPROACH

2.1 DEFINITION OF STUDY OBJECTIVES

Automatic processing is intended to ease the burden placed on the interpreter of multispectral imagery. In the absence of automatic processing, the interpretation of imagery from a 24-channel scanner would require the inspection of 24 sets of imagery with each showing the same spatial area. To ask the interpreter to deal objectively with the strictly spectral information that might be available to him, while simultaneously presented with a great deal of redundant spatial information much more easily interpreted than spectral information, is asking too much. If the interpreter makes use of the spectral information in the imagery at all, he is very likely to look for the broad categorical differences which have played a historical role in motivating the development of remote sensing instrumentation. He will look for thermal differences in the infrared (IR), evidence of moisture content in the near IR, reflectivity differences in the visible, and normally unobservable reflectivity differences in the ultraviolet. If multispectral imagery were to be interpreted manually, the usable spectral resolution would probably require far less than provided on a multispectral scanner.

The study of multispectral signature data is carried out with the objective of classifying each observation by type. One does not intend to measure a continuously variable parameter of a target category. Temperature is such a measurable parameter and might well be one of the components of multispectral observation. The classifications desired are the kinds of material which produced the observation such as corn, oats, granite, quartz, grassland, or forest. Multispectral data processing can be undertaken to produce enhanced imagery which is more easily interpreted by the user and which relieves him of the burden of inspecting a great many images, or to produce decision imagery in which case the machine makes the final categorical decision. Enhanced imagery is normally prepared with a continuous gray scale formed to sharpen the contrast between target categories which have characteristic spectral differences expressable on a single scale. Decision imagery has distinct gray levels or colors associated with each target category. The distinction, not a generally recognized one, is useful to emphasize an important difference in the type of classifications achievable with automatic methods of processing multispectral data.

The maximum likelihood approach to signature recognition can be demonstrated to provide optimum recognition. In this case, the use of the word optimum means that all spectral information available is used to assign new observations to the member of a set of alternative populations sampled earlier, which will be correct with highest probability, will minimize cost, or, with highest probability, will place an extreme value on any other quantifiable consequence of the classification.

Decision imagery is produced by applying Bayesian decision theory to observations, in the process comparing them with the signature distributions of reference data, and displaying the decision alone. This is done on a printer-plotter by associating a character with each category or on a color film recorder by associating a color with each category. In neither case is the relative confidence, as measured by the conditional probability used to make the decision, transmitted to the user via the image display.

Enhanced imagery transmits gray scales but cannot separate well, more than a few major categories; decision imagery transmits only a result obtained by using all the available information, but not the information used to obtain the result.

Ideal processed imagery, which the current study undertook to make possible, combines the best features of both types. Given a set of features or targets with well known spectral signature distributions, the ideal presentation of processed results in a single display would show the classification of new observations into categories described by the known distributions and transmit to the user the relative confidence of each decision. This can be accomplished by making classifications based on a Bayesian decision strategy and computing the conditional a posteriori probability for each classification under the hypothesis selected. The output of such an ideal processor is a time-varying classification vector, with one component for each permissible feature alternative. The output vector is precisely analogous to the input which may be regarded as a time-varying observation vector with one component for each spectral channel.

The classification vector is a continuous time stream of conditional probabilities which can be displayed in a video format just as the input vector. A single display can be devised to accommodate all components of the classification vector. Each component is identified with a color; thus, channel 1, corresponding to feature alternative A, is encoded to make a red spot on a color display; channel 2, feature alternative B, is encoded to make a green spot. The number of colors usable in this way is limited only by the minimum differences in hue distinguishable by the human eye.

As the display device sweeps the image plane to make the display, the color displayed is that determined by decision rule such as described in the following section. This result is ordinary decision imagery. The saturation of the selected color, however, is controlled by the magnitude of the conditional probability computed for the category selected. This causes the display to show the confidence that should be placed in the decision. The saturation of the selected hue varies between zero (white) and one (completely saturated) as the conditional probability is very low or very high.

This procedure is conceptually simple when employing digital computing methods. The probability can be computed separately for each observation and for each category considered. The output of this processing scheme is a vector with one component for each of the alternatives considered. The value of each component is the probability of the corresponding alternative. This can be turned into a display by directing the multicomponent output to a color film recorder such as that provided in the ground data station. A distinct color for each alternative, modulated in intensity from saturation to white as the probability of correctness is large or small, provides a display which makes all the information available to the interpreter.

Two problems are raised when Bayesian procedure is applied to multivariate observations from imaging sensors. The amount of information that must be stored to represent adequately the multivariate signature distribution for a single target increases exponentially with the number of channels because readings on the channels are not statistically independent. Thus, a multispectral scanner with 24 channels and only 100 resolvable levels on each channel would require 10^{48} words to store the signature distribution from a single target. Such a requirement exceeds the storage capacity of any reasonably sized computer. Signature distributions in this form cannot therefore be stored digitally in the core portion of the computer memory. Other methods of digital storage for this information are equally unacceptable and also have prohibitively long access times.

The second problem concerns the necessary processing rate. Imaging sensors collect observations at rates on the order of 10^5 per second. A proper job of classification requires that each of these observations be compared with signature distribution for each of the alternatives among which classification is desired. Digital methods can be used for signature recognition only if the information storage requirements can be reduced to acceptable levels. When the information storage requirement is so reduced, digital processing can be completed only at a rate much slower than the rate at which the data are provided. When analog processing techniques are employed, the rate at which data can be processed is acceptable, but the information required to apply decision rules and compute conditional probabilities is passed to the processing device at the expense of much additional analog circuitry.

The basic objectives of this study were:

1. To explore methods for minimizing the amount of information that must be stored to represent multispectral signature distributions, and to develop quantitative measures of the relative merits of various methods

2. To explore methods for maximizing the amount of information that can be stored

3. To evaluate the feasibility of analog methods for implementing the analytical methods explored to permit real data rate processing with analog equipment.

The scope of these objectives is extremely broad. To approach them in a manner which does not permit the illusion of progress to be represented as real progress is to conduct painfully slow research. Samples of a population are used as the ultimate representation of a population. This choice guarantees that assumptions concerning the distribution of the population will not influence the results obtained. The use of samples in this way, however, imposes a number of other problems which must themselves be studied before proceeding with the direct objectives.

2.2 DECISION RULES

Many different objective criteria have been postulated to initiate series of analytical deductions to specify rules for decision and classification of multivariate observations. This section concludes with a partial bibliography which cites references describing many such developments and ways of looking at the problem of multivariate decisions.

Decision rules devised in this manner always require knowledge of, or an assumption concerning, the nature of two or more multivariate distribution functions. The rules do not normally differ in important ways. Some permit the assignment of relative cost of errors and minimize net cost; some use the same information to minimize the maximum risk. In any case, the assignment of a priori probabilities and relative costs of various errors to the possible decisions, while necessary, is a user judgment problem which has no serious impact on the difficulty of the basic decision problem.

The peripheral trimming which often accompanies the development of decision rules is eliminated in the concise statement of three theorems proved by Rao.* These are quoted without proof but not without comment. The theorems are stated for decisions involving three categories; the generalization required for more categories is obvious. Regions, R_i, in multivariate space within the boundaries of which the best (most probably correct) decision is category i, are defined. The probability density function, $f_i(\vec{x})$, describes the distribution of observations, $\vec{x}$, drawn from the population of category i. The a priori probability of category i observations is P_i.

Theorem 1

The regions defined by ($\cap$):

$$R_1 \cap P_1 f_1 \geq P_2 f_2,\ P_1 f_1 \geq P_3 f_3$$

$$R_2 \cap P_2 f_2 \geq P_1 f_1,\ P_2 f_2 \geq P_3 f_3$$

$$R_3 \cap P_3 f_3 \geq P_1 f_1,\ P_3 f_3 \geq P_2 f_2$$

constitute the best possible system of mutually exclusive regions.

Such mutually exclusive regions are also exhaustive in the sense of assigning every point in the space to a region, R_i. In the event of equally probable categories, the best regions are:

$$R_i \cap f_i \geq f_j,\ j \neq i.$$

These regions may be used for classifying an observation when nothing is known about the P_i. "This is the maximum likelihood method in the problem of classification We choose that hypothesis for which the likelihood is a maximum."

A second theorem is developed to define regions, C_i, which maximize the probability of a correct decision subject to a maximum permissible probability of an incorrect decision. These regions can also be mutually exclusive but are not necessarily exhaustive of the space.

*C. Radhakrishna Rao, Advanced Statistical Methods in Biometric Research, John Wiley & Son, Inc., New York, 1952.

Theorem 2

Region C_1 satisfying the condition

$$\int_{C_1} f_1 \, d\vec{x} \text{ is a maximum subject to the restrictions}$$

$$Q \geq \int_{C_1} f_2 \, d\vec{x}$$

and $$Q \geq \int_{C_1} f_3 \, d\vec{x}$$

is defined by

$$f_1 \geq a f_2 + b f_3$$

where a and b are suitably chosen.

Similarly, regions C_i, each maximizing the probability of correct assignment to category i subject to a maximum probability Q of assignment to each of the other categories $j \neq i$, are defined by:

$$f_i \geq \sum_{j \neq i} a_j f_j .$$

Furthermore, if the a priori probabilities P_i are known, a single constant can be chosen such that the region C_i is defined by:

$$f_i \geq a \sum_{j \neq i} P_j f_j .$$

The practical implication of this result is that, given the ability to compute the density functions, f_j, for each new observation, $\vec{x}$, it can be determined whether the observation lies within a region C_i by evaluating a simple linear combination of the probability densities f_i for all j. The determination permits the degree of certainty to be retained. The linear combination may be selected to determine a division of multivariate space into mutually exclusive regions which limit the probability of incorrect decisions.

A third theorem describes a method for the determination of mutually exclusive, exhaustive regions such that the errors for each group are in an assigned ratio.

Theorem 3

The system of regions:

$$R_1 \cap a f_j \geq b f_2,\ a f_1 \geq c f_3$$

$$R_2 \cap b f_2 \geq c f_3,\ b f_2 \geq a f_1$$

$$R_3 \cap c f_3 \geq a f_1,\ c f_3 \geq b f_2$$

where a, b, and c are suitably chosen are the best possible if the errors of classification for the three groups are to be in an assigned ratio.

Rao's proofs describe methods for suitably choosing the constants a and b of Theorem 2, and a, b, and c of Theorem 3.

The point to be made here is that knowledge of the probability density functions permits the best possible classification to be accomplished by applying the decision rules developed in these theorems.

The thrust of the Signature Data Processing (SDP) Study was therefore directed to methods of representing general probability density functions and making them easily and rapidly available for the simple computations necessary to complete the desired classification. The difficulty of representing general density functions, i.e., representation of functions which cannot be easily described with parametric analytic functions, is the justification for the commonly applied assumption of multivariate normality. Study of solutions to this difficulty involves two parallel lines of activity; both were initiated during this study and are continuing. Analytical and computer studies of real data distributions are conducted

to learn about the properties of real distributions, how to represent them, and the consequences of using representations which are less general than population samples. The parallel activity to design, fabricate, and evaluate breadboard hardware for implementing decision processing was also carried to the extent necessary to define development specifications for an SDP System. The technical approach to the two activities is described in the following sections.

Before proceeding with a discussion of these activities, however, consider the nature of a display obtained by the decision rules of Theorems 1, 2, and 3. The rule described by Theorem 1, when applied to a three-category decision problem, can give three possible answers. Each observation tested by this rule will be assigned to one of the three categories by division of the space into mutually exclusive and exhaustive regions. Display of these results could take the form of a three-color map or the equivalent generated on a computer printer substituting characters for colors. More colors or characters are required to map larger numbers of categories. A three-color map of this type, produced during work supported by the Geographic Sciences Division, US Army Topographic Laboratories, Ft. Belvoir, Virginia, is illustrated in Figure 2-1.

From the user point of view, such a display is defective because it does not distinguish between regions in which the alternative categories are almost equally probable but one is selected by the decision rule, and regions in which the probability of erroneous assignment to the selected category is small. This is the problem addressed by Theorems 2 and 3. Theorem 3 modifies the mutually exclusive and exhaustive regions in such a way that the probabilities of error for the categories are in assignable ratios. While a decision rule of this kind is useful for many problems, Theorem 2 provides the best illustration of the flexibility desirable in a decision rule.

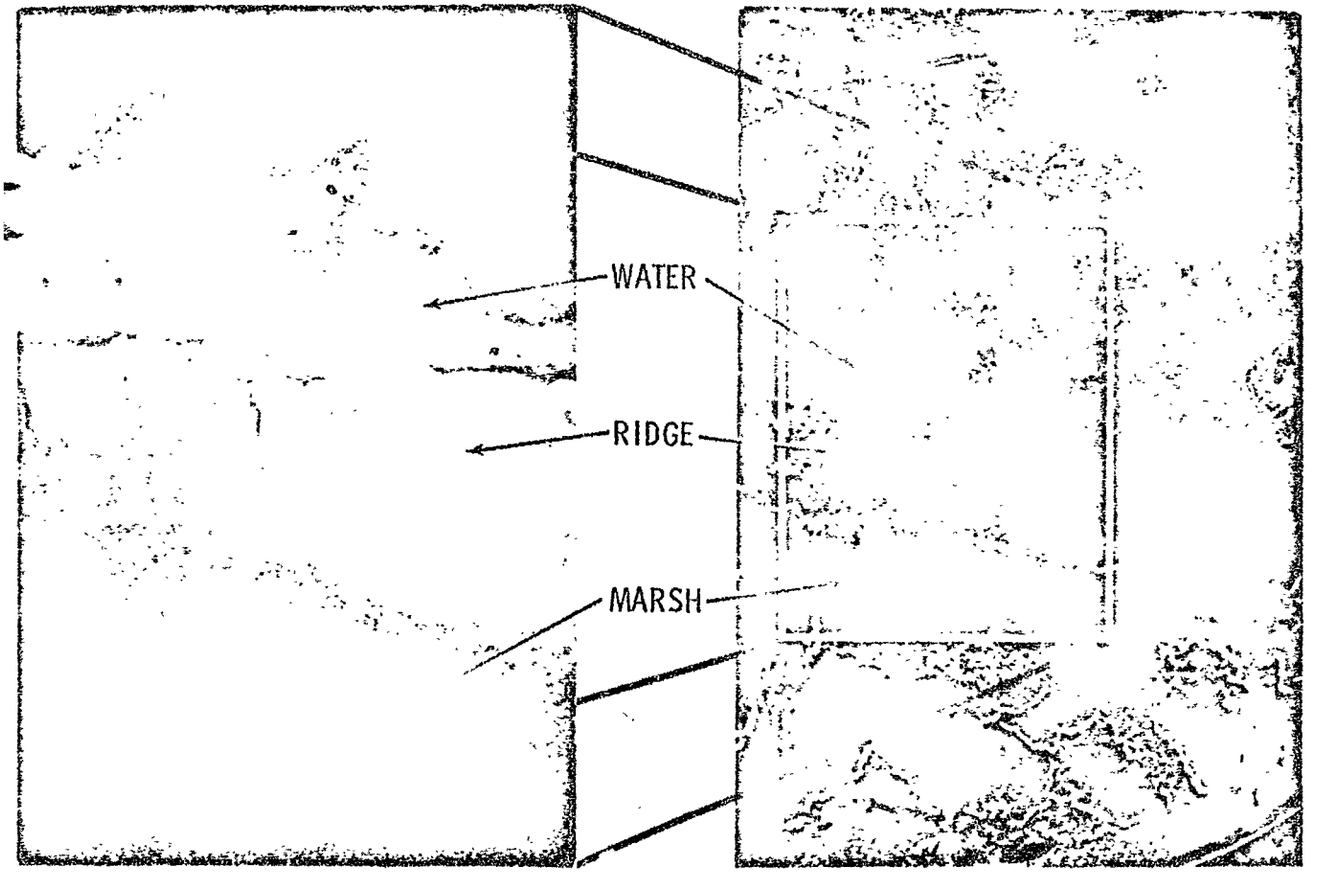
The regions into which space is divided for the purpose of classifying observations need not be exhaustive or mutually exclusive for that matter. The decision rule that region C_i is defined by

$$f_i \geq \sum_{j \neq i} a_j f_j$$

produces a region C_i within which the probability

$$Q_i \geq \int_{C_i} f_j \, d\vec{x}$$

False Color Classification of Multispectral Data From Computer Analysis



THIS BEACH RIDGE IS A POTENTIAL LOCATION OF SAND AND GRAVEL DEPOSITS. MAPPING OF POTENTIAL DEPOSITS SIMPLIFIES THE SEARCH FOR ENGINEERING CONSTRUCTION MATERIALS. THE PHOTOGRAPH WAS MADE ON AERIAL EKTACHROME FILM NEAR LAKE CHARLES, LOUISIANA AT AN ALTITUDE OF 2000 FEET ON MARCH 24, 1969.

THE COLOR CODED IMAGE IS BASED ON COMPUTER ANALYSIS OF MULTISPECTRAL SCANNER VIDEO OF THE SAME LOCATION. THE CLASSIFICATION OF MARSH, WATER, AND BEACH RIDGE IS PERFORMED AUTOMATICALLY TO MAP TARGETS OF INTEREST OVER EXTENDED AREAS.

NOTE: A color print not included in this copy of the report.

Figure 2-1

for all $j \neq i$, when the a_j are suitably chosen. Q may be chosen small enough that any region C_i shrinks to an arbitrarily small size. Q may also be chosen so large that the regions C_i overlap. A reasonable selection of Q for a three-category problem in two variables might produce regions

$$C_i = R_i - D_i$$

as illustrated in Figure 2-2. The regions $R_i = C_i + D_i$ are those defined by the rule of Theorem 1. The regions D_i are excluded by the rule of Theorem 2 for violation of the conditions on the probability of error.

Regions comparable to the D_i may be found on the three-color map of Figure 2-1. The white areas along the boundaries between colors are observations assigned to none of the three decision regions represented by a color. Decision is withheld.

Black areas may also be found on the map. These in fact appear only because of the darkroom techniques used to produce the three-color map. Black and white maps in transparency form were made separately and superimposed in the darkroom to produce the three-color version. The black areas could have been produced, however, by overlapping decision regions C_i. This could be desirable and achieved by selecting sufficiently large Q_i to define C_i. The regions C_i in this case are not mutually exclusive.

Selection of decision regions in this way - such that decisions are withheld unless clearly justified - transmits more information than simple categorical assignment on the basis of maximum likelihood, i.e., the Theorem 1 rule.

The basic defect is still not removed, however. The objection that some observations which satisfy the maximum likelihood criterion and are assigned to the appropriate category may not be clear-cut cases of an observation from that category is answered only by creating another category against which the same objection may be raised.

The process could be continued by creating another region within which the degree of uncertainty is greater still, and others within which the certainty of the decision given is higher. The basic defect is still there because of the attempt to display a continuous variable with a mechanism permitting the display of discrete variables only.

The decision rule of Theorem 2 states a concise test for deciding whether a given observation falls within a region defined to favor correct decisions and limit

PRECEDING PAGE BLANK NOT FILMED

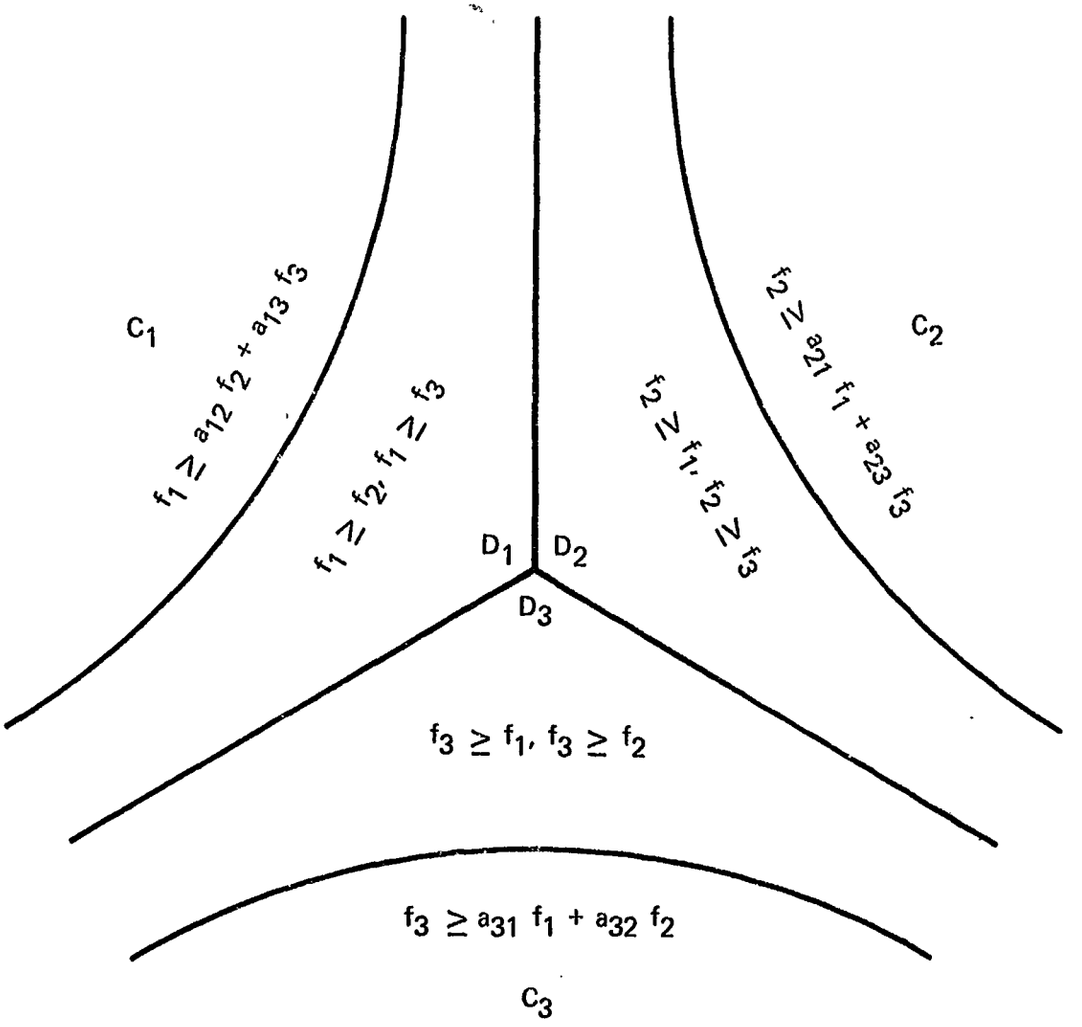


Figure 2-2 Partition of a Two-Space Into Three Category Types by Theorems 1 and 2

the number of errors. Bayes' Theorem on the computation of conditional a posteriori probabilities describes a method for determining the probability that a given observation is assigned to the correct category. The difference is important in terms of identifying a variable suitable for display purposes. If an observation is assigned to region C_i, the Bayesian conditional a posteriori probability that the assignment is correct is:

$$P(\vec{X} / C_i) = \frac{P_i f_i}{p_i f_i + \sum_{j \neq i} p_j f_j}$$

For display purposes, probability is more suitable than probability density (likelihood) or likelihood ratio.

The interpretation to be placed on a likelihood is not at all clear from the numerical value. Probability is commonly inferred only from integrated values of likelihood over a region. Bayes' Theorem gives a probability computed from the conditional probability densities without integration. Display of a probability, expressed on a map in terms of gray scale or saturation of a color selected to represent the category to which an observation is assigned, gives a continuous variable which admits to an intuitive interpretation. The range of the variable is sufficiently predictable to allow scaling of the signal to an analog display device to accomodate all possible values. In cases where probability variations are confined to a small range, and variations within the small range are of interest, it could be desirable either to increase the gain at the expense of dynamic range or elect to display a logarithm of probability.

The importance of admitting continuously variable instead of discretely variable information to the display mechanism is based partly on the reluctance of people to allow machines to draw conclusions without assistance. This reluctance is soundly based, since recognition techniques based on spectral information alone do not take into account the additional information in spatial configuration. The spatial information is only made available if the spectral information is completely displayed. Human interpreters handle spatial information with far greater facility than machines. The whole experience of sight and living contributes to the ability of a human to identify shapes. Machines which perform spectral signature recognition make no attempt to process the spatial information, but they do not necessarily destroy it. The spatial information content is not eliminated until a display device is selected which cannot properly present it for interpretation. Subsequent developments will permit processing for the recognition of spatial patterns. The utility of this technology, however, will be impeded by the use of display devices which present conclusions based on spectral patterns alone without also presenting the continuous variables which permit spatial patterns to appear without smoothing, thus making them available for manual or intuitive interpretation and for subsequent machine processing.

REFERENCES

1. Theodore W. Anderson, An Introduction to Multivariate Statistical Analysis New York, John Wiley and Sons, 1958, 374 pp.

2. Maurice S. Bartlett, "The Vector Representation of s Sample", Proc. Cambridge Phil. Soc., 30, 1934, pp. 327-340.

3. David Blackwell and M. A. Girshick, Theory of Games and Statistical Decisions, New York, John Wiley and Sons, 1954, 355 pp.

4. Irwin D. J. Bross, Design for Decision, New York, The Macmillain Company, 1953, 276 pp.

5. Evelyn Fix and Joseph L. Hodges, Jr., Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties, Report No. 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

6. Joseph L. Hodges, Jr., Discriminatory Analysis: 1. Survey of Discriminatory Analysis, Report No. 1, USAF School of Aviation Medicine, Randolph Field, Tex., 1950.

7. Paul G. Hoel and Raymond P. Peterson, "A Solution of the Problems of Optimum Classification", Ann. Math Statistics, 20, 1949, pp. 433-438.

8. Stanley L. Isaacson, "Problems in Classifying Populations," Statistics and Mathematics in Biology, Ed., O. Kempthorne et. al., Ames, The Iowa State College Press, 1954, pp. 107-117.

9. Maurice G. Kendall, A Course in Multivariate Analysis, New York, Hafner Publishing Co., 1957, 195 pp.

10. C. Radhakrishna Rao, Advanced Statistical Methods in Biometric Research, New York, John Wiley and Sons, 1952, 390 pp.

11. Phillip J. Rulon, "Distinctions Between Discriminant and Regression Analyses and a Geometric Interpretation of the Discriminant Function," Harvard Educ. Rev., 21, 1951, pp. 80-90.

12. Claude E. Shannon and Warren Weaver, Mathematical Theory of Communication, Urbana, Univ. of Illinois Press, 1949, 117 pp.

13. Donald F. Specht, Generation of Polynomial Discriminant Functions for Pattern Recognition, Technical Report No. 6764-5, Systems Theory Laboratory, Stanford University, Stanford, Calif, 1966.

14. Maurice M. Tatsuoka and David V. Tiedeman, "Discriminant Analysis," Rev. Educ. Res., 24, 1954, pp. 402-420.

15. Abraham Wald, Statistical Decision Functions, New York, John Wiley and Sons, 1950, 179 pp.

16. Norbert Wiener, Cybernetics, Cambridge and New York, The Technology Press and John Wiley and Sons, 1948, 194 pp.

SECTION 3

STUDY PLAN

The storage of non-parametric probability density functions required for accomplishing recognition and classification of spectral signatures is identified as the major practical impediment which prevents the immediate construction of an optimum processing system. The approach now taken is to develop the characteristics of a system which will perform in a satisfactory, though perhaps not optimum, fashion under the storage constraints imposed by the state of the art. This problem is addressed with three specific objectives: (1) reduction of the amount of data that must be stored, (2) increasing the total amount of information that can be stored, and (3) increasing the throughput speed with which information can be recovered for data processing.

The study plan, as initially conceived, offered an approach to these objectives based on linear transformations. The soft spot in practical signature recognition systems is representation of the conditional probability density function. By determining the best of many approximate methods for representing the density function, the plan as initially represented would call for selection of a best approach. There are very few criteria for judging a best representation. Judgments based on results obtained are not reliable because all decision boundaries are not necessarily affected by the adequacy of a representation. A classification result says only that the relative sizes of the densities are satisfactory near the decision boundary. When a new decision is required, the boundary does not necessarily fall in a region where the densities have been found satisfactory. The divergence between pairs of distributions is helpful but somewhat difficult to interpret on an absolute scale. As a measure of the relative difficulty of the distinctions in a multiple category classification problem in which similar representations are used for the density of each category, it is unambiguous. Because the measure is sensitive to so many parameters, it is difficult and expensive in terms of computer time to maximize information content by parametric adjustment.

The study as actually conducted differs for these reasons from the plan initially conceived. The comparisons originally undertaken could be carried out with the analytical and computational tools developed during the study. Because results obtained in this way are not conclusive, however, the analytical and computational tools originally conceived and developed were supplemented with methods which directly addressed the problem of finding a satisfactory method of representing conditional probability densities.

For the very reasons that the study was necessary, it was not possible to lay out an initial study plan which both satisfied the control objective of giving specific direction and pacing events and the technical benefit of hindsight.

The material covered in Sections 3.1 and 3.2 represents the structure of the ideas upon which the original study plan was based. Section 3.3 summarizes the unanticipated study findings and developments; more thorough discussions of these findings are found in other sections of the report to which the reader is directed.

3.1 MATHEMATICAL METHODS

The first two objectives - minimizing the amount of data that must be stored and maximizing the amount of information that can be stored - are approached by devising linear transformations of the raw data that permit dimensionality reduction without loss of information. If the useful information contained in eight channels of the raw multispectral observations can be completely mapped into a smaller number of new variables by a transformation, the amount of data that must be stored is proportionately reduced. Because the storage capacity thus made available can be used to store other information, the amount of information that can be stored is increased.

Transformations used for dimensionality reduction, however, must be chosen in such a way that the amount of useful information contained in the variables deleted is as small as possible. One linear transformation considered for this purpose is obtained during factor analysis. Among the properties of such a transformation is one which may allow the space required to store certain kinds of information to be reduced even if dimensionality reduction is not possible. The transformation which results from diagonalizing the correlation matrix of a data sample is known in the literature of factor analysis as a "principal component" transformation and the study of this transformation is known as "principal component analysis." It can be shown that the transformation has the following properties:

1. The new variables, known as the principal components, are uncorrelated. This property is exactly true only for the principal components of the data sample analyzed in this way.

2. The transformation can be reversed to recover exactly the original data. No information is lost by applying the transformation.

The first property does not guarantee the statistical independence of principal components, but the property is more likely to be true of the variables transformed in this way than of the original data. The second property is true for any data to which the transformation is applied; it is non-singular.

One potential value of this transformation lies in the extent to which the principal components are not only uncorrelated but statistically independent also. If this is so, the information storage requirement becomes a linear function of the number of channels instead of an exponential one. Storage of the most general signature distribution, of the form,

$$f = f(X_1, X_2, \ldots, Xm) \qquad (3.1\text{-}1)$$

requires

$$N = R^m \qquad (3.1\text{-}2)$$

words for each of the alternatives among which a new signature is to be classified, where N is the storage requirement in words of a length in bits necessary to distinguish R resolvable levels. For an eight-channel device with 100 resolvable levels on each channel, the total requirement is

$$N = (100)^8 = 10^{16} \text{ words} \qquad (3.1\text{-}3)$$

for each alternative. If however, the signals on the eight channels are statistically independent, the storage of the most general distribution can be expressed in the form

$$f(X_1, X_2, \ldots X_8) = f_1(X_1)\, f_2(X_2) \cdots f_8(X_8) \qquad (3.1\text{-}4)$$

and requires only

$$N = mR = 8 * 100 = 800 \text{ words} \qquad (3.1\text{-}5)$$

for each alternative.

In the event that the principal component transformation produces variables which are uncorrelated but not statistically independent, it may nonetheless be possible to identify pairs of these variables which are more nearly statistically independent of the other pairs than each member of a pair is of the other. Since principal component analysis removes linear dependence (which is a lack of correlation) but cannot remove simultaneously the nonlinear dependence, the pairing of dependent principal components brings together the strongest residual nonlinear dependencies.

A bivariate joint distribution describes each such pair. In this situation, the complete multivariate distribution for each alternative can be written as:

$$f(X_1, X_2, \ldots, X_8) = f_1(X_1, X_2)\, f_2(X_3, X_4)\, f_3(X_5, X_6)\, f_4(X_7, X_8). \qquad (3.1\text{-}6)$$

If digital methods are used to store this information, the reduction in storage requirement is less impressive than in the case of completely independent variables, equation 3.1-5. The requirement now is proportional to the number of independent bivariate distributions. For the eight-channel example,

$$N = (1/2)n\,R^2 = 4\,(100)^2 = 4 * 10^4 \text{ words} \qquad (3.1\text{-}7)$$

3.1.1 Principal Component Analysis (Factor Analysis)

The role of the principal component transformation is to express the raw data in a new coordinate system in which the new variables, principal components, are more nearly independent than originally. Factor analysis, as performed in modern times by computer, often begins with a principal component or principal factor analysis of the data. The reader is referred to the literature of factor analysis * for further discussion of the terms and techniques of that important subject. A principal component transformation is obtained as follows:

1. Prepare the data sample for analysis. Express this as a matrix with m rows (corresponding to multivariate observations with m channels) and n columns (corresponding to n observations of the target described by the data sample). Require that n > m. Call this

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & & X_{2n} \\ \vdots & & \ddots & \vdots \\ X_{m1} & & & X_{mn} \end{pmatrix}$$

2. The data may or may not be put in standard form. It has been found desirable to do so. Compute the mean and standard deviation of each row of the data matrix. From each element, subtract the mean value of its own row and divide by the standard deviation. Define this as

$$Y = \frac{1}{\sqrt{n}}\, S^{-1} (X - \bar{X})$$

* Harry H. Harman, Modern Factor Analysis, 2nd Edition Revised, The University of Chicago Press.

which is the matrix of the data sample in standard form. S is the diagonal matrix of standard deviations for each row:

$$S = \begin{pmatrix} S_{11} & & & \\ & S_{22} & & 0 \\ & & \ddots & \\ & 0 & & S_{mm} \end{pmatrix}$$

and

$$S^{-1} = \begin{pmatrix} 1/S_{11} & & & \\ & 1/S_{22} & & 0 \\ & & \ddots & \\ & 0 & & 1/S_{mm} \end{pmatrix}.$$

$\bar{X}$ is a matrix of n identical columns; each row is the mean value of a row of X repeated n times:

$$\bar{X} = \begin{pmatrix} \bar{X}_1 & \bar{X}_1 & \cdots \\ \bar{X}_2 & \bar{X}_2 & \cdots \\ \vdots & \vdots & \\ \bar{X}_m & \bar{X}_m & \cdots \end{pmatrix}.$$

3. Form the correlation matrix of the data sample. This is:

$$R = YY^T.$$

4. Diagonalize the correlation matrix using standard procedures.* The result is:

$$D = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \ddots & \\ & 0 & & \lambda_m \end{pmatrix} = URU^T.$$

The real symmetric matrix, R, is diagonalized to a form of eigenvalues, $\lambda_1, \ldots, \lambda_m$, by the unitary matrix U, formed from the m eigenvectors of R. U^T = U transposed.

5. Multiply both sides of the above by the inverse of D expressed in the form:

$$D^{-1} = D^{-1/2} D^{-1/2}$$

where

$$D^{-1/2} = \begin{pmatrix} \lambda_1^{-1/2} & & & \\ & \lambda_2^{-1/2} & & 0 \\ & & \ddots & \\ & 0 & & \lambda_m^{-1/2} \end{pmatrix}.$$

Substitute yy^T for R and the result is:

$$I = D^{-1/2} UYY^T U^T D^{-1/2}$$

since $D^{-1/2}$ is commutative.

* I. S. Sokolnikoff and R. M. Redheffer, Mathematics of Physics and Modern Engineering, McGraw-Hill Book Co., 1958.

6. Define

$$F = D^{-1/2}UY$$

$$= \frac{1}{\sqrt{n}} S^{-1}D^{-1/2}U(X - \bar{X}).$$

This is the matrix of principal components in standard form. It is m • n just as is the data sample matrix. In nonstandard form, the principal component matrix is:

$$G = F + \frac{1}{\sqrt{n}} S^{-1}D^{-1/2} U\bar{X}$$

$$= \frac{1}{\sqrt{n}} S^{-1}D^{-1/2} UX.$$

Note that the principal components are uncorrelated because the correlation matrix of the principal components is equal to the identity matrix.

$$FF^T = (D^{-1/2}UY)(D^{-1/2}UY)^T$$

$$= (D^{-1/2}UY)(Y^TU^TD^{-1/2})$$

$$= I.$$

7. The principal component transformation is thus identified as the m • m matrix.

$$T = \frac{1}{\sqrt{n}} S^{-1}D^{-1/2}U$$

and

$$G = TX$$

or

$$F = T(X - \bar{X}).$$

A computational scheme to perform principal component analysis of multispectral data has been developed. A flow chart describing the optional methods of analysis available and a computer program logical flow chart is presented in Volume II of this report.

3.1.2 Bayesian Decision Processor

The decision rules discussed in Section 2 are general and equivalent to the Bayesian approach described here. The coefficients and parameters used in the decision rule which minimizes the cost of an experiment admit to a direct intuitive interpretation. A computational method for applying this decision rule is described in Volume II with a computer program flow chart.

The criterion is minimum cost. The expected cost of an experiment is:

$$E(w) = \sum_{k=1}^{k} \sum_{j=1}^{k} w(j, k)\, P(j, k) \qquad 3.1\text{-}8$$

where

w = cost of the experiment

$w(j, k)$ = cost of assigning an object of class j to class k

$P(j, k)$ = probability of assigning an object of class j to class k.

The probability $P(j, k)$ can be expressed as the product of an a priori and a conditional probability

$$P(j, k) = P(j)\, P(k \mid j) \qquad 3.1\text{-}9$$

where

$P(j)$ = a priori probability of occurence for objects of class j

$P(k \mid j)$ = probability of assigning an object to class k provided it belongs to class j

The problem is to define values of the observed signature which cause it to be attributed to a class k. Let Γ_k be the set of signatures (i.e., a region in the space of signatures) which is attributed to class k. Then, by definition:

$$P(k \mid j) = \int_{\Gamma_k} f_j(\vec{x})\, d\vec{x} \qquad (3.1\text{-}10)$$

where

$\vec{x}$ = an observed signature

$f_j(\vec{x})$ = the probability density function of signatures for objects of class j

Γ_k = acceptance region for assignment to class k

or

$$P(k \mid j) = \int_{\Omega} \Gamma_k(\vec{x})\ f_j(\vec{x})\, d\vec{x} \qquad (3.1\text{-}11)$$

where

Ω = the entire signature space

$\Gamma_k(\vec{x})$ = 1 for $\vec{x} \in \Gamma_k$

= 0 for $\vec{x} \in \Omega - \Gamma_k$

The expected cost is then:

$$E(w) = \sum_{k=1}^{k} \sum_{j=1}^{k} w(j, k)\, P(j) \int_{\Omega} \Gamma_k(\vec{x})\, f_j(\vec{x})\, d\vec{x}$$

$$= \int_{\Omega} \sum_{k=1}^{k} \Gamma_k(\vec{x}) \left[\sum_{j=1}^{k} w(j, k)\, P(j)\, f_j(\vec{x}) \right] d\vec{x}. \qquad (3.1\text{-}12)$$

To minimize the expected cost, evaluate the cost function in square brackets for each k and set $\Gamma_k(\vec{x}) = 1$ for the k associated with the minimum cost. The decision rule implied is to assign the observation $\vec{x}$ to the category k which minimizes the cost of the decision.

The computational scheme which applies this decision rule and is described in Volume II requires a reliable method for computing the probability densities $f_j(\vec{x})$. During the study, more effort is directed to the study of how reliably to describe probability densities than to applying the Bayesian Decision Processes.

3.1.3 Information Divergence

Divergence is a measure of the amount of information which the differences between two distributions contribute to the ability to classify a spectral signature. The divergence is defined as the sum of two information measures.

Suppose a signature $\vec{x}$ arises from one of two object classes, j or k. The likelihood ratio method of classification is to evaluate the likelihood function

$$L(\vec{x}) = \frac{f_j(\vec{x})}{f_k(\vec{x})}$$

and assign the signature to object class j if L exceeds some criterion value, and to class k otherwise. The amount of information in the likelihood ratio for making the decision is log L. (The base of the logarithm does not matter since the results vary only by a multiplicative constant; however, $\log_2 L$ gives the information in bits.)

To obtain the average information per observation for evaluating the hypothesis that the signature belongs to object class j against the alternative class k, it is necessary to take the expected value over distribution j:

$$I(j;k) = \int (\log \frac{f_j(\vec{x})}{f_k(\vec{x})}) f_j(\vec{x}) d\vec{x}$$

$$= \int \log L(\vec{x}) f_j(\vec{x}) d\vec{x}.$$

Similarly, the average information for evaluating the hypothesis that the signature belongs to object class k against the alternative class j is:

$$I(k;j) = \int (\log \frac{f_k(\tilde{x})}{f_j(\tilde{x})}) f_k(x)dx$$

$$= -\int \log L(\tilde{x}) f_k(x)dx.$$

The divergence between the two populations is:

$$J(j, k) = I(j;k) + I(k;j).$$

Note that J(j, k) applies to differences between distributions and not to the sample of signatures to be classified. Note also that I(j;k) and I(k;j) are integrals, but the integration takes place over different distributions, viz, the distributions of the two object classes. Finally, note that the information measures, while integrated over only one of the two object class distributions, are not defined unless the alternative class is specified; i.e., each is defined for the distribution, but only against a well defined alternative.

To evaluate the divergence between distributions of object classes, an average of the log likelihood functions over the reference data sets replaces the integral. Th.. distributions are assumed to be characterized by the reference data sets and the algorithms for evaluating the density functions.

Let N_j and N_k be the number of observations in each of the two reference samples. The required computations are:

$$I(k;j) = \frac{1}{N_j} \sum_{i=1}^{N_j} \log \frac{f_j(\tilde{x}_i)}{f_k(\tilde{x}_i)}$$

$$I(k;j) = \frac{1}{N_K} \sum_{i=1}^{N_k} \log \frac{f_k(\tilde{x}_i)}{f_j(\tilde{x}_i)}$$

$$J(j, k) = I(j;k) + I(k;j).$$

The density functions are computed with one of the optional methods provided for all computational routines which require the use of a density function. The computational scheme is documented in detail with a computer program flow chart in Volume II.

3.1.4 Probability Density Functions

Signature distributions, particularly storage of the information which describes them, make the task of implementing any multivariate decision strategy a difficult one. The intended approach requires that alternative methods of expressing and storing these distributions be evaluated by comparing the performance of Bayesian decision strategy using the various methods. Five methods of expressing multivariate distributions of observations are provided:

1. General Form; Multivariate Histogram

$$f(X_1, X_2, \ldots X_m) = f(\vec{X})$$

2. Independent Bivariate Histograms

$$f(X_1, X_2, \ldots X_m) \cong f_1(X_1, X_2)\, f_2(X_3, X_9) \ldots f_{m/2}(X_{m-1}, X_m)$$

3. Independent Univariate Histograms

$$f(X_1, X_2, \ldots X_m) \cong f_1(X_1)\, f_2(X_2) \ldots f_m(X_m)$$

4. Independent Univariate Normal Distributions

$$f(X_1, X_2, \ldots X_m) \cong \left(\frac{1}{(2\pi)^{m/2} S_1 S_2 \ldots S_m}\right) \exp\left(-\frac{1}{2} \sum_{i=1}^{m} \left(\frac{X_i - \bar{X}_i}{S_i}^2\right)\right)$$

5. Multivariate Normal

$$f(X_1, X_2, \ldots X_m) = \frac{\left|R^{-1}\right|^{1/2}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} X^T R^{-1} X\right)$$

Each of the histogram forms is smoothed according to the method suggested by Specht. The general expression for the smoothing technique is:

$$f(\vec{X}) = \sum_{i=1}^{N} g(\vec{X} - \vec{Y}_i)$$

The $\vec{Y}_i$ are members of the reference sample set which is taken as representative of the population of interest. To evaluate a density function at a point $\vec{X}$ which does not coincide with an observation $\vec{Y}_i$ in the reference sample set, a smoothing function, $g(\vec{X} - \vec{Y}_i)$, is used. The functions $g(\vec{X} - \vec{Y}_i)$ can be normalized so that:

$$\int_{\infty} f(\vec{X})\, d\vec{X} = 1 .$$

This approach, a concept equivalent to the use of a smoothed and normalized histogram, is used for the histogram forms of the density function. The immense storage requirements imposed by the actual use of a general multivariate histogram are avoided by instead storing the sample from which such a histogram would be computed. In this way, an impractically long computational time requirement is substituted for an impossible storage requirement; the objective, however, is strictly for evaluation.

Donald F. Specht* suggested the use of an exponential form for $g(\vec{X} - \vec{Y}_i)$.

If a single multivariate observation with m channels is expressed as an m-dimensional vector, $\vec{X}$, the multivariate distribution function for $\vec{X}$ can be written:

$$f(\vec{X}) = \frac{1}{n(2\pi S)^{m/2}} \sum_{j=1}^{n} e^{-\frac{(\vec{X} - \vec{Y}_j)^2}{2S}}$$

*D. F. Specht, IEEE Transactions on Electronic Computers, Vol EC-16, No. 3, June 1967.

This function can be shown to have the mathematical properties of a distribution function (see ref by D. F. Specht). The vectors $\underset{\sim}{Y}_j$, j=1, 2, ... n, are the n multivariate observations which comprise the data sample representing the signature distribution of a single target alternative. The constant, S, is a data-dependent parameter set by the user. The difference vector, $\underset{\sim}{X} - \underset{\sim}{Y}_j$, appears in the argument of an exponential function. The absolute value of this vector causes points $\underset{\sim}{Y}_j$ a large distance from $\underset{\sim}{X}$ to contribute small amounts to the distribution function at the point $\underset{\sim}{X}$, and those nearby to contribute relatively large amounts.

> "From a practical point of view...the estimator...has two serious disadvantages: (1) it requires storage of the entire training set for use in classification, and (2) the amount of computation required for each classification is roughly proportional to the size of the training set. For most practical problems for which a statistical approach is warranted, the [data sample] will easily be large enough to make these characteristics prohibitive." From Research Bulletin, Laboratory for Agricultural Remote Sensing, Vol 3, No. 844, September 1968.*

While the computation time required to employ this function is large even for evaluation purposes, it does have the virtue of giving the best results that can be expected from any decision strategy. By paying the penalty of computational time during the evaluation phase, the degradation in performance which accompanies the use of other approximations can be quantitatively assessed.

The general form multivariate histogram can also be computed using functions which decay linearly instead of exponentially. The computational time saved with the use of linear functions is considerable. During the study, it became necessary to devise such a method. The numerical methods used to compute probability densities are performed in a subroutine with options. All programs which require such a computation have the same options available. The computations are discussed and a computer program flow chart is presented in Volume II.

Two additional forms of the density function are provided because they are frequently used in other work and very often provide a reasonably accurate description of the world.

The independent univariate normal density function, fourth on the list, should be equivalent to the use of the fifth, multivariate normal, whenever the input variables, X, are uncorrelated. This is true in the case of principal components under some conditions. The expression for the general multivariate normal form uses matrix notation in the argument of the exponential. The definitions that

*This document goes on to suggest an expansion of the estimator in polynomial form as a solution to the disadvantages. The solution is not entirely satisfactory, however, for reasons also discussed in the document. The polynomial method was rejected for use during this study to avoid such problems as evaluating the truncation error superimposed on that of selecting smoothing parameters.

apply are: X and X^T are the data matrix and the transpose thereof, and R^{-1} is the inverse of the correlation matrix. Use of the form presumes earlier subtraction of mean values from the data matrix elements.

The computational scheme and computer program flow charts used for each of these is presented in Volume II.

3.1.5 Review of Mathematical Methods

Fig. 3-1 is a block diagram illustrating the various operations performed on multivariate observations to accomplish signature data processing. The 13 blocks are numbered for reference in the discussion which follows.

The basic mathematical methods of principal component analysis, Bayesian decision strategy, information divergence, and computation of probability densities are performed in several of the blocks in Figure 3-1. Blocks 1 through 5 are data input operations only. Block 7, scatter diagrams, represents only a display operation. Blocks 6, 8, and 9 apply the mathematical methods discussed. The computation of probability densities is performed as required in Blocks 8 and 9.

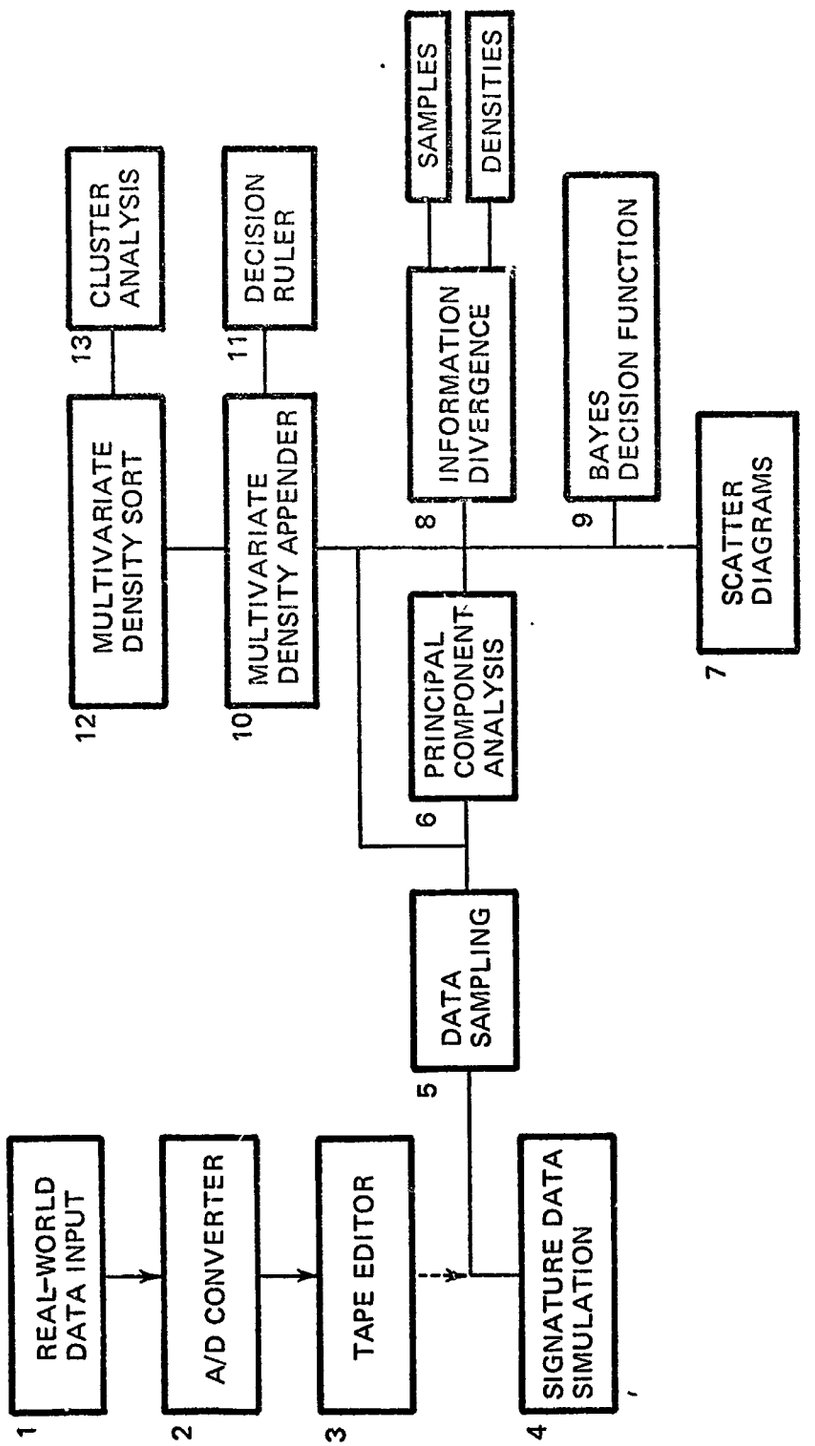
Because of the time consuming nature of the density computations, Block 10 was added to append the density computed with a given set of parameters to each observation of a data set. The operations performed subsequently are performed directly on the previously computed probability densities at a considerable saving in time. These operations are discussed in Section 3.3.

3.2 HARDWARE CONCEPTS

A new hardware concept, applied to the processing of multivariate data, is intended not only to increase the amount of information that can be stored in a processor—the second of the study objectives—but also to raise the throughput rate at which multispectral video can be processed.

For two dimensions, an analog equivalent of a bivariate histogram in digital storage is a scatter diagram . A scatter diagram is a method for presenting a perfectly general bivariate distribution. When observations are plotted in the plane of two factor scores, the bivariate distribution is faithfully described including all residual nonlinear relationships between them. Observations from a single target category plotted in this way represent part of the multivariate signature distribution for that alternative.

A general multivariate distribution can be written as a product of bivariate distributions if each pair of variables is statistically independent of the other pairs.



8059-36

Figure 3-1 Flow of Information Through Signature Data Processing

This condition is more likely to be true of pairs of the transformed variables (the factor scores) than of the pairs of the original variables. The distribution of pairs showing the strongest evidence of residual nonlinear relationships is described on a single scatter diagram. Assuming that the bivariate distribution described on each scatter diagram is statistically independent of the others which apply to the same alternative, they can be multiplied to form the general multivariate distribution. The only remaining question is the nature of an analog device to recover the value of the bivariate distributions from the scatter diagrams.

A flying spot scanner system for this purpose is being evaluated. This differs from the usual flying spot scanner in that deflection of the beam is not in a raster pattern. Rather, the same two video linear combinations plotted on the scatter diagram are used to deflect the beam. The scatter diagram is photographed and reduced in size; in the form of a negative transparency it is then placed over the CRT faceplate with the origin located at the undeflected spot position. Opaque portions of the transparency are those without points; positions of high probability density are those where the transparent spot density is high. The analog linear combinations used to deflect the beam are equivalent to the coordinates of the scatter diagram; thus, with the origin aligned and the deflection gain properly controlled, the flying spot scanner output is a direct measure of a general bivariate probability distribution. When the beam is deflected into a region of high-point density—also high-probability density—more light passes through the transparency to the photomultiplier. The beam spot size is enlarged to provide smoothing of the distribution function.

This provides rapid analog computation of a general conditional joint probability that an observation arose from the alternative described by the signature distribution on the transparencies. With the addition of more flying spot scanners, the signature distribution for several alternatives can be stored and computed rapidly and simultaneously.

Analog summing circuits and dividing networks can then be used to apply any desired decision rules to the probability densities which have been provided in analog form. A functional schematic of a single analog table look-up device is illustrated in Fig. 3-2. Section 6, Hardware Development, describes in some detail the breadboard version of the Analog Table Look-Up device constructed during the study. Two distinct functions are performed by the breadboard version. A principal component transformation is applied to the video with an analog transformation network. This provides the input deflection variables for the flying spot scanners which produce bivariate probability densities as analog output. The two functions are: (1) the transformation and (2) the computation of bivariate probability density.

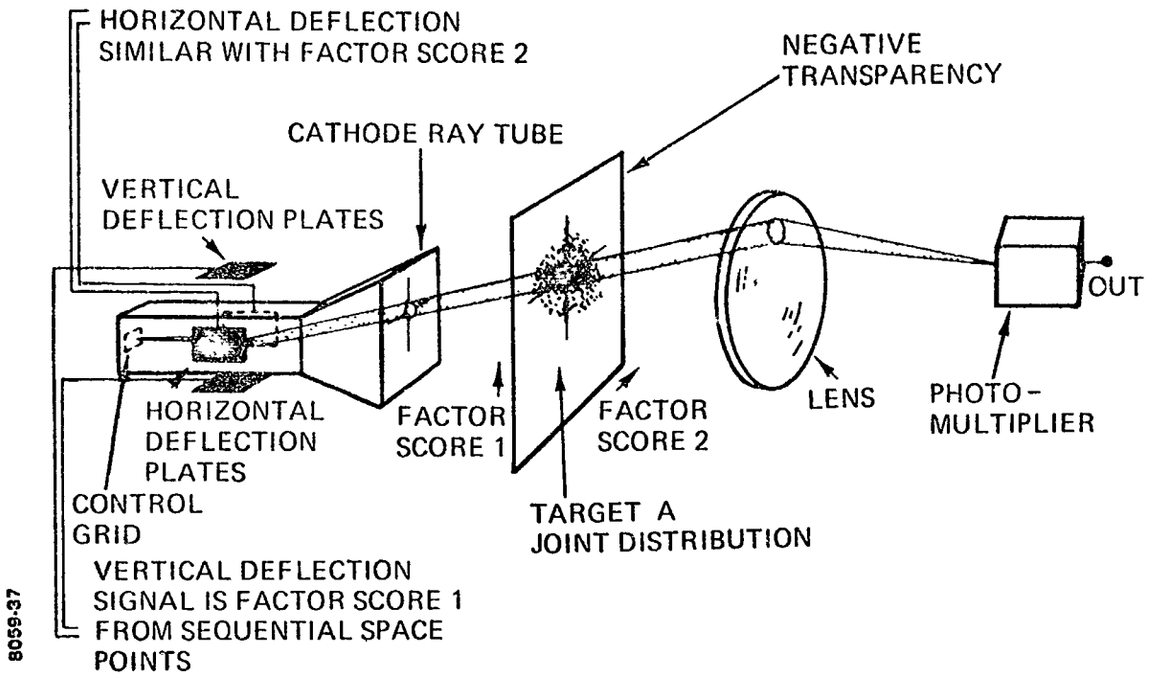


Figure 3-2 Functional Schematic of Analog Table Look-Up Device

3.3 STUDY REVIEW

The analysis portions of the study began with completion of the software necessary to conduct the comparisons required in the original study plan. The review presented here calls attention to the mathematical methods and computational schemes developed during the study that were not initially foreseen. They are described in greater detail else where in the report. Some perspective on the total accomplishment is gained by recognizing the unforeseen developments together.

The concepts discussed include:

1. Cluster analysis
2. Density divergence
3. Smoothing functions
4. Smoothing parameters
5. Canonical analysis.

A new concept in cluster analysis has been developed to locate cluster centers as modes in a hyperspace. While tested and proved with simulated data, the program appears to work equally well on real-world data. This tool eventually can be used to obtain an evaluation of the separability of the alternative target signature distributions which contribute to distribution of all signatures found on a common set of raw data. Cluster analysis applied in this way may eventually relieve the necessity for complete and accurate ground truth. Because the primary study activities did not require the use of cluster analysis, an analytical description of the method used is included in an Appendix to Volume I. The same description is accompanied by a computer program flow chart in Section 1.4, Volume II.

Information divergence has been used largely as a measure of the differences in the probability density functions of two populations. In fact, two samples drawn from the populations of interest are used to allow the lengthy integration otherwise required for a divergence computation to be replaced with a summation over two samples, one from each population. The substitution of sum for integral is performed in Section 3.1. It became necessary during the study to devise a method for measuring the differences in two ways of representing the probability density of a single sample. The concept of a density divergence is therefore distinguished from a sample divergence as noted in Figure 3-1. The density divergence is derived and applied in Section 4 to determine the sample size and smoothing parameters necessary to represent a parametrically describable population.

The exponential form of the smoothing function originally proposed by Specht and contemplated for use during this study, was found to be unnecessarily time consuming. A linear form, computationally faster by using only add-and-substract

arithmetic, was devised to replace it. The linear form is defined and applied in Sections 4 and 5. The linear form maintains the normalizability required of a density function.

The selection of smoothing parameters for use with the smoothing function was a far more difficult task than anticipated. This is a problem not yet solved satisfactorily. Specht suggested a single smoothing parameter for a complete multivariate distribution. During an early program review, it was suggested that the smoothing parameter should be different for each variable. Further developments required directional smoothing in directions not necessarily parallel with the axis of the coordinate system in which the data are represented. Hence, a smoothing matrix is required which reflects directionality similar to that described by a correlation matrix. The relationship between sample size and satisfactory smoothing parameters remains complicated. For sufficiently large samples, it would probably be less so. Development of a generally applicable method for automatic selection of smoothing parameters, allowing for the variable density of observations in hyperspace, is one of the most important areas for further work identified during the study.

Linear transformations produced by principal component analysis appear to be nonoptimum for the purpose of dimensionality reduction. While such transformations permit more efficient storage of information by allowing the probability density to be factored, they are not necessarily the best such transformations for reducing dimensionality. The eigenvalue problem from which they are devised does not distinguish between variance owing to noise-like influences and variance due to spectral information content. The method of canonical analysis described in Section 5.3 is believed to be original. Preliminary evaluations of the method tend to confirm the hypothesis that information of value is more concentrated in the first few canonical variables than in the first few principal components.

SECTION 4

PARAMETRIC STUDIES

The simulated data program, described in Volume II, was used to generate multivariate data samples from populations with known parameters. Computations performed on these samples serve as a test of the software generated during the study and as a guide for interpretation of results. The interpretive role applies particularly to the use of the information divergence programs and the selection of smoothing parameters for use in the Specht function.

When the divergence in units of bits or nits is computed between two populations with multivariate normal distributions, the interpretation of signal-to-noise (S/N) ratio can be placed on the result. This is demonstrated in Section 3. The divergence between a population and a sample drawn from the population is a measure of how well the sample represents the population. The S/N interpretation placed on the value of divergence between two dissimilar populations provides an intuitive interpretation for this measure of similarity.

Divergence computations with simulated data can be performed between populations and samples to help determine the sample size necessary to represent a population. Because samples are used in the study to represent populations and for the direct computation of probability density, the samples must be smoothed with the Specht function concept described in Section 3. The problem of picking suitable smoothing parameters is also investigated with simulated data during the parametric studies.

The diagram in Figure 4-1 shows the flow of data during the parametric studies. Each box on the diagram represents one of the programs applied during the study. These are described in Volume II. The portion of the investigation devoted to software checkout is routine. This section deals with the portion conducted to learn how to apply the software.

The box, i.e., program called Density Lister was written specifically for the detailed study of the Specht function and the selection of smoothing parameters. The Decision Ruler applies various decision rules to the output of the Density Lister. The Decision Ruler was written specifically to realize economies in computation

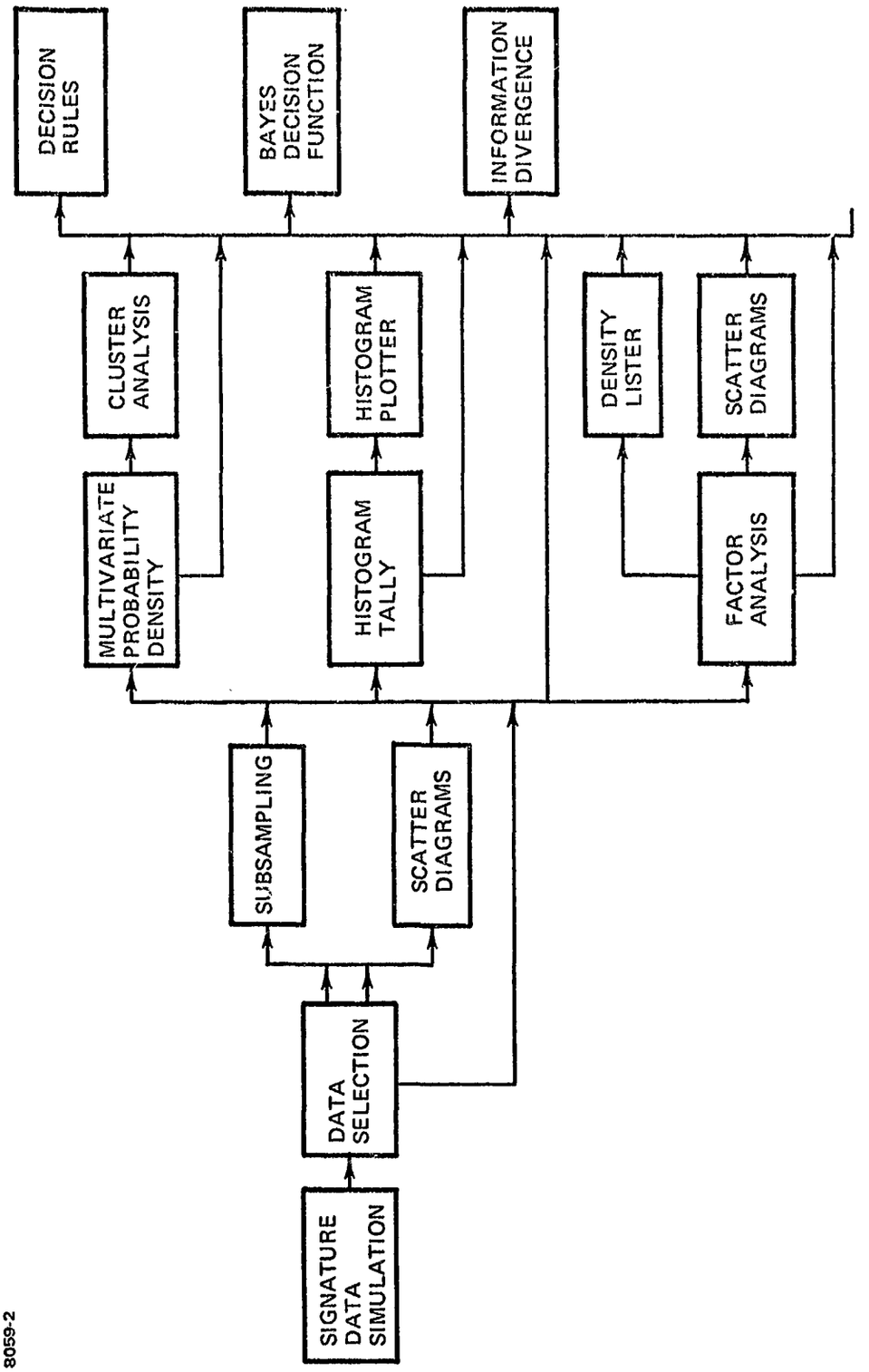


Figure 4-1 Flow of Simulated Data

time by avoiding recomputation of probability densities when the same data are used repetitively in the conduct of the study.

The Density Lister and the Decision Ruler were not used during the parametric study; they were created for use during the study of real world data to satisfy requirements identified during the parametric study.

4.1 DESCRIPTION OF SIMULATED DATA SETS

Four-channel data were used in this parametric study. It is sufficiently complicated for testing the analytical techniques and can be economically displayed in the form of scatter diagrams for visual interpretation and insight. Thirty-two points were chosen to serve as mean values for each of the eight clusters of four-channel data. Standard deviations and eigenvalues for each channel were likewise chosen, the latter to utilize the random correlation option to the Simulated Data Program.* Eight data clusters of 1000 signatures each resulted from the choice of parameters in Table 4-1. The Data Selection Program, followed by the Subsampling Program, was then used to form data set SDPS10 containing three training sets of 200 signatures each and eight sample sets of 100 signatures each. The three training sets were organized as follows:

1. A sampling from a single cluster, cluster 1

2. A sampling from two nearby clusters, clusters 2 and 3, typical of a target affected by different conditions

3. A sampling from five widely scattered clusters, clusters 4 through 8, to represent heterogeneous background area.

Each of the eight sample sets, candidates for assignment to one of the three targets during testing of the Bayes Decision Function Program, was sampled from each of the original clusters.

The scatter diagrams of SDPS10, Figure 4-2, in which two of the possible six channel pairs of the combined training sets are displayed, can be regarded as a two-dimensional projection of the four-dimensional distribution of the signatures. The scatter diagrams of Figure 4-2 show all three training sets comprising SDPS10 plotted on the same diagram. The evidence of clustering is obvious. Figure 4-3 shows similar scatter diagrams for each of the three training sets above.

*A description of the Simulated Data Program options is found in Volume II.

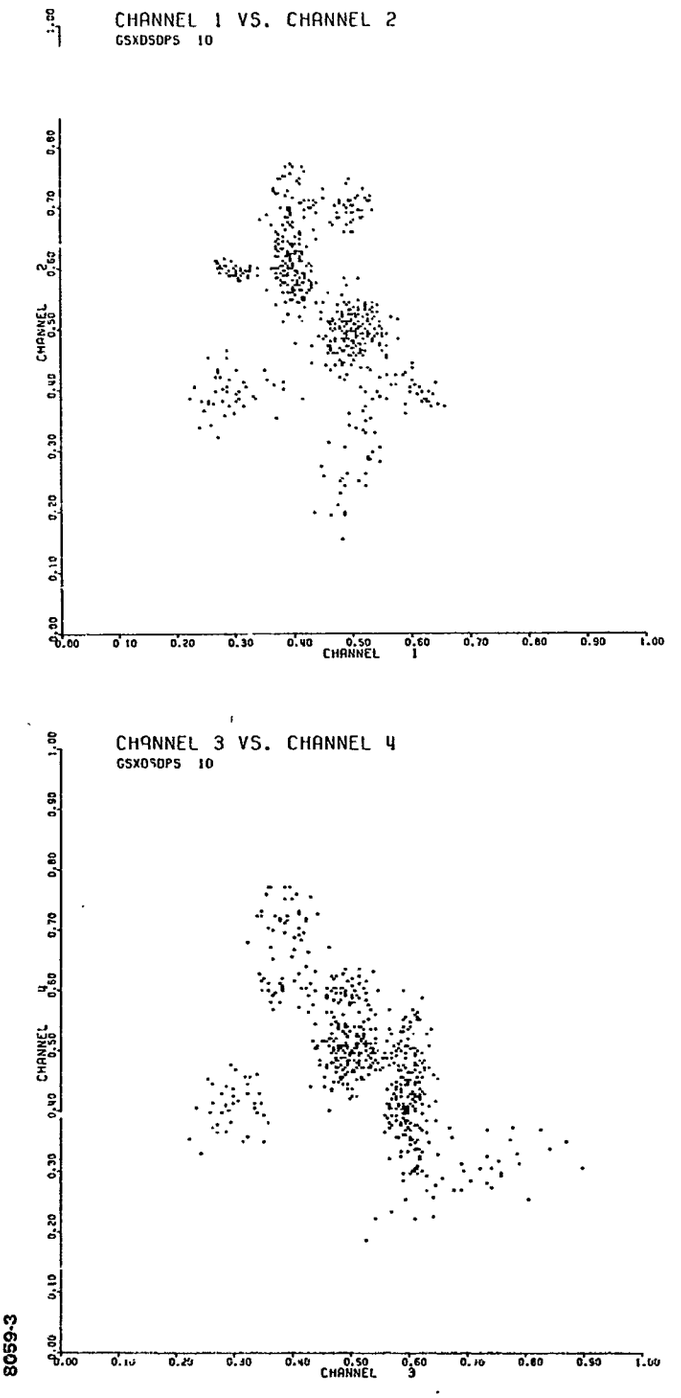
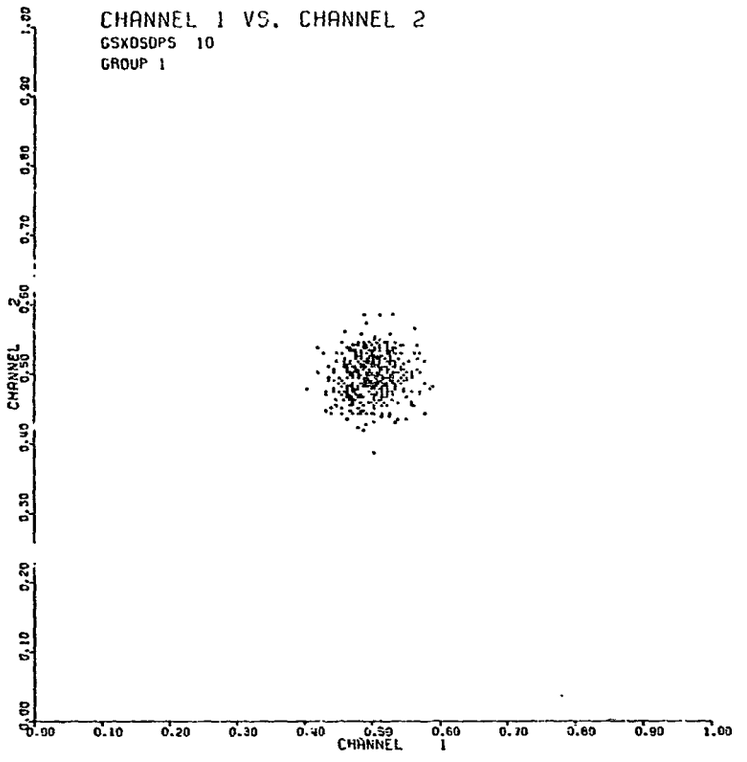
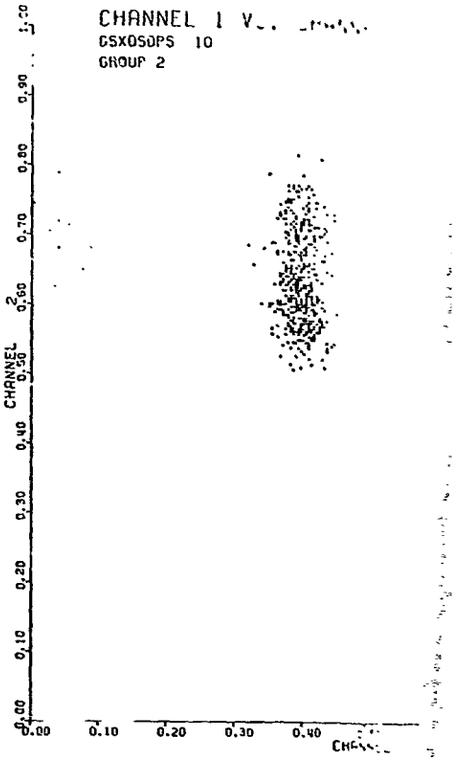
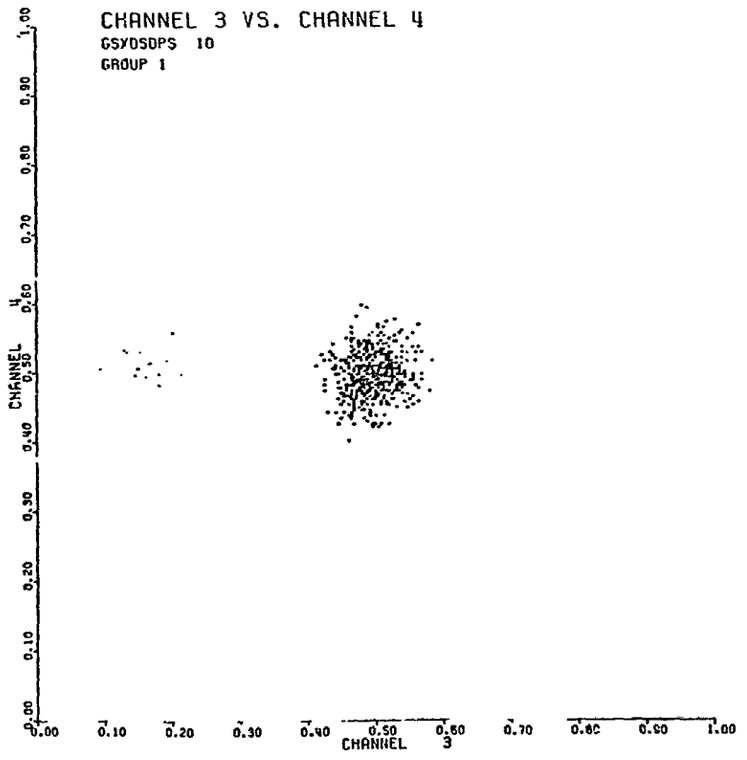
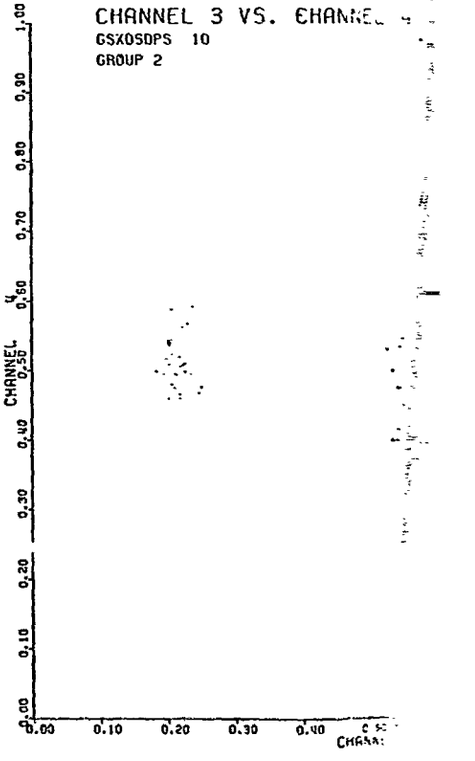


Figure 4-2 Scatter Diagrams, SDPS10





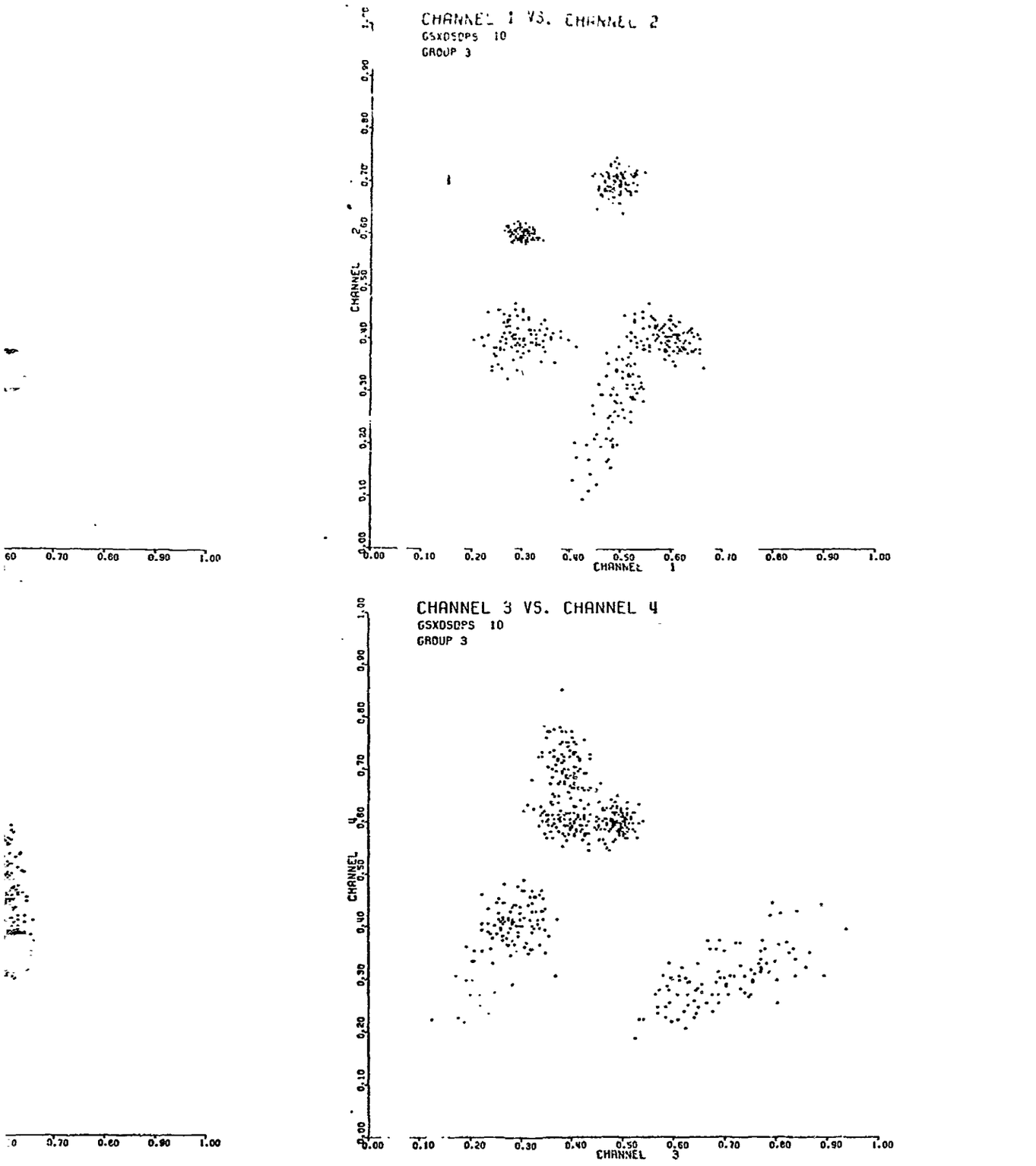




8059-4

FOLDOUT FRAME 1

BSR 2949



FOLDOUT FRAME 2

Figure 4-3 Training Set Groups of SDPS10

TABLE 4-1

SIMULATED DATA PARAMETERS, SDPS10

Cluster	Description	Channel			
		1	2	3	4
1	Mean	0.50	0.50	0.50	0.50
	Standard Deviation	0.03	0.03	0.03	0.03
	Eigenvalue	0.85	1.05	0.95	1.15
2	Mean	0.40	0.60	0.60	0.40
	Standard Deviation	0.02	0.04	0.02	0.04
	Eigenvalue	1.30	0.90	1.10	0.70
3	Mean	0.40	0.70	0.60	0.50
	Standard Deviation	0.02	0.04	0.02	0.04
	Eigenvalue	1.00	1.00	1.00	1.00
4	Mean	0.60	0.40	0.40	0.60
	Standard Deviation	0.03	0.02	0.04	0.02
	Eigenvalue	1.50	1.20	0.80	0.50
5	Mean	0.30	0.40	0.50	0.60
	Standard Deviation	0.04	0.03	0.02	0.02
	Eigenvalue	0.50	0.80	1.20	1.50
6	Mean	0.50	0.70	0.30	0.40
	Standard Deviation	0.02	0.02	0.04	0.04
	Eigenvalue	1.10	1.30	0.70	0.90
7	Mean	0.30	0.60	0.40	0.70
	Standard Deviation	0.02	0.01	0.03	0.04
	Eigenvalue	0.60	0.80	1.10	1.50
8	Mean	0.50	0.30	0.70	0.30
	Standard Deviation	0.04	0.08	0.10	0.06
	Eigenvalue	2.70	0.40	0.70	0.20

TABLE 4-2

EXPECTED CLASSIFICATION RESULTS (TRUTH TABLES)

Unknown Samples	1	2	3	4	5	6	7	8
Target 1	100	0	0	0	0	0	0	0
Target 2	0	100	100	0	0	0	0	0
Background	0	0	0	100	100	100	100	100

Visual inspection shows the clusters to be easily separable; the decision boundaries are so distinct that correct target assignment by the Bayes Decision Function Program is easily accomplished.

To create a sufficiently difficult problem for further study, data sets SDPS15, SDPS20, SDPS24, and SDPS30 were created with 1 1/2, 2, 4, and 10 times the cluster standard deviations, respectively, of SDPS10, and otherwise identical parameters and organization. Progressively more difficult tasks of target separation are demonstrated by scatter diagrams for these data sets, Figure 4-4, comparable to those of SDPS10 in Figure 4-2. Bayes Decision Function and Information Divergence Program results from all five data sets under comparable parameters follow the visual interpretation, and are summarized in Figures 4-5 and 4-6.

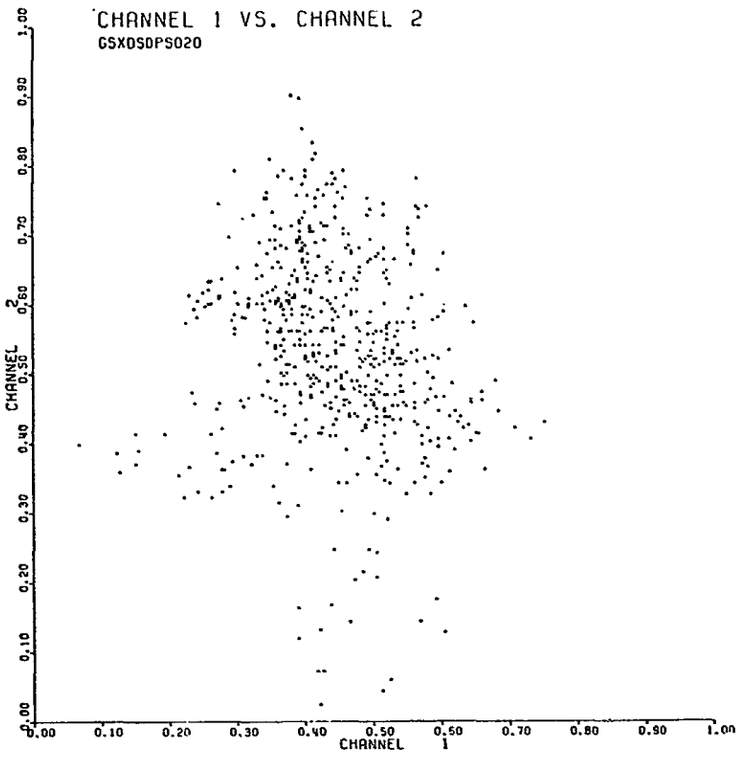
The observations treated as unknown are assigned to one of the three training set groups. The set of unknown observations consists of 100 observations drawn from each of the eight populations described in Table 4-1. The composition of the unknown group was thus 100 from training set group 1, 200 from training set group 2, and 500 from training set group 3.

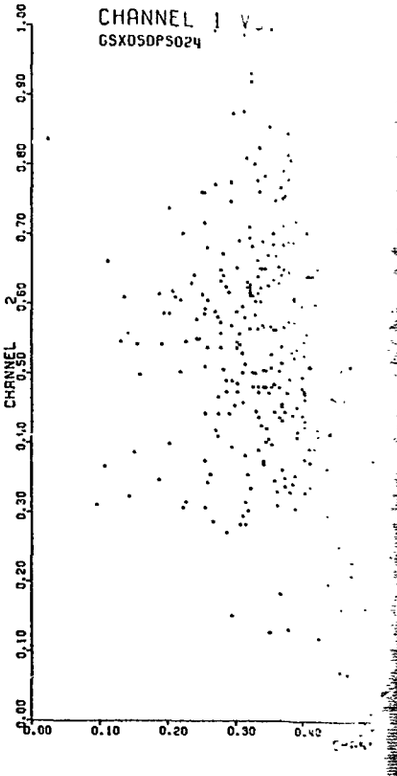
While classification results for SDPS10 were near perfect, the most degenerate success (300 misclassifications) occurs for SDPS30 with all classification of samples as background, the most highly probable category as specified by the a priori probabilities. Training set divergences for SDPS30 were only 0.39 bit between targets 1 and 2, 0.21 bit between target 1 and background, and 0.415 bit between target 2 and background. Truth tables for the classification results summarized in Figure 4-5 are shown in Table 4-2.

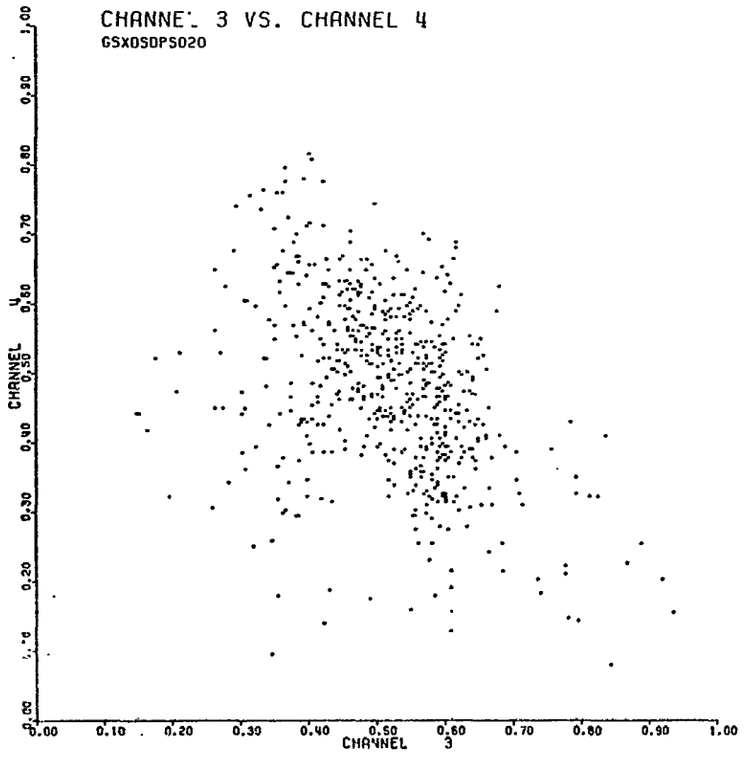
4.2 PRELIMINARY ANALYSIS OF SIMULATED DATA

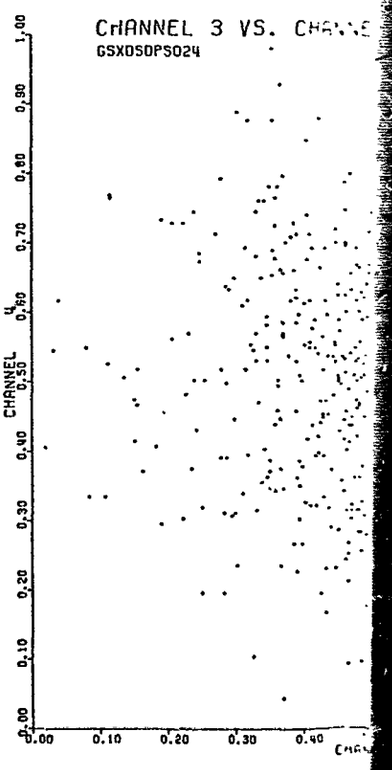
Trial of the Bayes Decision Function Program and the Information Divergence Program was made with training set data in digitized scanner form, using the multivariate Specht smoothing function option * for probability density computation. This suggested areas requiring further study of the analytical techniques. A priori probabilities of occurrence of targets 1, 2, and 3 of data set SDPS10 were specified as 0.2, 0.2, and 0.6, respectively, and the costs of misclassification were equal. The results in Table 4-3 support the conclusion that evaluation of the merits of a particular representation of the probability density on the basis of classification results alone is inconclusive and can be misleading. To obtain correct classification, it is necessary only that the relative values of the probability density functions be correctly ordered; in regions far from a decision boundary, this is not difficult to do. Thus, a representation which produces good results may correctly order the likelihoods near the decision boundaries and also in regions far from decision boundaries for the problem solved without providing a representation equally good close to and far from the decision boundaries. The perfect classification of the eight samples in both runs shown is in spite of the poor choice of Specht smoothing parameters. Decision boundaries can thus be quite insensitive to the probability density representation chosen.

*A discussion of the Specht smoothing function is presented in Section 3.

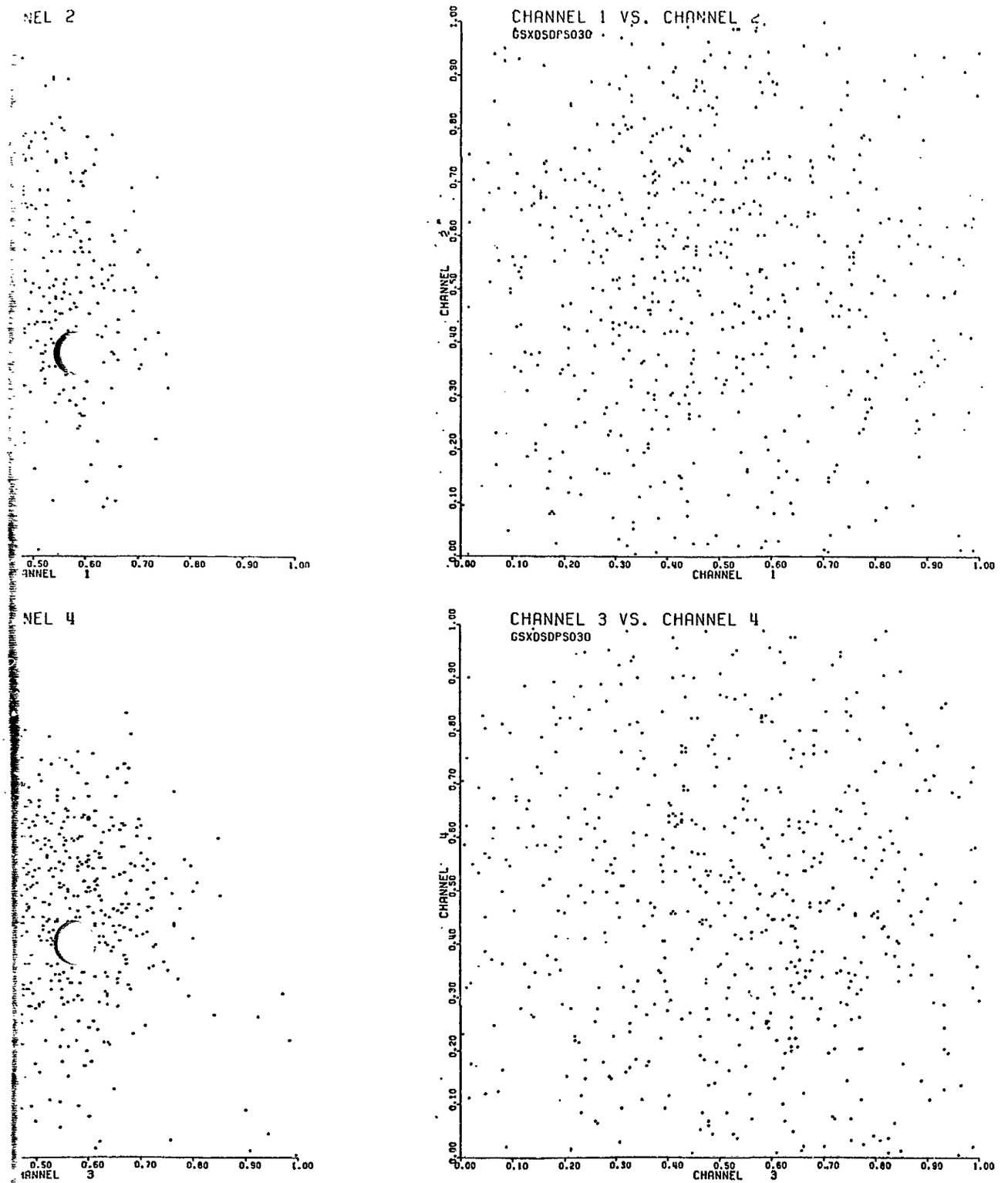








FOLDOUT FRAME 1



FOLDOUT FRAME 2

Figure 4-4 Simulated Data Scatter Diagrams

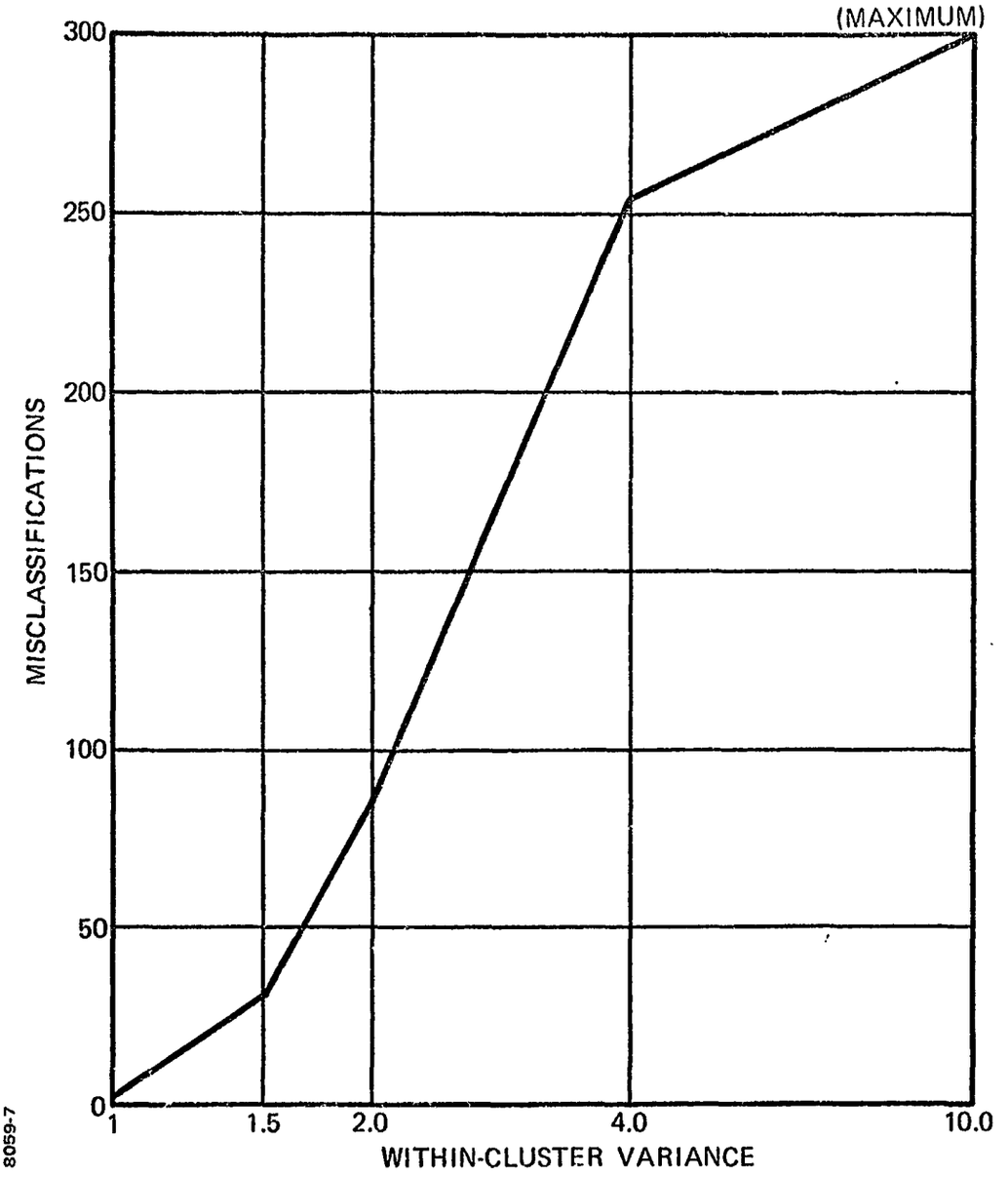


Figure 4-5 Signatures Misclassified as a Function of Within-Cluster Variance (Fixed Cluster Means)

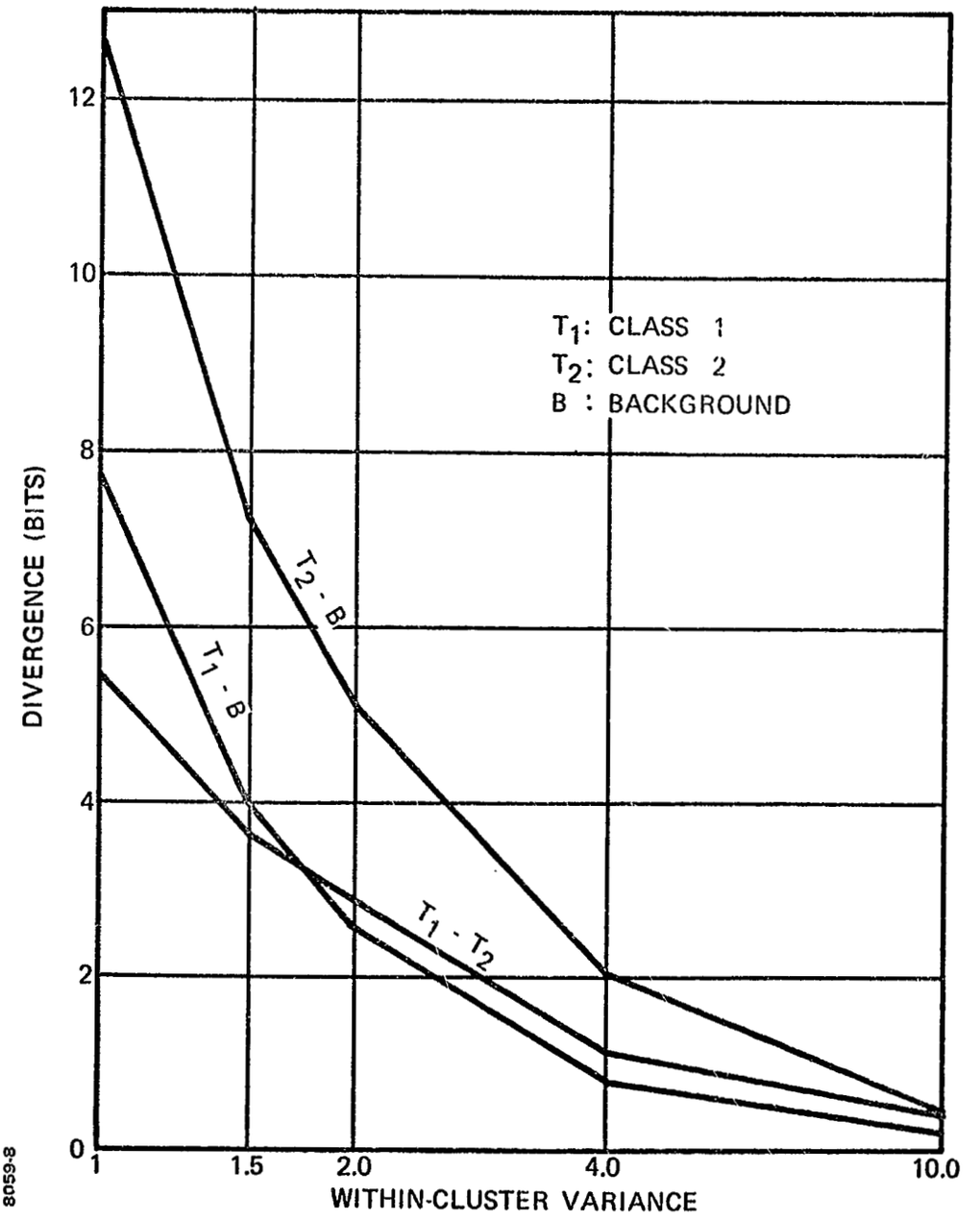


Figure 4-6 Divergence as a Function of Within-Cluster Variance (Fixed Cluster Means)

TABLE 4-3

BAYESIAN DECISIONS AND INFORMATION DIVERGENCE, SDPS10

Run No.	Specht Smoothing Parameters					Classification Results									Divergence Results (bits)		
	Channel	1	2	3	4	Sample	1	2	3	4	5	6	7	8	Tar 1-Tar 2	Tar 1-Back	Tar 2-Back
1	Tar 1	0.0018	0.0018	0.0018	0.0018	Tar 1	100	0	0	0	0	0	0	0			
	Tar 2	0.0013	0.0013	0.0013	0.0013	Tar 2	0	100	100	0	0	0	0	0	26.60	62.42	130.33
	Back	0.0025	0.0025	0.0025	0.0025	Back	0	0	0	100	100	100	100	100			
2	Tar 1	0.01	0.01	0.01	0.01	Tar 1	100	0	0	2	0	0	0	0			
	Tar 2	0.01	0.01	0.01	0.01	Tar 2	0	100	100	0	0	0	0	0	5.40	7.68	12.65
	Back	0.01	0.01	0.01	0.01	Back	0	0	0	98	100	100	100	100			

Factor analysis is a method of multivariate statistical analysis for finding coordinate systems in which the data exhibit certain desirable properties. The solutions are not unique but depend on the model selected by the investigator. Standard factor solutions are available which identify orthogonal factors and have the property that each of the successive factors selected explains a maximum amount of the unexplained variance. The orthogonal solution transforms the data into independent variables by removing those correlations between variables which are the result of linear dependence. Nonlinear relationships which contribute to the correlation between variables cannot be removed in this way. The orthogonal representation also eliminates redundant information, resulting in a moderate reduction in storage requirements. The original observations can be recovered from the transformed data, provided the number of dimensions is the same. If the information content is small for several factors, they may be eliminated from the factor solution with little effect on reproducibility of data or on the subsequent analysis. The model for the orthogonal factor solution is presented in matrix form in Section 3.

The four Bayesian Decision processing and divergence computations were performed with the probability density computation as the Specht smoothing function:

$$f(x_1, x_2 \ldots . x_m) = \frac{1}{(2\pi)^{m/2} N(\prod_{i=1}^{m} S_i)^{1/2}} \sum_{j=1}^{N} \exp\left(-\frac{1}{2}\sum_{i=1}^{m} \frac{(x_i - y_{ij})^2}{S_i}\right)$$

where m is the dimension of the sample and training vectors, x_i is the ith component (factor) of the sample vector, S_i is the smoothing parameter for the ith dimension (factor), y_{ij} is the ith component (factor) of the jth training vector, and N is the number of training vectors. With the factor scores ordered according to decreasing proportion of variance explained by the factor, and with constant S_i, i=1, m, the function was used as follows:

1. Four dimensionally with m = 4 and factors 1, 2, 3 and 4 of the training and sample set vectors

2. Three dimensionally with m = 3 and factors 1, 2, and 3 of the training and sample set vectors

3. Two dimensionally with m = 2 and factors 1 and 2 of the training and sample set vectors

4. One dimensionally with m = 1 and factor 1 of the training set and sample set vectors.

Figure 4-7 graphically represents the results of the four tests. It is helpful and reasonable, since information is additive, to examine the sum of the information divergences between training sets (T1-T2) + (T2-B) + (T1-B). A total of 79 bits of information from such an addition, based on the two-dimensional Specht computation divergence results, would seem to be sufficient to keep classification results satsifactorily stable with those based on the four-dimensional computation. Yet, the one-dimensional computation sends misclassification soaring from five misclassifications with two factors to 195 misclassifications; the total from summation of divergences for the one-dimensional run was only 21 bits. This conclusion, however, cannot be generalized because the information divergence computation is heavily influenced by the selection of smoothing parameters. A conclusion of general validity must therefore be deferred until the problem of selecting suitable smoothing parameters can be objectively treated. This problem is discussed in Section 4.3.

Experience with the Bayes Decision Function Program and interpretation of the results soon made it clear that the confidence to be placed in a classification made with the Bayesian strategy would be useful supplementary information. Computation of the conditional probability that a particular observation may be correctly classified as one of the alternatives presented is easily provided as a by-product of applying the Bayesian decision strategy or any comparable decision rule. Given an observation in multivariate space, the probability density is determined for each target under consideration. If K alternative categories are to be considered, the conditional probability that an observation is correctly classified as belonging to the Jth category is:

$$P(\vec{x} \mid J) = \frac{p_J f_J(\vec{x})}{\sum_{I=1}^{K} p_I f_I(\vec{x})}$$

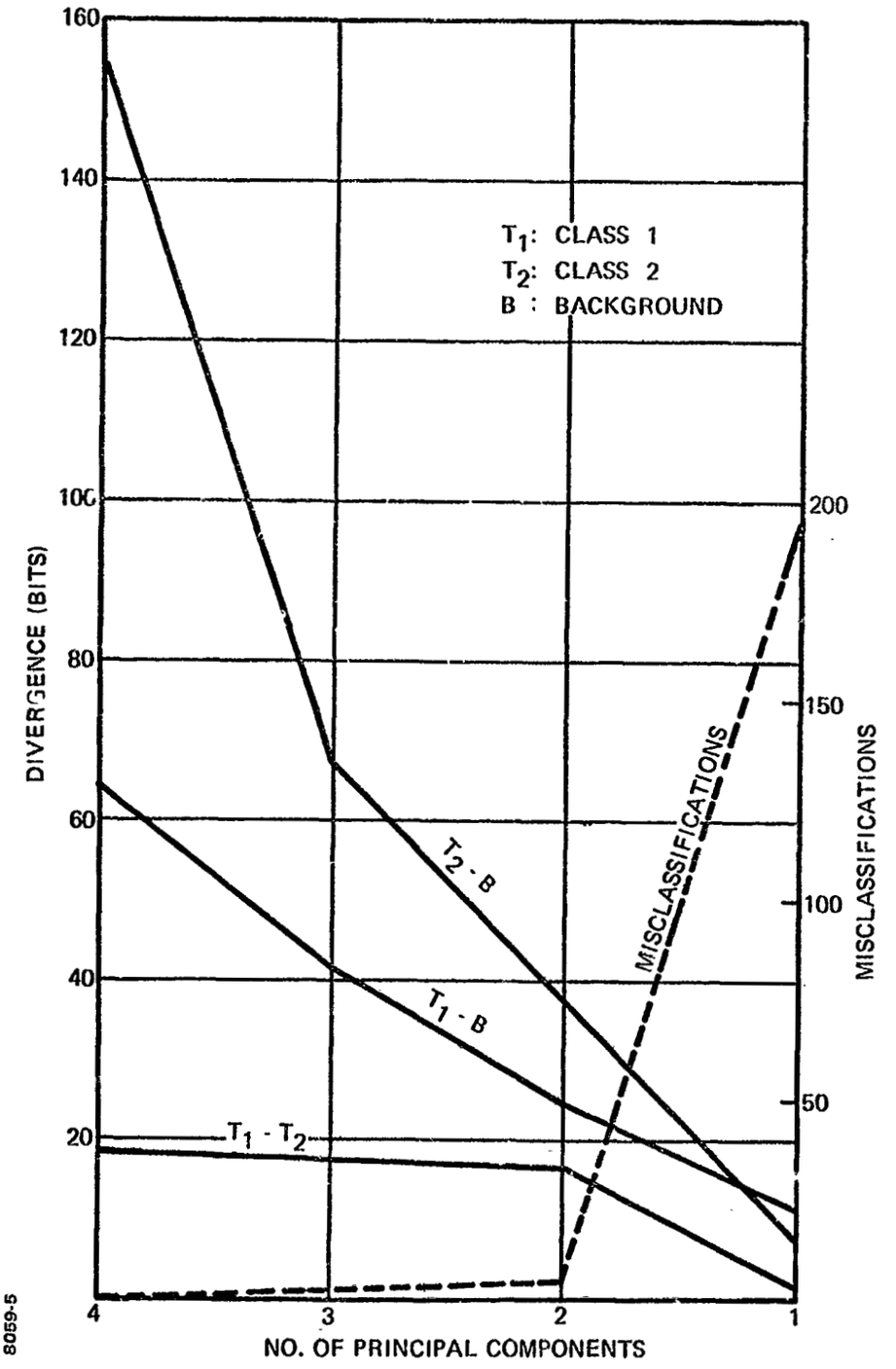


Figure 4-7 Divergence and Number of Misclassifications as a Function of Number of Principal Components (Factor Scores) Used to Represent Spectral Signatures, SDPS10

where p_I is the a priori probability of occurrence of category I and $f_I(\vec{x})$ is the probability density function of signatures of class I. The sample output from the Bayes Decision Function Program in Figure 4-8 shows the usefulness of this added information. For example, sample signature number 601 was classified as belonging to cateogy 1 with the Bayesian a posteriori probability of 0.75907 that the classification is correct. The second and third a posteriori probabilities are the conditional probability that assignment of the same signature to categories 2 and 3, respectively, would be correct. A higher confidence could be placed in the classification of signature number 1400 as category 3 since the probability of correct classification of such is 0.99917.

4.3 SAMPLE SMOOTHING

The best estimate of the probability density function for sample data is based on the actual distribution of the sample. The assumption is that the sample represents the population, and that the true distribution can be approximated with any desired accuracy with a sufficiently large random sample. A method is needed for estimating the probability density of an arbitrary observation which is independent of assumptions as to distributional form. The method described here is based on a function proposed by Specht.* For purposes of computation, the sample may be expressed in the form of a histogram, or the computations can be made directly from the individual observations.

The method consists of attributing a probability density function of arbitrary form to each point in the training sample. The density function at an arbitrary point is obtained by adding up the contributions from all points in the data set and renormalizing. When the probability density function attributed to each point in the training sample is multivariate normal with the identity matrix for channel-to-channel correlation, the form studied by Specht is obtained and the contribution of each point to the total decreases exponentially with the distance from the point at which the density is evaluated.

The density function for a normally distributed variable in a space of m dimensions (channels) is:

$$f(x_1, x_2, \ldots, x_m) = \frac{1}{(2\pi)^{m/2}\sigma_1\sigma_2\cdots\sigma_m} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}\frac{(x_i - a_i)^2}{\sigma_i^2}\right)$$

provided the variables are uncorrelated. Here, a_i is the mean and σ_i the standard deviation in the ith channel. By substituting for a_i an σ_i the points of a training sample, y_{ij}, and the smoothing parameters, $S_j^{1/2}$, and summing the contributions

*Donald F. Specht, "Generation of Polynomial Discriminant Functions for Pattern Recognition," Stanford Electronics Laboratories, Stanford, California, Technical Report No. 6764-5, May 1966.

SIGNATURE CONTRIBUTIONS

INDEX OF SIGNATURE IN SAMPLE DATA SET	SAMPLE GROUP INDEX	OBJECT CLASS ASSIGNMENT	BAYESIAN A POSTERIORI PROBABILITY OBJECT CLASS 1	2	3
601	1	1	0.75907	0.09761	0.14333
602	1	1	0.74840	0.07424	0.17736
603	1	1	0.74457	0.08343	0.17199
604	1	1	0.73821	0.05915	0.20263
605	1	1	0.73955	0.15076	0.10969
606	1	1	0.79936	0.10257	0.09807
607	1	1	0.78364	0.14524	0.07113
608	1	1	0.78323	0.08984	0.12694
609	1	1	0.72844	0.17740	0.09416
610	1	1	0.76493	0.05719	0.17787
611	1	1	0.78607	0.07674	0.13720
612	1	1	0.79440	0.11244	0.09316
613	1	1	0.76196	0.13108	0.10696
614	1	1	0.78454	0.05711	0.15835
615	1	1	0.72387	0.21569	0.05544
616	1	1	0.71305	0.23228	0.05467
617	1	1	0.77660	0.07416	0.14925
•					
•					
•					
1390	8	3	0.01616	0.00798	0.97586
1391	8	3	0.05705	0.09752	0.84543
1392	8	3	0.00258	0.00161	0.99581
1393	8	3	0.01607	0.02821	0.95571
1394	8	3	0.02522	0.01238	0.96240
1395	8	3	0.00021	0.00015	0.99964
1396	8	3	0.00316	0.00050	0.99634
1397	8	3	0.00067	0.00042	0.99891
1398	8	3	0.01364	0.00799	0.97837
1399	8	3	0.01563	0.00755	0.97682
1400	8	3	0.00063	0.00020	0.99917

NUMBER OF SIGNATURES IN SAMPLE GROUPS ATTRIBUTED TO RESPECTIVE CLASSES:

CLASS	SAMPLE GROUP: 1	2	3	4	5	6	7	8
1	100	0	0	2	0	0	0	0
2	0	100	100	0	0	0	0	0
3	0	0	0	98	100	100	100	100

8059.9

FINISH

Figure 4-8 Bayesian Decision Function Program Output

of all points $\vec{y_j}$ in the training sample to the density function at an arbitrary point $\vec{x}$, the following result is obtained for the density function at the arbitrary point $\vec{x}$:

$$f(\vec{x}) = \frac{1}{(2\pi)^{m/2} N(S_1 S_2 \cdots S_m)^{1/2}} \sum_{j=1}^{N} \exp\left(-\frac{1}{2} \sum_{i=1}^{m} \frac{(x_i - y_{ij})^2}{S_i}\right)$$

The index i extends over the m components of the arbitrary point $\vec{x}$ and the points $\vec{y_j}$ in the training sample. The parameter "S_i" is a smoothing parameter which determines the extent to which the density is concentrated about the observation vectors of the sample.

For large values of S_i, the density function is quite smooth. As the S_i approaches zero, the density function approaches the reference sample with an impulse function of amplitude 1/N at each observed point. By choosing suitably small values of S_i, the reference sample distribution can be smoothed without sacrificing the underlying characteristics of the population distribution.

It soon becomes clear to a user, however, that the selection of smoothing parameters which are "suitably small" is no easy task. Different amounts of smoothing are required in different directions. These directions are not necessarily those parallel to the axes used to describe the location of observations in the chance hyperspace defined by the spectral boundaries of channels. The amount of smoothing also depends on the local point density in the sample of observations used to define the distribution of the population. The local point density in turn depends, at any given region in the distribution, on the total number of points in the sample.

The importance of these complications to the direct use of samples to represent populations is illustrated by the results of some preliminary studies with simulated data which are described below. The use of samples in this way was undertaken with full knowledge that the computational time required to describe population densities would make the scheme operationally impractical. A search for a computationally faster method of smoothing samples was begun when it became clear that a major investigation of the methods of selecting smoothing parameters would be necessary. One of the most time-consuming operations in the smoothing function described by Specht is evaluation of the exponential function. To complete this operation requires many multiplications and divisions and a total use of several hundred microseconds per exponentiation. Addition, subtraction, and logical operations are performed relatively fast by the computer. A smoothing function was therefore implemented which can be

evaluated with only add and logical operations once the data have been divided by the smoothing parameters. The difference between the two functions can be imagined by considering the difference between a three-dimensional Gaussian surface and a pyramid.

A generalization of the expression which results from using a smoothing function, $g(\vec{x}, \vec{y}_j)$, to smooth a sample is:

$$f(\vec{x}) = \frac{1}{D} \sum_{j=1}^{N} g(\vec{x}, \vec{y}_j)$$

where $\vec{x}$ is an arbitrary vector, y_j is the jth training vector, N is the number of observations in the training sample, and m is the number of channels. The form of $g(\vec{x}, \vec{y}_j)$ proposed by Specht is:

$$g(\vec{x}, \vec{y}_j) = \exp\left(-\frac{1}{2} \sum_{i=1}^{m} \frac{(x_i - y_{ij})^2}{S_i}\right)$$

and

$$D = (2\pi)^{m/2} N(S_1 S_2 \cdots S_m)^{1/2} .$$

These expressions use smoothing parameters with the dimensions of variance, i.e., the dimension of the x_i, squared.

A pyramid smoothing function, generalized from two dimensions and called hyperpyramid smoothing, has been implemented to shorten the computational time requirements from those of exponential smoothing. The $g(\vec{x}, \vec{y}_j)$ takes the form:

$$g(\vec{x}, \vec{y}_j) = 1 - \max \left\{ \left| \frac{x_i - y_{ij}}{S_i} \right| , \; i = 1, m \right\} \text{ for}$$

$$\max \left\{ \left| \frac{x_i - y_{ij}}{S_i} \right| , \; i = 1, m \right\} \leq 1,$$

$$g(\vec{x}, \vec{y}_j) = 0 \text{ for } \max \left\{ \left| \frac{x_i - y_{ij}}{S_i} \right| , \; i = 1, m \right\} > 1.$$

The form of D for this function is $D = 2^{m-1} N(S'_1 S'_2 \ldots S'_m)$. Here, the S'_i's have the dimensions of the standard deviation, i.e., the same dimensions as the variables x_i. ($(S'_i)^2 = S_i$.) The expression $\left|\frac{x_i - y_{ij}}{S_i}\right|$ is the absolute value of the difference between the jth training sample observation and the same component of the point, $\vec{x}$, at which the density is to be evaluated. The value of i used in the expression for $g(\vec{x}, \vec{y}_j)$ is that single value of i which minimizes the value of $g(\vec{x}, \vec{y}_j)$, but $g(\vec{x}, \vec{y}_j)$ is never allowed to become negative.

The computational time required for the hyperpyramid function evaluation is approximately half that required for the exponential smoothing. Typical computation times for samples used in this study, however, are long even with this saving. Running time is a function of the sample size, number of components, and the number of points at which the density must be evaluated.

4.3.1 Specialized Applications of Divergence

The concept of information divergence is normally applied to compute the divergence between two distributions as represented by two different samples, or between two samples from the same distribution. If the samples are drawn from the same population, this may be used as a measure of the sample size required to represent the population. This divergence in bits is a measure of the difference between two samples from the same population. When this divergence becomes small, it is fair to infer that the samples are large enough to represent the population.

A more direct computation of the divergence between a population with known parameters and a sample drawn from the population is derived below. This modified divergence computation is used when it is desired to determine the divergence between two representations of the same population. It is thus useful in evaluating any given selection of smoothing parameters.

The integral evaluated to compute divergence is:

$$J = \int f_1 \log \frac{f_1}{f_2} \, d\vec{x} + \int f_2 \log \frac{f_2}{f_1} \, d\vec{x}.$$

To compute the divergence between a parametrically computed distribution, f_1, and the Specht representation, f_2, of the same distribution, the integral is first converted to a sum over the sample used for the Specht representation. Note also that:

$$dP = f_1 \, d\vec{x}$$

and

$$\int f_1 d\vec{x} \longrightarrow \frac{1}{N} \sum \text{ ; i. e., integration of the density}$$

function f_1 over $\vec{x}$ is replaced with summation over the training sample. Both procedures are legitimate methods for computing expected values.

Then define

$$\int f_2 d\vec{x} \longrightarrow d\vec{x} = \frac{dP}{f_1} .$$

Then note that

$$\frac{1}{N} \sum \frac{f_2}{f_1}$$

and the expression for divergence can be rewritten as:

$$J = \frac{1}{N} \sum_{i=1}^{N} (\log \frac{f_1}{f_2} - \frac{f_2}{f_1} \log \frac{f_1)}{f_2}$$

$$= \frac{1}{N} \sum_{i=1}^{N} (1 - \frac{f_2)}{f_1} \log \frac{f_1}{f_2} .$$

When f_1 is a parametric expression of the density of a known distribution and f_2 is a smoothing function using a sample from the distribution, the choice of smoothing values used in f_2 may be evaluated. This expression is used in the parametric study to determine the divergence between a multivariate normal population and a sample drawn from the population which is used to represent it via the Specht smoothing function.

In cases of real world data, where the population distribution parameters are unknown or complicated in form, the parameters of smoothing and the sample sizes are selected to provide two samples to represent the population. The samples should differ from each other no more than a sample differs from a parametrically represented population.

4. 3. 2 Investigation of Smoothing Parameters

To use a sample to represent a population as in all of this analysis, a smoothing method must be employed. This smoothing must be local, but not so local that the spaces between sample points are considered to be less probable

than they really are. A systematic examination of smoothing with the Specht function was undertaken with the use of SDPS20, a data set which apparently presents a problem of reasonable difficulty.

Eight attempts at classification and divergence computation were made with the smoothing parameters for the multivariate Specht function representing the following proportions of within-cluster variance: 1/4, 3/8, 1/2, 1, 1 1/2, 2, 3 and 4. While minimizing of percentage error of confusion of target 1 and target 2 occurs with smoothing parameters in the range 3/4 to 3 times the within-cluster variance, the same occurs in the ranges 1/2 to 1 times the cluster variance for confusion of target 2 and the background, and 3/8 to 1/2 times the cluster variance for confusion of target 1 and the background. Also to be noted in Figure 4-9 is that target 1 and the background are to be more easily separable than targets 1 and 2. This conclusion is suggested by the continually larger divergence between the former pair of targets than between the latter pair. Yet, the confusion of samples to be classified as target 1 and the background is continually greater than that of targets 1 and 2. An explanation of this is that the training sample size was too small for representation of the population. Contrary indications are obtained from divergence between these inadequately sized training samples and classification results.

The selection of smoothing parameters on the basis of misclassification would be unwise, particularly in view of the generalization needed to move to real problems. After consideration of the many possible interpretations of these results, it was concluded that the intricacies of smoothing will not permit the selection of smoothing parameters on the basis of an easily established rule; a more careful study was then undertaken.

The amount of smoothing required differs with the shape and number of modes in the actual distribution, the number of points in the sample, the number of channels, and the direction along which smoothing takes place. A pair of correlated variables, for example, with a distribution shape not parallel to either axis, as in Figure 4-10 would be misrepresented by smoothing in the direction of the axes. The distortion caused by such smoothing is indicated by dotted lines on target B. There are infinitely many points well outside such a cluster which would not be treated as such by smoothing only in the directions of the axes. The density evaluated at such a point by the Specht function method is far higher than the distribution warrants. Two methods may be applied to accomodate the requirement for directional smoothing. A smoothing matrix, similar in concept to the inverse covariance matrix used on the multivariate normal expression, could be devised. If this option is adopted, the number of smoothing parameters that must be specified leaps from N = the number of channels to N^2. Furthermore, there is no convenient intuitive interpretation available to select easily the off-diagonal smoothing parameters. Factor analysis, however, provides ready-made software for a simpler

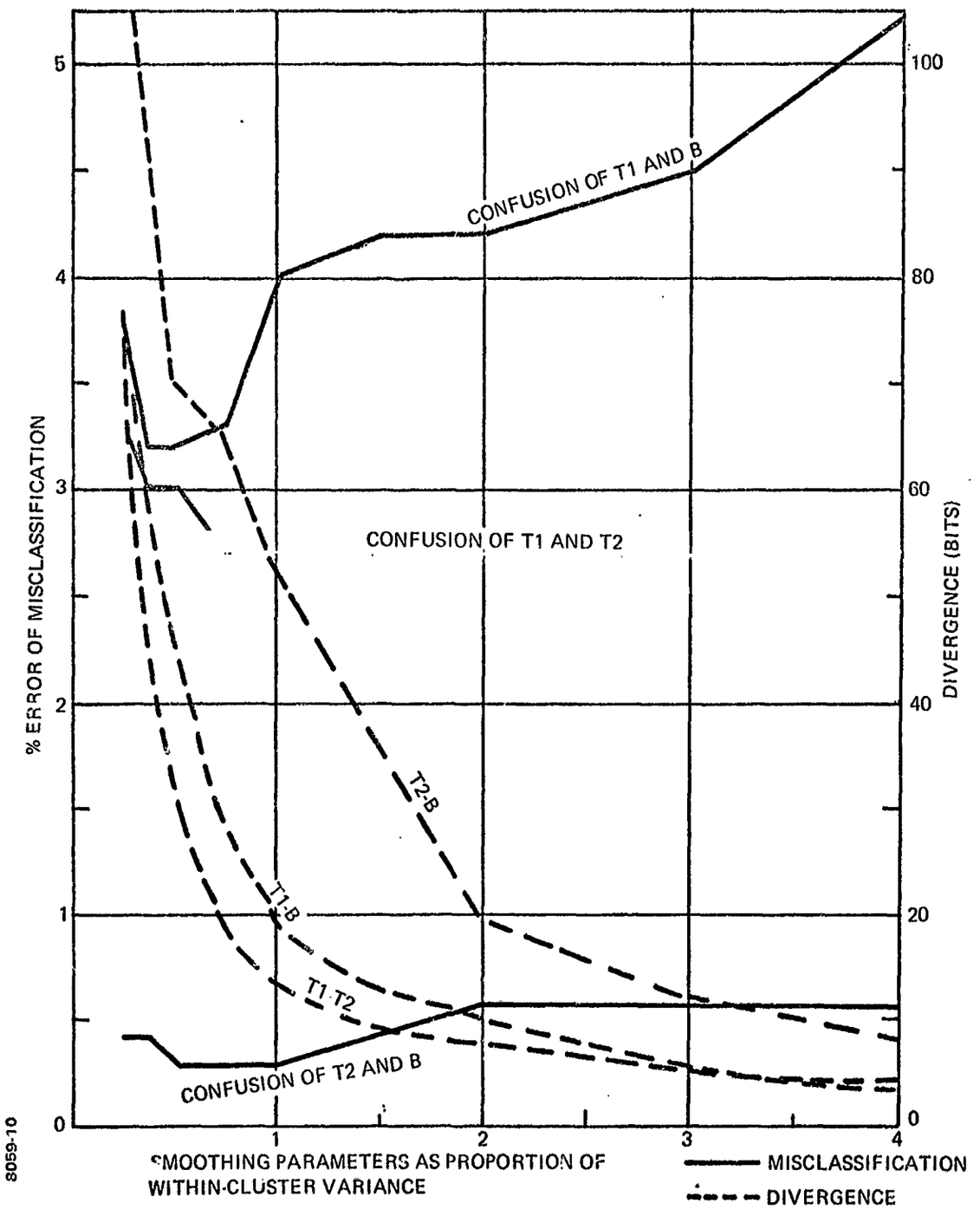


Figure 4-9 Percentage Error of Misclassification vs. Smoothing Parameters as Proportion of Within-Cluster Variance

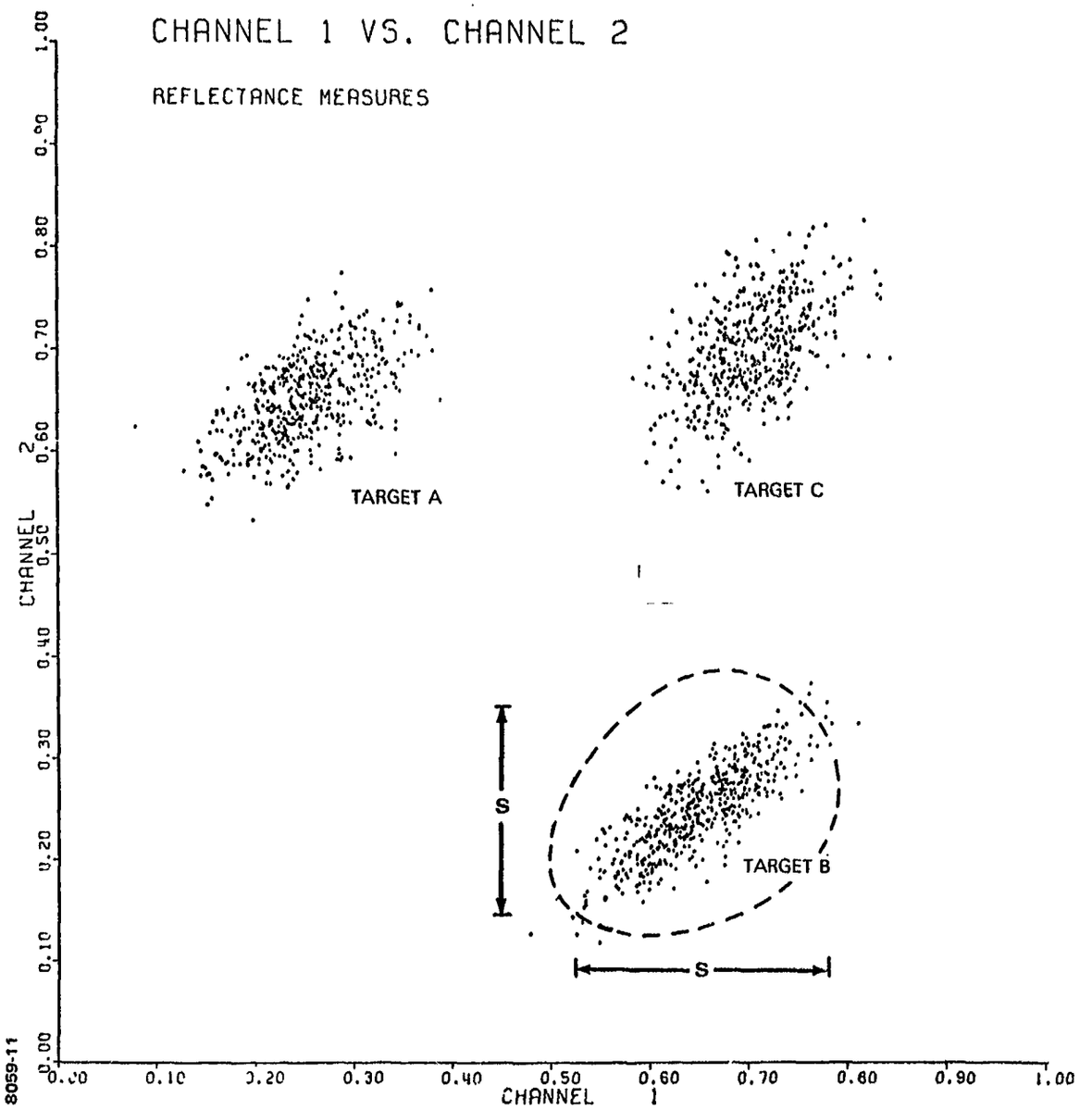


Figure 4-10 Correlated Data Scatter Diagram

approach. Transformation of the data into a coordinate system in which the new variables are uncorrelated factor scores, Figure 4-11, eliminates the need for off-diagonal elements of a smoothing matrix. This approach is used for the balance of the smoothing study.

Factor analysis has the further benefit of permitting a net reduction in the number of variables employed. Because the transformation is used to concentrate information in a smaller number of variables than would otherwise be necessary, and because these variables are uncorrelated, the number of smoothing parameters that must be specified is less than N for real data.

The hyperpyramid form of the smoothing function, discussed in Section 4.3, was used for the balance of the parametric study to realize the savings in computer time. The form of this function is repeated:

$$f_k(\vec{x}) = \frac{1}{2^{m-1} N_k \prod_{i=1}^{m} S_{ik}} \sum_{j=1}^{N_k} g(\vec{x}, \vec{y}_j)$$

where N_k is the number of observations in the kth training set, S_{ik} is the smoothing parameter for the ith channel of the kth training set, and

$$g(\vec{x}, \vec{y}_j) = 1 - \max \left\{ \left| \frac{x_i - y_{ijk}}{S_{ik}} \right| , \; i=1, \; m \right\} \quad \text{for}$$

$$\max \left\{ \left| \frac{x_i - y_{ijk}}{S_{ik}} \right| , \; i=1, \; m \right\} \leq 1,$$

$$g(\vec{x}, \vec{y}_j) = 0 \text{ for } \max \left\{ \left| \frac{x_i - y_{ijk}}{S_{ik}} \right| , \; i=1, \; m \right\} > 1.$$

y_{ijk} is the ith component of the jth signature of the kth training set, and x_i is the ith component of $\vec{x}$. This hyperpyramid function can be implemented with the use of logical and add instructions, eliminating the costly exponentiation of the previous smoothing function.

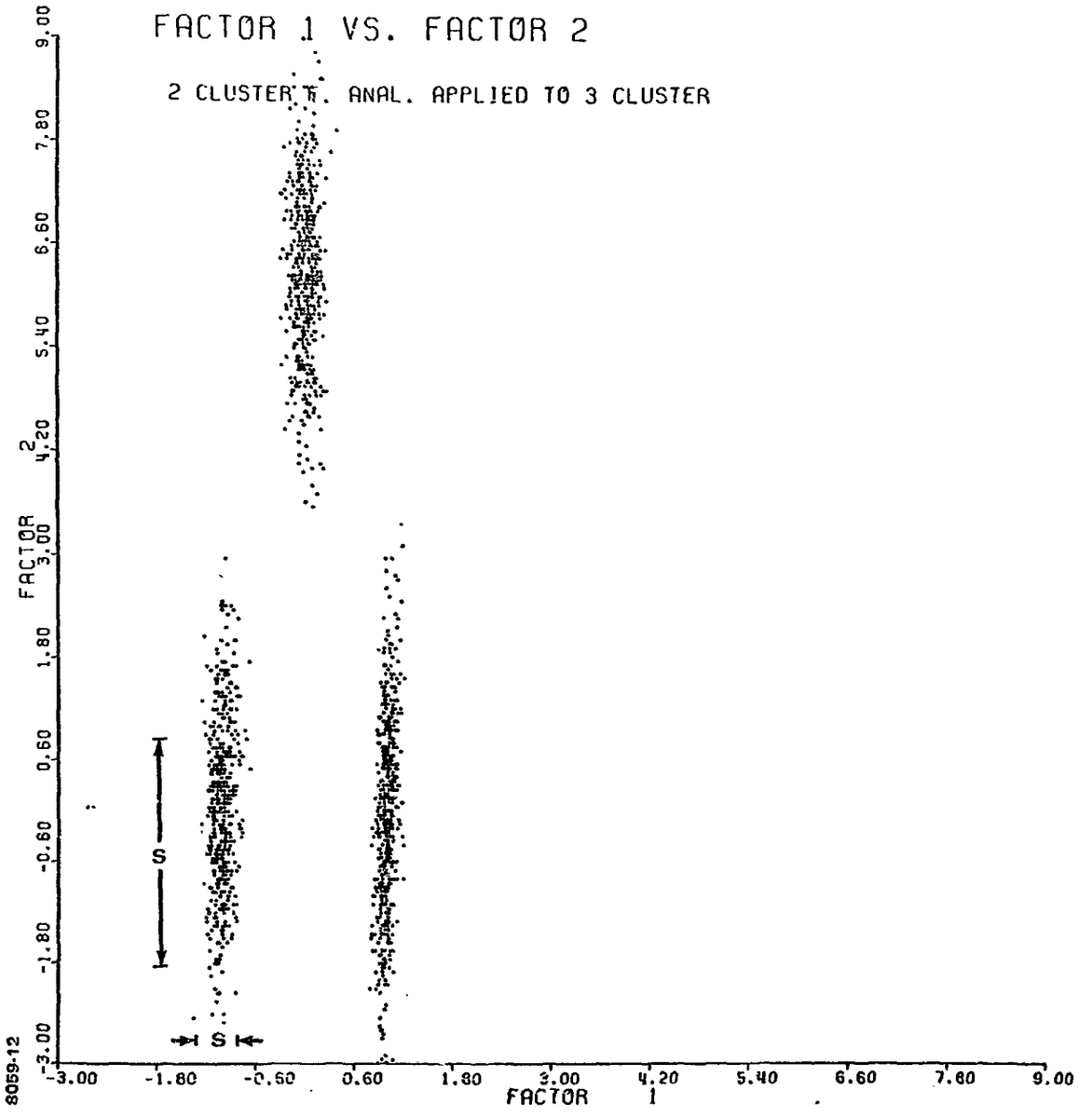


Figure 4-11 Factor Score Data Scatter Diagram

A modified divergence computation * was developed for determining the bit difference between the distribution of a population with known parameters and a representation of that distribution obtained from a smoothed sample. This divergence computation, referred to in the remainder of this section as density divergence, is:

$$J = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{f_2}{f_1}\right) \log \frac{f_1}{f_2}$$

where f_1 is computed as multivariate normal from the known distribution parameters and f_2 is the smoothing function. In conjunction with this modification, specialized use of the Information Divergence Program was made by computing the divergence between two samples from the same population. * A further intricacy, the relation of smoothing to point density, was then studied along with the problem of choosing a large enough sample size for representation of a population.

Using the population parameters of SDPS20, cluster 1, three data sets were generated, each containing two samples of size 200, 400, and 600 observations from the unimodal multivariate normal population. A basis for concluding that a sample is large enough to represent a population is a small divergence between two samples of that size. Results from this type of computation, using the hyperpyramid smoothing function for probability density computation, are shown in tabular and graphic form in Figure 4-12.

A sample size of 400 observations is seen to be much more reasonable for representation of the population than 200 observations. For 1-sigma smoothing, the divergence is further decreased to the order of 1 bit for a sample of 600. One-half sigma smoothing appears to be inadequate. The figure of about 1 bit or 1 nit of divergence is desirable. It can be interpreted crudely as an S/N ratio. In the case of two unimodal and normally distributed populations, the interpretation of signal and noise takes the following form: signal is the distance (generalized to the number of dimensions required) between modes, and noise is the square root of the sum of the two variances.

The reason for variation in the divergence results for $1/2\sigma$ and 1σ smoothing parameters in Figure 4-12 is an initial indication of the relationship of smoothing to sample size. The parameter should be just large enough to smooth the transition between adjacent training set points so that new points between these are assigned density values comparable to those assigned to the training points themselves. When, however, the parameter is too small, small peaks are introduced into the estimated density, determined by randomness in the selection of the training

*Derivation and discussion presented in Section 4.3.1.

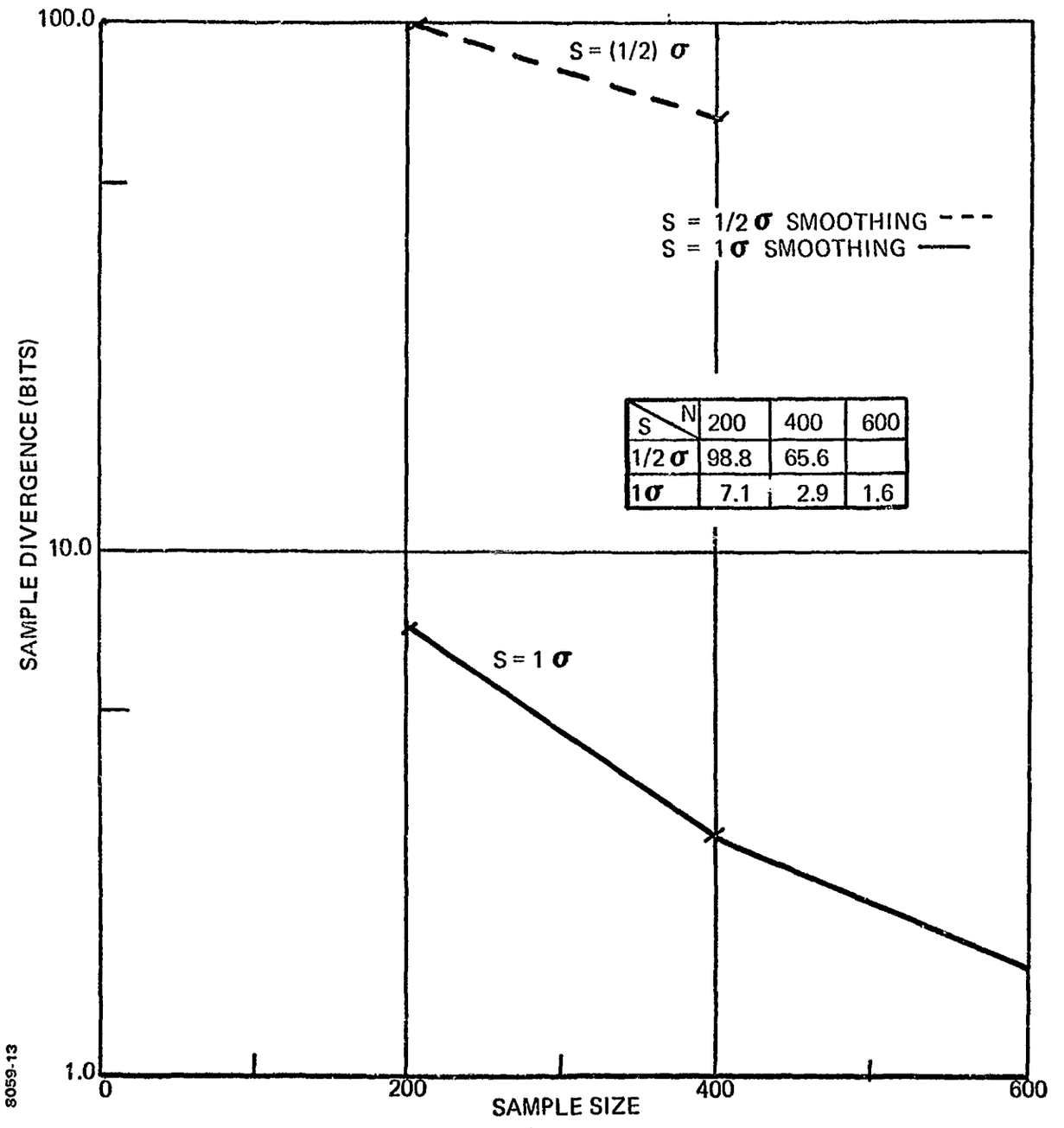


Figure 4-12 Sample Divergence as a Function of Sample Size

sample rather than by the actual distribution. That $1/2\sigma$ for smoothing was indeed too small in this problem is supported by the results from the density divergence computation shown in Figure 4-13. The following inverse covariance matrix was computed from one of the samples containing 200 observations used above:

$$\begin{pmatrix} 342.55 & -41.73 & -36.66 & 43.12 \\ -41.73 & 280.24 & 32.77 & -6.46 \\ -36.66 & 32.77 & 307.71 & -31.98 \\ 43.12 & -6.46 & -31.98 & 284.60 \end{pmatrix}$$

With f_1 based on the multivariate normal density function and this inverse covariance matrix, and f_2 as the hyperpyramid smoothing function, the density divergence computation was performed on three samples of sizes 200, 400, and 600 observations, respectively. Smoothing parameters of $1/2\sigma$ and 1σ were used. The divergence, then, between the actual multivariate normal density and the estimated density for all three sample sizes is much larger for a $1/2\sigma$ smoothing parameter than for the 1σ smoothing parameter. The density divergence is increasingly worse for both $1/2\sigma$ and 1σ smoothing as sample size decreases from 400 to 200 observations, but the density divergence for 600 signatures, point three on each graph, was worse than that for 400 signatures. This is explained by examination of the inverse covariance matrix computed from 600 observations in comparison to that from 200 observations presented above:

$$\begin{pmatrix} 291.54 & 29.10 & -13.26 & -23.73 \\ 29.10 & 316.58 & 5.71 & -22.39 \\ -13.26 & 5.71 & 278.54 & 33.28 \\ -23.73 & -22.39 & 33.28 & 268.82 \end{pmatrix}$$

An increase from 200 observations to 600 observations changes the inverse covariance matrix of the samples to the point that representation of the multivariate normal density of the 600-observation sample with the original matrix did not lend to an accurate measure of the divergence between the hyperpyramid estimate of the density and the actual density of the sample. As expected, use of the second matrix in the multivariate normal probability density computation brought the density divergence computation below that for the sample of size 400, point 4 on each graph.

The parametric representation used in f_1 for the point 3 computations was not describing the population from which the 600 samples were drawn. When the proper parametric representation was used, the computation gave the result plotted as point 4.

Note that the divergence between samples, plotted in Figure 4-12, is comparable for all sample sizes to the divergence between population and sample, plotted in Figure 4-13.

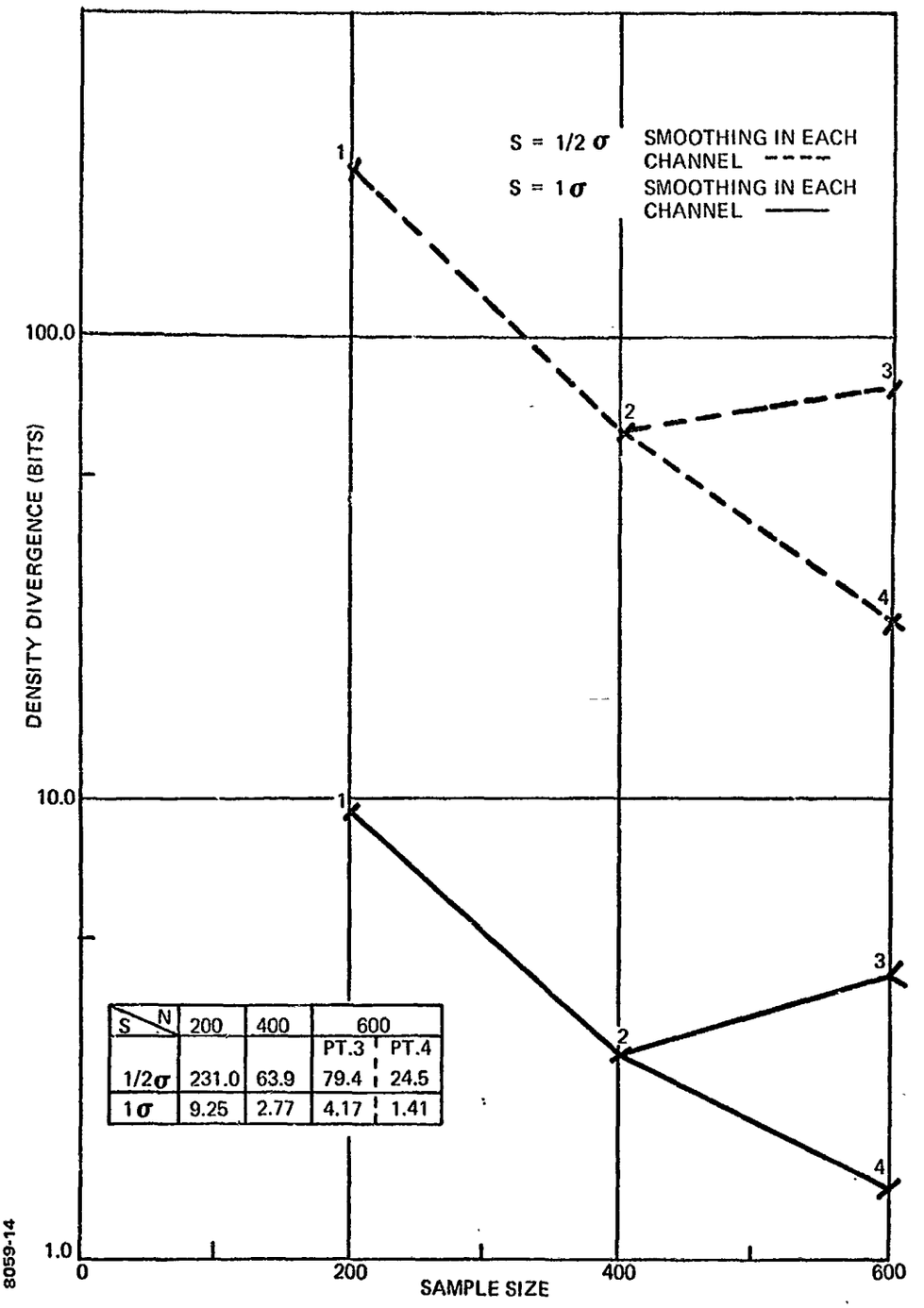


Figure 4-13 Density Divergence as a Function of Sample Size

While this study could be profitably carried much further, to investigate the effects of multiple modes, for example, and still larger samples, it is tentatively concluded that a smoothing parameter of dimension approximately equal to the dimension of a cluster is suitable smoothing. A smoothing parameter equal to the standard deviation appears to be about right. Given a suitable smoothing parameter, the divergence may be decreased smoothly by increasing the sample size. A sample of size 600 gives a reasonably small divergence for four-channel data. This conclusion is more critically evaluated when the result is applied to observations of real world data with multiple modes. It then becomes necessary to select smoothing parameters based on the size of clusters.

Note also that the simulated data were not factor analyzed before the divergence computations were performed. The smoothing was carried out, however, as though the variables were uncorrelated; the directionality of the smoothing was ignored. The results obtained in Figure 4-13, showing the divergence between population and sample, are therefore somewhat larger than if the data had been properly smoothed, i. e., factor analyzed.

The study of simulated data, presented in this section under the title Parametric Studies, is presented to document the work conducted to evaluate the software prepared and used in the balance of the study. Problems identified at this stage for later resolution have been studied first with simulated data to gain confidence that the software is properly performing the intended operations. Problems to be studied further in the following section include selection of smoothing parameters and the size of training samples used. These problems relate to both aspects of the study objectives. The two aspects are: (1) a computer evaluation of processing schemes based on maximum likelihood methods, using only samples of data drawn from a much larger population, and (2) the development of analog processing hardware to operate on the unknown parent populations to perform maximum likelihood processing.

The selection of smoothing parameters best for use in the computer portion of the study, if the results are to be implemented usefully, must correspond to realizable smoothing parameters realizable by the analog processing hardware. The linear transformation specified by computer analysis is also performed by the analog hardware. Every operation performed by the computer has an analogous operation performed on parent populations by the hardware to be specified. Section 6 discusses the latter aspect of the problem. The computer study described in Section 5 serves as a guide for design and performance objectives of the hardware system which will evolve.

SECTION 5

REAL-WORLD STUDIES

Studies performed on data taken from the real world parallel those performed on simulated data with known distributions. Proportionately more emphasis is placed on the studies of actual data; conclusions about the potential role and value of multispectral data processing ultimately depend upon the distributions nature provides. It is only by understanding the information contained in the differences between distributions of signatures drawn from different populations of interest that objective specifications for a data collection system and a data processing system can be written.

The spectral differences that contribute collectively to the increased information obtainable with multispectral methods of collection and processing do not necessarily bear the same relationship to an allowable noise level on the corresponding channel of the data collection instrument as would be required for a single-channel instrument. An appropriate approach to the problem of objectively stating performance requirements for collection and for processing is to think in terms of the effect of instrument noise on the information content of an observed sample distribution. When the dispersion contributed by the instrument noise is small with respect to the dispersion contributed by nature, as determined by multispectral measures of information content, then a further improvement in instrument performance will not improve the ability of the data processor to recognize and classify targets. The answer obtained by applying this criterion will of course change with the difficulty of the signature recognition problems of current interest. When problems of one level of difficulty are satisfactorily solved, attention will be directed to the next more difficult problems. It is reasonable, therefore, to acquire the best performance in a data collection device that current technology and available resources can provide.

As data from such devices become available, the problem of learning to extract whatever information is present and present it in usable form will be studied with the best possible data. The study of such data will show the hierarchic difficulty of those problems which can be solved. Some of them will be identified as relatively simple and solvable with data collected by an instrument which does not perform to state-of-the-art specifications. Operational instrumentation should then be designed to the most cost-effective specifications; this will mean reduced performance standards in some cases.

Conclusions drawn from analysis of data currently available are less general in applicability. The sources of dispersion in the data include contributions from instrument noise and from nature. While the contributions from instrument noise appear to be smaller than those from nature, there is no way to determine quantitatively the relative sizes of the contributions. Two things are certain: (1) the recognition of agricultural crop type signatures, which comprise the real data used in this study, is not among the most difficult of signature recognition problems, and (2) the instrument used to collect the data studied does not have the best performance that current technology can provide - its performance, however, appears to be adequate for the problem considered.

The techniques developed during this study are to evaluate the information content of data. Performance achieved is not to be generalized as an indication of the limits of multispectral processing. This generalization will require analysis and experience with a much larger variety of data and the use of data collection instruments which approach technology-limited performance.

This section discusses the study of real-world data performed for evaluative purposes. Sections 5.1 through 5.3 indicate how data sets were generated, the statistical analysis applied to the data sets, and the results of the analysis.

5.1 PILOT DATA SETS

Multispectral data collected over agricultural areas near Weslaco, Texas, were examined to provide sufficient known signatures for analysis. The data collection instrument was an eight-channel multispectral line scanner with channels from the UV to the IR as illustrated in Figure 5-1. Bendix Aerospace Systems Division, under contract to NASA/MSC (Contract No. NAS 9-9434), conducted four flights during the 1969 growing season over the test sites maintained near Weslaco, Texas, by the Agricultural Research Service. Figure 5-2 shows the optical schematic of the scanner; Figures 5-3 and 5-4 are photographs of the scanner. Table 5-1 gives some of the performance characteristics.

Of interest in the scanner system is the technique used for calibration, control, and dynamic range compression of the electronic signals. The technique used is considered a major technological advance in this type of scanner. Calibration sources are provided for dark reference and illumination level reference. Signals from these calibration sources are injected into every channel on every scan line. The illumination level reference used is diffuse solar (and sky) irradiance brought into the scanner from above the aircraft through a fiber-optics bundle terminating below a diffusing surface.

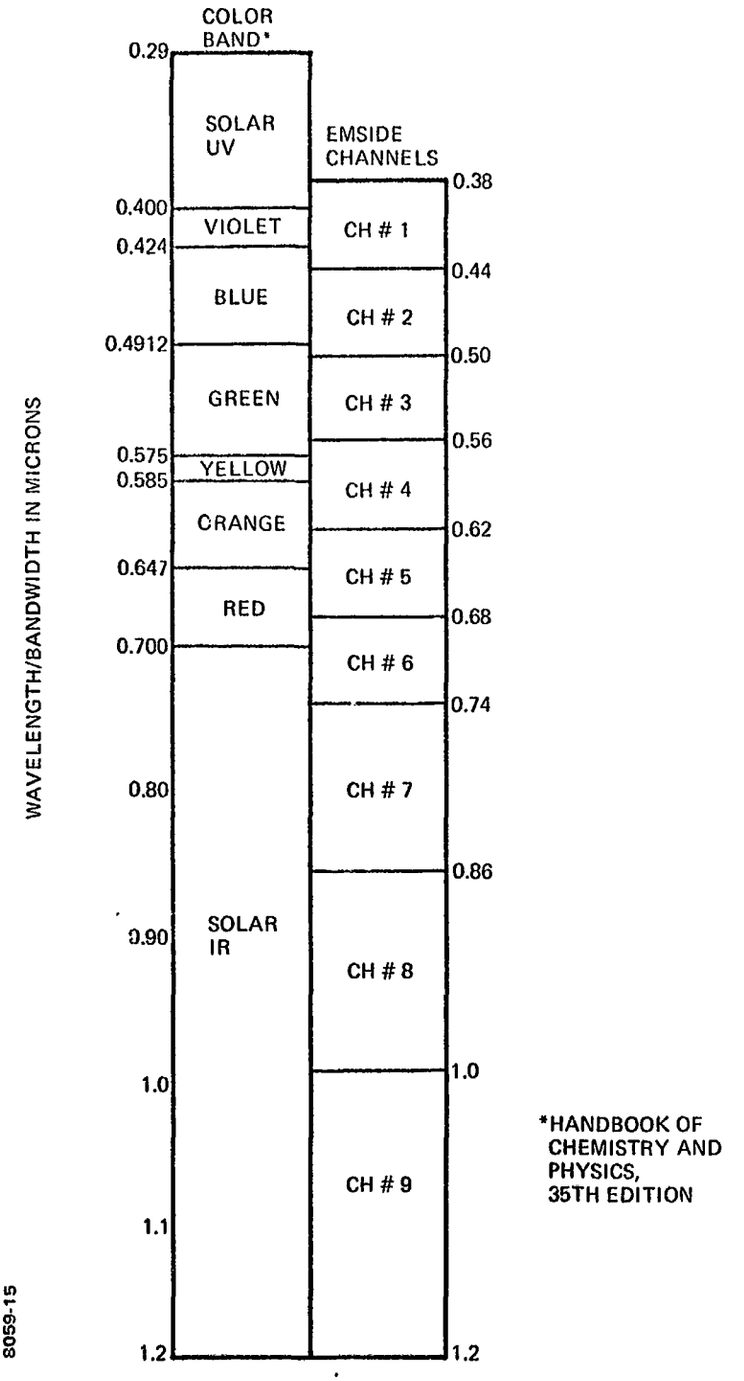


Figure 5-1 EMSIDE Channel Locations vs. Color Band

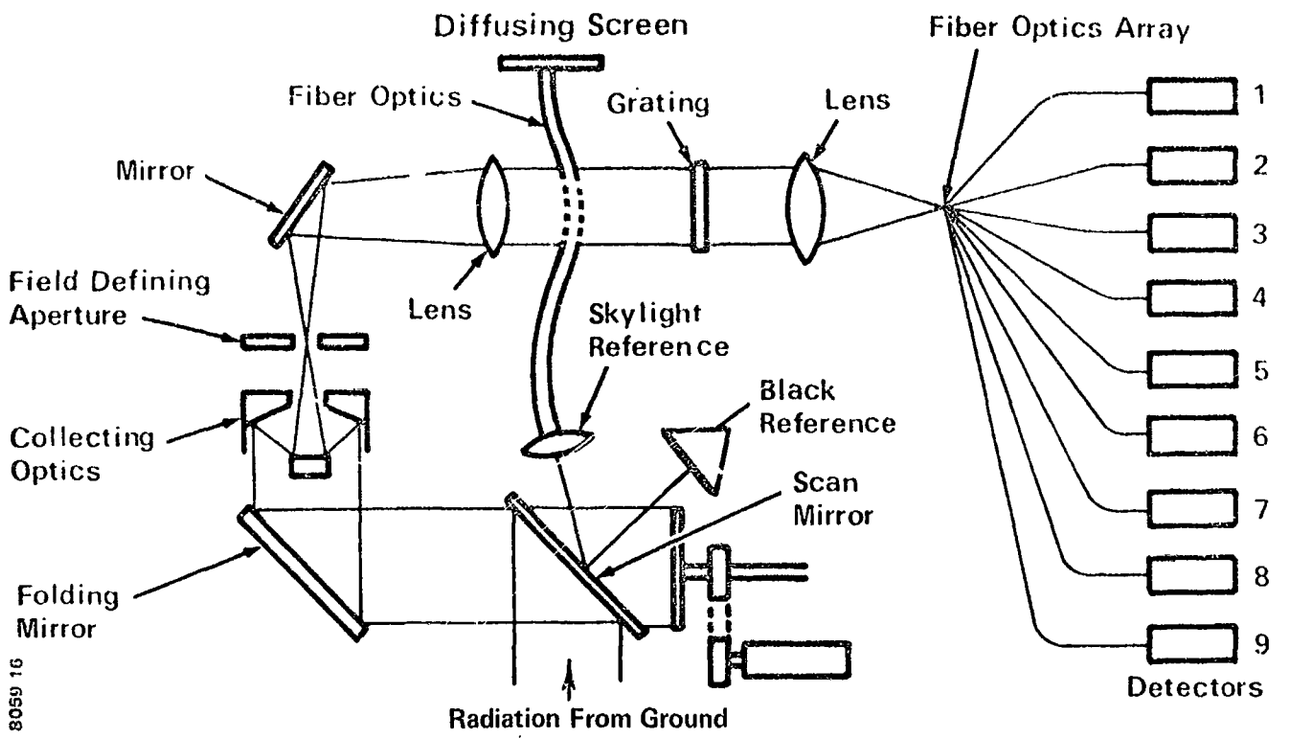


Figure 5-2 Multispectral Scanner Pictorial Layout

REPRODUCIBILITY OF THE ORIGINAL PAGE IS POOR

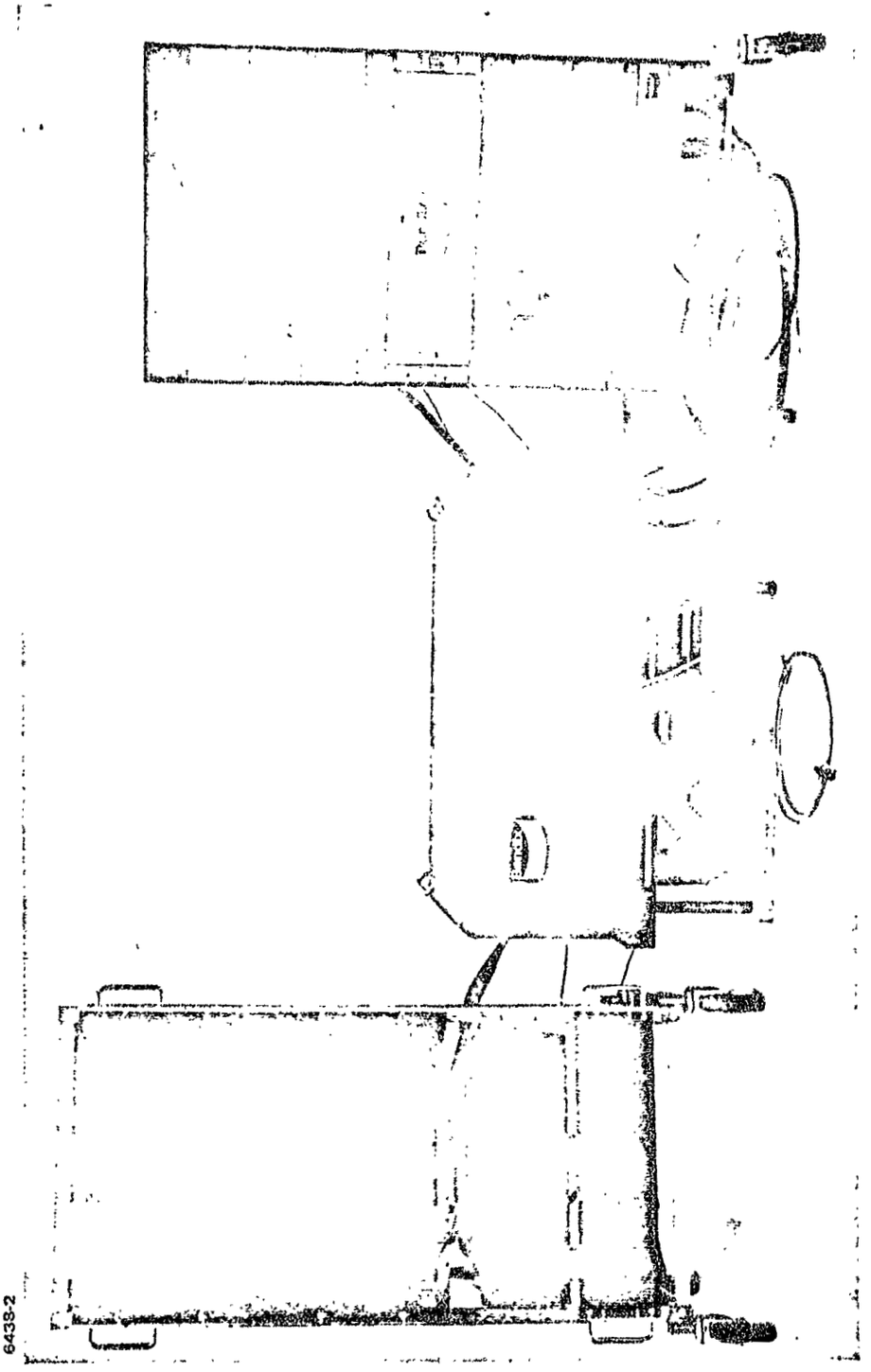
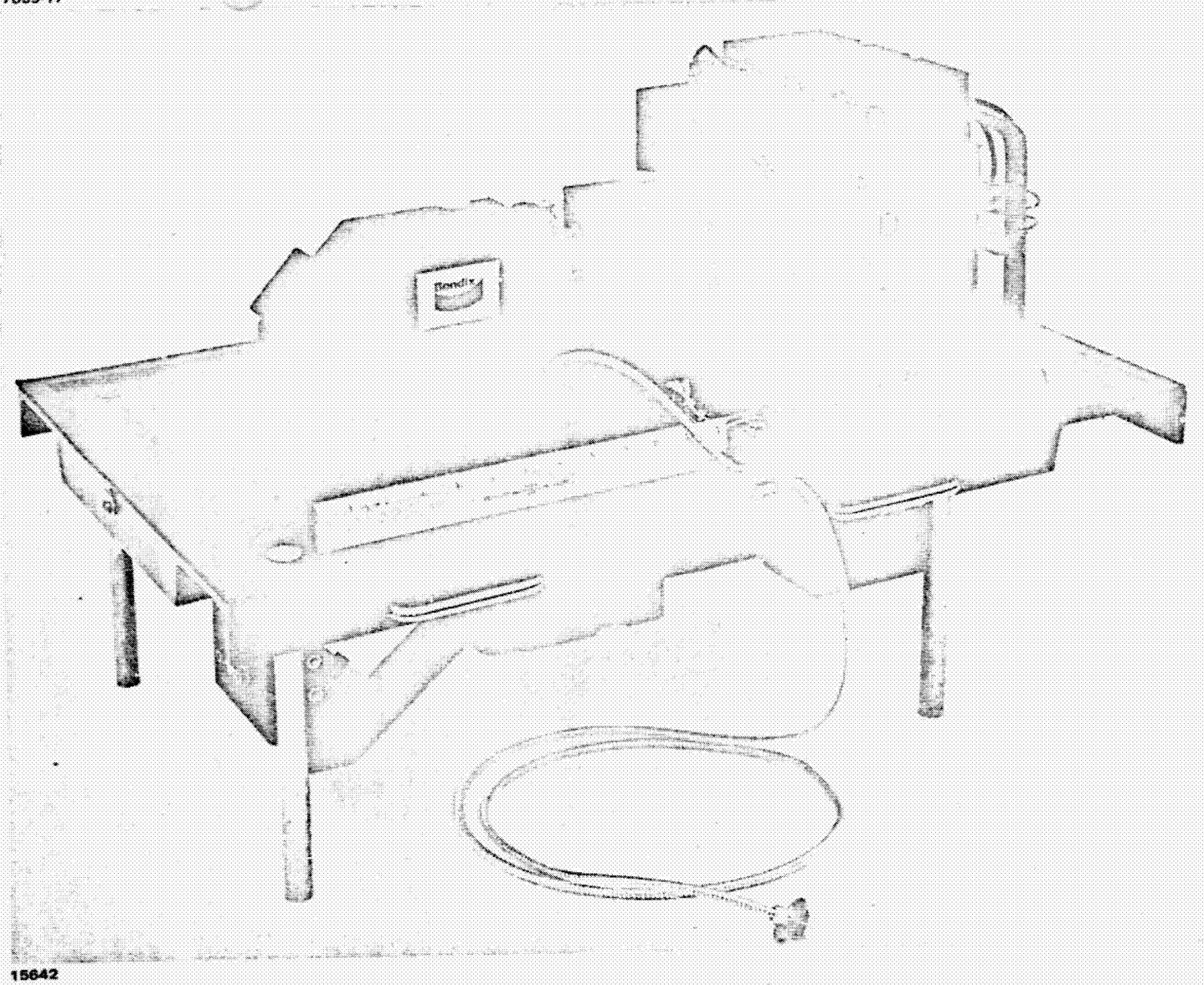


Figure 5-3 Multispectral Scanner Airborne System



15642

Figure 5-4 Multispectral Scanner, Cover Removed

TABLE 5-1

PERFORMANCE CHARACTERISTICS

Channels (Number)	Spectral Band (Microns)	Detector
1	0.38 – 0.44	S-20 PM
2	0.44 – 0.50	S-20 PM
3	0.50 – 0.56	S-20 PM
4	0.56 – 0.62	S-20 PM
5	0.62 – 0.68	S-20 PM
6	0.68 – 0.74	S-20 PM
7	0.74 – 0.86	S-1 PM
8	0.86 – 1.0	S-1 PM

Scanning rate (nominal)	100 revolutions/sec
Instantaneous Field-Of-View (Spatial Resolution)	2.5 milliradians
Lateral Scan Angle	120 degrees
Output Each Channel	Terrain Spectral Reflectance
Spectral Reflectance Accuracy	1.0 percent
Design V/h (aircraft velocity f.p.s/ altitude in feet)	0.25
Analog Tape Recorder	Ampex AR 1600
Tape Running Time (per reel)	30.0 min.
System Weight (Scanner, Electronics, Recorder)	400 lb.
Power Requirements (Scanner, Electronics, Recorder)	27 a @ 28V DC 7.5 a @ 115V, 400 cps Single Phase

Calibration of each channel is accomplished by a closed-loop electronic servo which automatically compensates for spectral and intensity variations in terrain illumination and variations in the optical and electronic characteristics of the scanner. The video output from each of the nine channels is calibrated to yield directly terrain spectral reflectance. System electronics permit selection of the total reflectance range sent to the tape recorder. This is a vital feature because tape recorder dynamic range is the limiting factor of the system dynamic range.

The Agricultural Research Service provided ground truth for the flight data used during the study. A summary of the information is found in Report 69-03.* A more detailed report on the ground truth for this period is found in a report by the same authors and organization.**

An initial survey was made of the number and conditions of various crops in the area as indicated by the ground truth summary of Site 32. A simple histogram was compiled showing total number of fields of various crops in the six flight lines where data were acquired. The three most numerous classes were cotton, sorghum, and citrus. These three, with the addition of bare soil, were chosen for primary analysis.

Imagery was generated from the analog data, and several fields of interest were located with the aid of corollary aerial photography in the ground truth summary. "Fields of interest" were defined as those given a "good" or "excellent" rating in the summary. Samples were taken from the analog data and written in IBM-compatible format as a digital tape.

Imagery is prepared during the sampling procedure. During the sampling period, a video blanking signal causes a line to appear on the imagery. Examples

---

* Alvin H. Gerbermann, Joe A. Cuellar, and Ross W. Leamer; Ground Truth Summary, April - July 1969, Site 32, Weslaco, Texas; Report 69-03; Rio Grande Soil and Water Research Center, Soil and Water Conservation Research Division, Agricultural Research Service, US Department of Agriculture, Weslaco, Texas; 1 Dec 1969

** Alvin H. Gerbermann, Joe A. Cuellar, and Ross W. Leamer; Detailed Ground Truth, April - July 1969, Site 32, Weslaco, Texas; Report 70-01; Rio Grande Soil and Water Research Center, Soil and Water Conservation Research Division, Agricultural Research Service, US Department of Agriculture, Weslaco, Texas, 2 March 1970.

of imagery prepared during digital sampling are shown in Figure 5-5. The thin black line which appears along the length of most of the imagery is due to a video blanking signal. The video blanking signal is applied to all channels of video once each scan line for a 5-μsec period. The multiple channel A/D converter samples the video from all channels simultaneously at this point in the duty cycle. The blanking signal identifies in the real-time laboratory imagery display the line along which digital samples are collected. The position of this line in the imagery may be adjusted until the cursor line produced by the blanking signal passes through the regions which are to be sampled. The blanking signal is also applied to the film recorder input with one exception. During the actual sample recording times, the blanking signal is removed. The locations sampled and recorded on the digital tape are thus found in the illustrations on a straight line filling the gap in the cursor line, i.e., that line due to the video blanking signal.

The numbers and letters on the fields in Figure 5-5 are the identification provided on the aerial photography in the Ground Truth Summary. They were annotated on the imagery produced from the scanner video in Figure 5-5. The line numbers are flight lines drawn on the map controlling the conduct of the mission and along which aerial survey data and ground truth data were collected. The examples are taken from lines 12 and 13. The ID numbers are data control numbers. These are appended to each digital sample recorded on tape. A single sample is comprised of an ID number determined by thumb wheel settings on the digitizer, and an 8-bit number from each of eight channels simultaneously sampled. SDP software edited the digital tape, rejecting all samples containing one or more channels with an under or overflow. The 22,000 unedited samples were compared with imagery made during the time of sample for the purpose of isolating those samples representing the four classes of interest. A scan line count, as well as a usable sample case number, is supplied by the tape editing software. This is called a case number. Boundaries were identified for each field. Case numbers of samples clearly established as having originated from a particular field were obtained for further data selection.

The four classes of interest, cotton, sorghum, citrus, and bare soil, existed in varying conditions as indicated in the following summary.

The "crop type" refers to ground truth information provided by the Weslaco Agricultural Research Service:

1 = Excellent
2, 4 = Less than optimum crop coverage
2, 5 = Uneven plant height
2, 12 = Recently watered
2, 13 = Debris present
2, 14 = Crusted surface.

The ID number is written on the magnetic tape by the A/D converter at the time of digitizing.

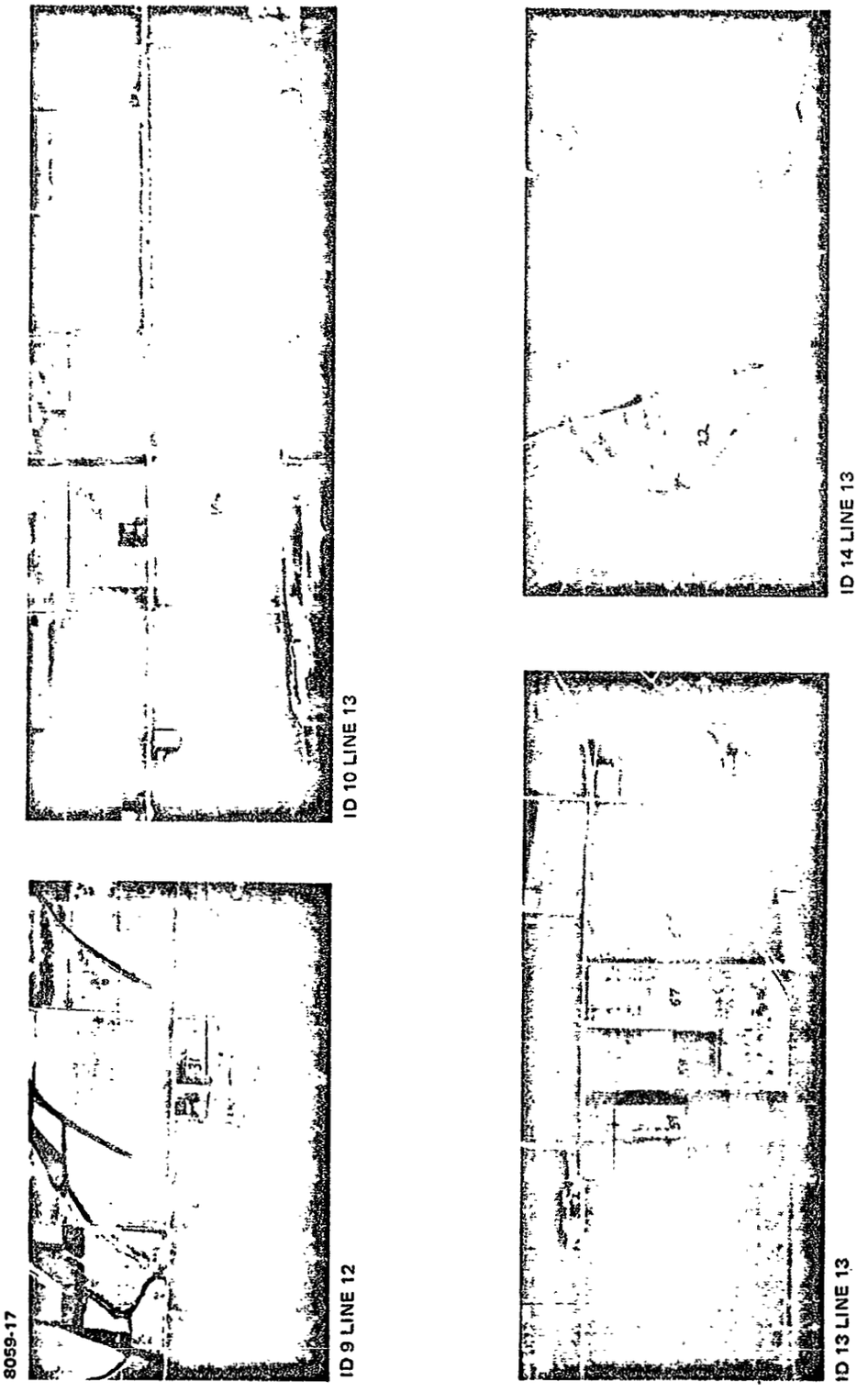


Figure 5-5 Digital Sampling Imagery

The field and line number were provided by the ground truth summary. They indicate the exact position of the various fields. The case number indicates the order of the digital samples after the digital tape has been edited. (Editing consists of discarding all spectra which contain a zero or 252, corresponding to underflow or overflow of the A/D converter).

With this information, each sample may be uniquely associated with a particular field and thus identified as per the ground truth information.

Tape Editor Output Ground Truth: Data Set ASD. P124. WESL0001

Crop	Crop Type	ID	Field	Line	Case No.
Cotton	1	10	83	3	25-91, 124-165
	(2099)	11	82	3	313 - 489
		16	38	3	2320 - 2452
		31	32	12	12655 - 13194
		32	23	12	13305 - 13921
		35	103	13	18523 - 19006
		30	89 B	13	21484 - 21530
Cotton	2, 5	13	69	3	1605 - 1649
	(1409)	18	13	3	4141 - 4449
		23	18	11	8798 - 9005
		33	10	12	14445 - 14621
		37	107	13	20143 - 20242
		30	89 A	13	21645 - 22220
Cotton	2, 4	16	39 B	3	2320 - 2452
	(426)	23	23	11	8108 - 8292
		27	2 B	11	10538 - 10648
Cotton	2, 12	18	10	3	4493 - 4801
	(1853)	18	9 C	3	4823 - 4999
		19	12	3	5052 - 5139
		29	43	12	11360 - 11498
		30	38	12	11955 - 12258
		31	34	12	12347 - 12567
		34	118	1	16501 - 10662
Total Cotton = (5787)		35	102	13	19029 - 19489

Tape Editor Output Ground Truth: Data Set ASD. P124. WESL0001 (Cont.)

Crop	Crop Type	ID	Field	Line	Case No.
Sorghum	1	12	74 A	3	1048 - 1312
	(2566)	15	52	3	2060 - 2195
		16	39 A	3	2474 - 2562
		20	54	11	5862 - 5950
		23	20	11	8592 - 8752
		25	13	11	9490 - 9896
		27	2A	11	10670 - 10758
		28	49	12	10859 - 11079
		29	41 D	12	11613 - 11705
		33	8	12	14709 - 15083
		33	73	12	15171 - 15589
		37	109	13	19901 - 20099
Sorghum	2, 5	15	53	3	2262 - 2319
	(497)	26	9 A	11	9897 - 10337
Sorghum	2, 12	23	21 B	11	8293 - 8361
	(1233)	27	2 C	11	10472 - 10505
		29	41 B	12	11843 - 11889
Total Sorghum = (4296)		33	12	12	14115 - 14401
		34	115	13	16708 - 17490
Citrus	1, 23	17	25	3	3708 - 3862
	(585)	17	22	3	3972 - 4060
		21	39	11	6873 - 7245
Citrus	2, 23	22	28	11	7959 - 8074
	(709)	20	52	11	5972 - 6170
		20	50	11	6214 - 6610
			77	3	586 - 1014
	2, 4, 23 (468)	12			
	2, 4, 23 (368)	22	31	11	7568 - 7936

Total Citrus = 2091

Tape Editor Output Ground Truth: Data Set ASD.P124. WESL0001 (Cont.)

Crop	Crop Type	ID	Field	Line	Case No.
Bare Soil (748)	1	24	17	11	8990 - 9430
		39	97	13	21148 - 21434
		39	96 B	13	21457 - 21479
Bare Soil (1851)	1, 14	34	119	13	15696 - 16455
		34	114	13	17651 - 17973
		36	113	13	19547 - 19877
		38	104	13	20657 - 21097
Bare Soil (276)	1, 16	35	106 A	13	18201 - 18477
(184)	2, 13	35	106 B	13	17994 - 18178
(374)	2, 12, 17	17	27 A	3	3114 - 3488

Total Bare Soil = (3433)

5.2 PRINCIPAL COMPONENT TRANSFORMATIONS (FACTOR ANALYSIS)

Linear transformations are used in this study of the information content and methods for recognizing it in the belief that such transformations, properly chosen, simplify many of the data handling problems associated with multispectral data. The linear transformations applied in this section are the principal components obtained during a factor analysis. The new variables obtained by applying such a transformation are linear combinations of the original variables of the measurement space.

The properties of the principal components, or factor scores as they are sometimes called in this report, that make them of interest are:

1. They are uncorrelated, which simplifies the selection of smoothing parameters and improves the possibility of expressing multivariate probability density functions as the product of univariate or bivariate probability density functions. The property of uncorrelatedness, if accompanied by statistical independence, allows the density functions to be factored. Neither of these benefits is realized in the common transformation work which is described first. When a common transformation is determined by factor analyzing together, samples drawn from several categories, the resulting transformation does not produce variables which are uncorrelated in any single category. The smoothing parameters best suited to each category do not, under such conditions, coincide in direction with the direction of the coordinate system axes. The matrix of smoothing parameters is not diagonal. This also has the effect of causing the probability density function to be unfactorable. These benefits are realized, however, in the multiple transformation studies described in Section 5.2.2.

2. They are also useful in some degree for reducing the dimensionality of the multivariate classification problem. To the extent that the directions of maximum variance in the data are also the directions of highest information content, the dimensionality of the multivariate problem can be reduced by using only enough principal components or factor scores to account for a large percentage of the variance in the data. This procedure certainly retains information of value. When the largest contributions to the total variance in the data are due to the spectral differences of real value for the desired categorical distinctions, the largest principal component direction also contains the largest amount of useful information. This condition is more nearly realized in the common transformation work than in the multiple transformation work.

In the common transformation work, a single factor analysis is conducted. The sample so analyzed contains samples from each of the categories of interest. In the multiple transformation work, samples from each category are analyzed separately.

It has not been possible to explore adequately all questions concerning the relative merits of the transformations that are available. At the initiation of the study program, the plan was to compare the relative merits of various methods of extracting information from multivariate observations and determine the best of them. This would have required picking the best of a bad lot if strictly followed. As the deficiencies of each approach considered became more obvious, it became rather unproductive to attempt a comparison. Attention was shifted instead to the development of transformations and procedures not suffering from such deficiencies.

The order in which the transformations and processing methods are described is the same order in which the corresponding methods were conceived and implemented.

The deficiencies identified during the work with principal component transformations, both common and multiple, are directly addressed in Section 5. 3. The data provided for canonical analysis are sufficient to allow the statistical analysis to distinguish between noise-like variance and variance which expresses useful information. The procedure devised is similar to the multiple transformation type of factor analysis.

5.2.1 Common Transformation

The initial investigations with real-world signature data center around establishment of the reasonableness of results obtained from Bayesian decision processing and information divergence computations. In instances where transformed data are analyzed, the factor space in which they are expressed, hereafter referred to as the common transformation space, is that which is resultant from the factor analysis of a data set containing samples of all four targets. That is, an orthogonal factor analysis is performed on a data set containing 500 randomly selected signatures of each of the four crop types: cotton, sorghum, citrus, and bare soil. The resultant factor score coefficient matrix is then used to transform linearly signatures into a new frame of reference, ready for Bayesian decision function and information divergence computations aimed at separation of the targets.

Such an analysis produces a first principal component which accounts for 74% of the combined target variance, with the second describing an additional 19%. If high variance content ensures high information content, it would be expected that this transformation to the factor space allows a reduction in dimensionality of the data without significant information loss.

The scatter diagrams of 1000 observations of each crop after transformation into the common space are presented in Figures 5-6, 5-7, 5-8, and 5-9. Note the trimodal nature of bare soil and the similarity of cotton and sorghum.

5.2.1.1 Classification and Divergence Results

In preparation for Bayesian decision classification, random number methods are used to select two groups of 200 signatures each from all four crops. This is done for both transformed and nontransformed data. Group one of each crop serves as a training set to represent the crop population in the analysis. The 800 signatures resulting from the combination of the second groups from each crop serve as sample observations to be classified.

At the base of the Bayesian and information divergence analyses is the computation of the probability density of a sample signature when conditionally attributed to each target. This computation may be performed in one of several ways, some requiring assumptions about the nature of the target population. For example, an assumption that the target distribution is multivariate normal leads to the applicability of the following function for the probability density:

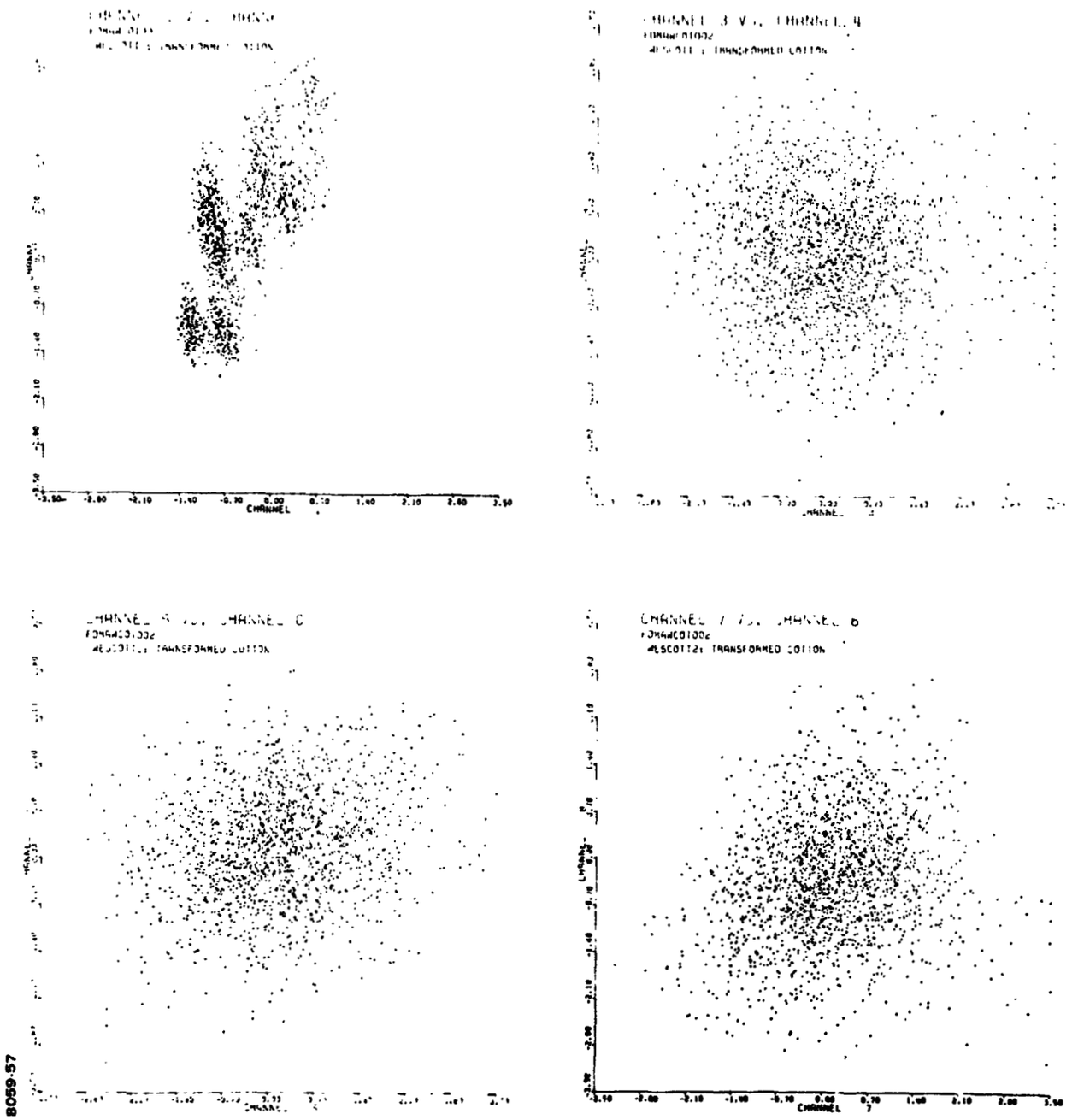


Figure 5-6 Scatter Diagrams of Transformed Cotton

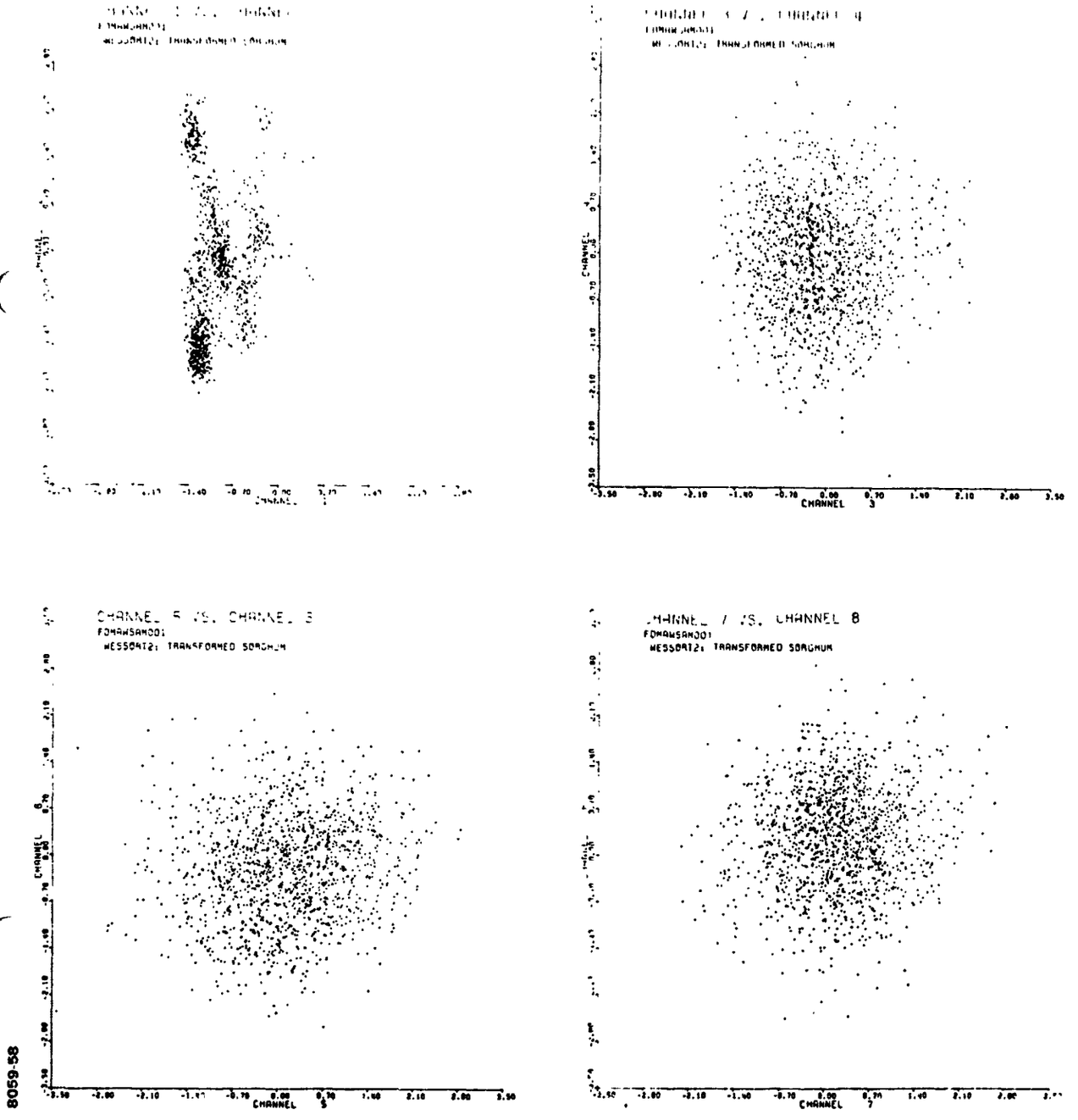


Figure 5-7 Scatter Diagrams of Transformed Sorghum

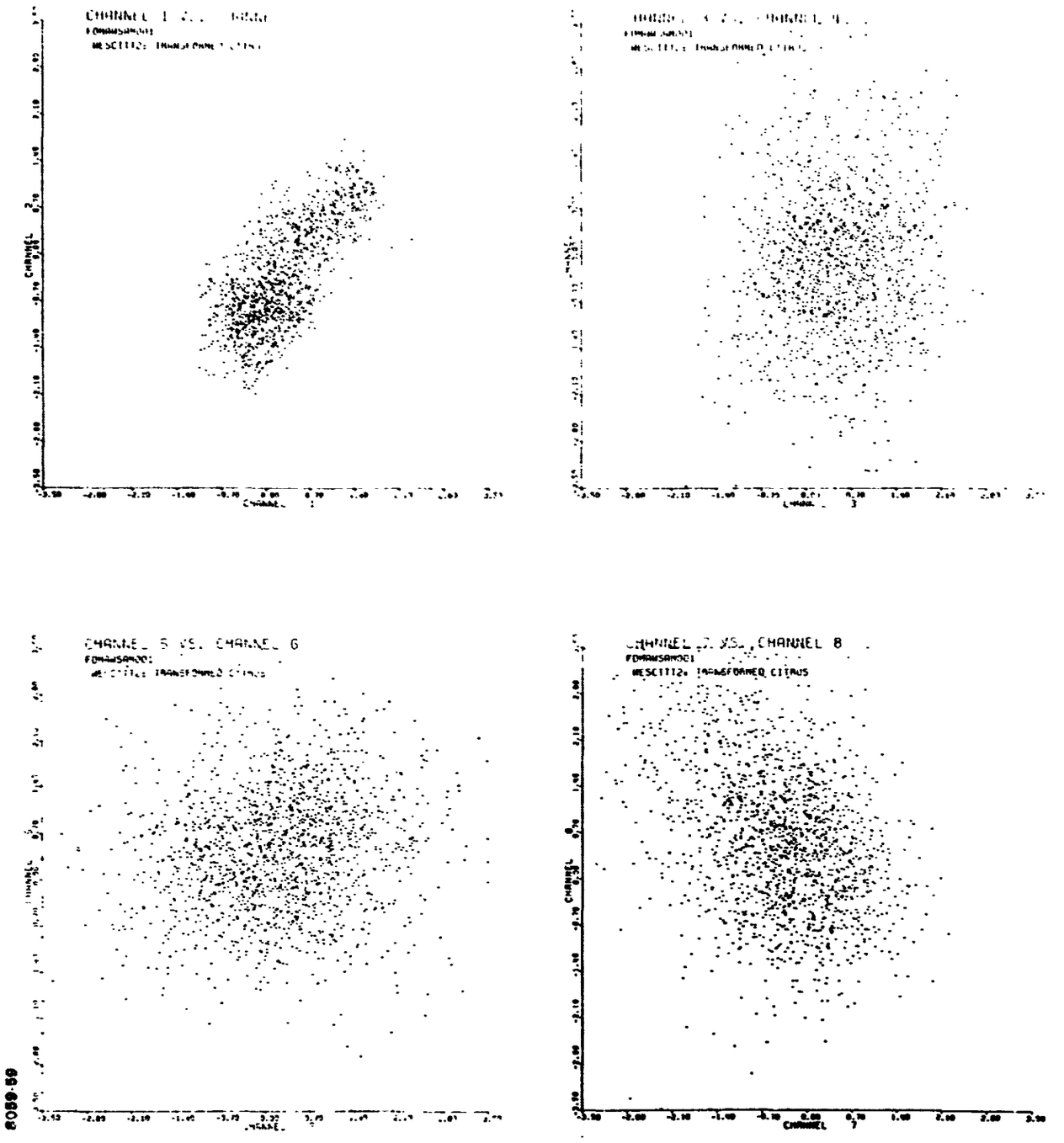


Figure 5-8 Scatter Diagrams of Transformed Citrus

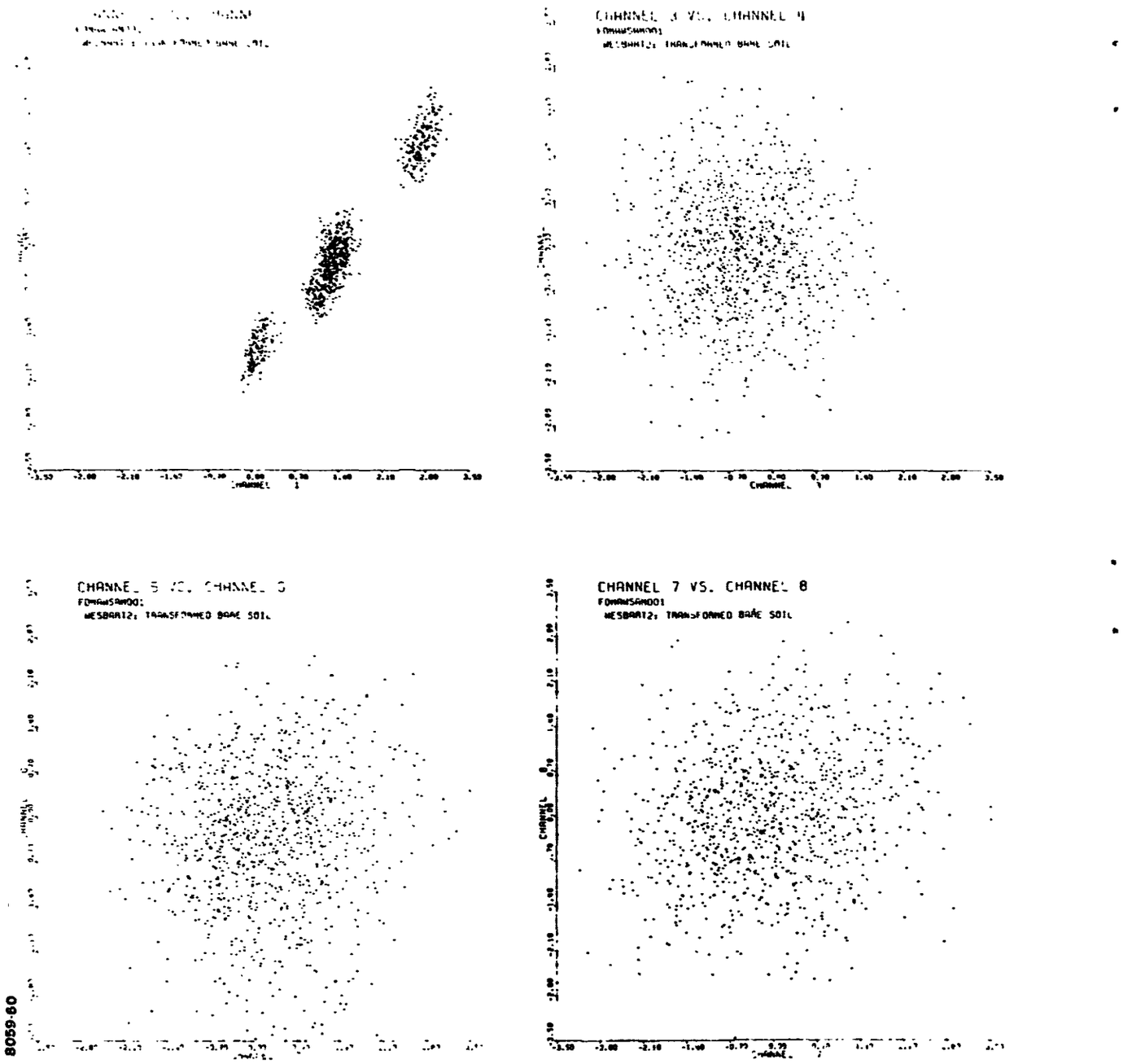


Figure 5-9 Scatter Diagrams of Transformed Bare Soil

$$f(\vec{x}) = \frac{\det(A)^{1/2}}{(2\pi)^{m/2}} \exp\left[-1/2\ Y^T\ A\ \vec{Y}\right]$$

where m is the number of channels, $\vec{Y}$ is the signature $\vec{x}$ expressed about the training set mean, and A is the inverse covariance matrix of the training set. Computation of the probability density based on a training set without any such underlying assumptions however, is obtained by the use of a smoothing function and smoothing parameters, as presented in Section 4.3. The Specht function is a smoothing method for weighting the contribution of each point in a training sample to the total density function evaluated at a point not in the sample. The Specht function is:

$$f(\vec{x}) = \frac{1}{(2\pi)^{m/2}\, N\, (S_1\, S_2 \cdots S_m)^{1/2}} \sum_{j=1}^{n} \exp\left[-\sum_{i=1}^{m} \frac{(X_i - Y_{ij})^2}{2S_i}\right]$$

N is the number of signatures in the training set, m is the number of channels, Y_{ij} is the i^{th} component of the j^{th} signature in the training set, and S_i is the smoothing parameter for the i^{th} component of the training set signatures.

Choice of Specht smoothing parameters for this study neglects careful consideration of local point density, shape and number of modes in the distribution, or direction along which smoothing should take place. Results obtained with the common transformation are discouraging and an evaluation of status, based on these results, dictated the more careful study described as the multiple transformation in the next section. The discouraging results are not due to the nature of the transformation alone. In retrospect, the small size of the training samples used is probably a more important explanation. In processing of both transformed and raw data with the smoothing function, S_{ik} is chosen such that $S_{ik} \approx \sigma_{ik}$, where i is the channel index and k is the training set index. With larger sample sizes, smaller smoothing parameters could have been used.

Classification results for non-transformed data with probability density computation based upon all eight channels entering into the Specht function are shown in the top of Table 5-2, with true categorization indicated in the middle. The Bayesian strategy was based upon equal a priori probabilities of 0.25 for each target and equal costs of misclassification. Accuracy of the results varies from 54% for cotton to 90% for sorghum. The corresponding table of confusion for transformed data, that is with both training and sample data represented in the common transformation space and computation based upon eight variables in the Specht function, are shown in the bottom of Table 5-2. The differences in the results of the raw data and the transformed data in Table 5-2 are not due to the transformation itself,

TABLE 5-2

RAW DATA CLASSIFICATION, 8-CHANNEL SPECHT FUNCTION

Source / Assignment	Cotton	Sorghum	Citrus	Bare Soil
Cotton	108	18	5	0
Sorghum	89	180	47	1
Citrus	3	2	146	20
Bare Soil	0	0	2	179
	54%	90%	73%	89%

Truth Table for Accurate Classification

Source / Assignment	Cotton	Sorghum	Citrus	Bare Soil
Cotton	200	0	0	0
Sorghum	0	200	0	0
Citrus	0	0	200	0
Bare Soil	0	0	0	200

Transformed Data Classification, 8-Factor Specht Function

Source / Assignment	Cotton	Sorghum	Citrus	Bare Soil
Cotton	117	22	12	0
Sorghum	65	167	12	0
Citrus	13	11	160	20
Bare Soil	1	0	5	170
	58%	83%	80%	85%

but are attributable to the differences in the effect of smoothing parameters in the two spaces. Smoothing in the direction of the axes of the hyperspace defined by the spectral boundaries of channels, has a differing effect than smoothing in the direction of principal component axes owing to differing directional orientation of variation in the two hyperspaces.

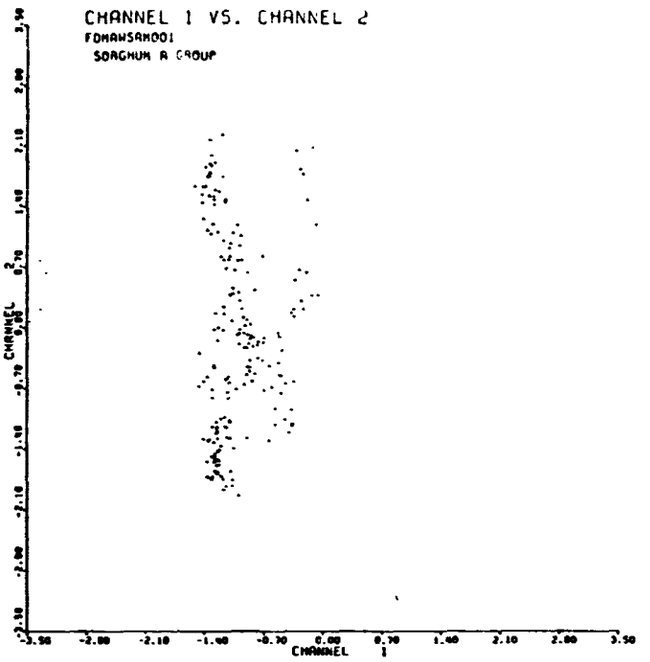
nprovement of the above results based on sm~othing function probability density computation depends in part upon choice of training set sample size. Certainly, 200 observations for representation of target populations is inadequate as can be supported alone by a comparison of the sparse training set scatter diagrams, Figure 5-10, with Figures 5-6 through 5-9.

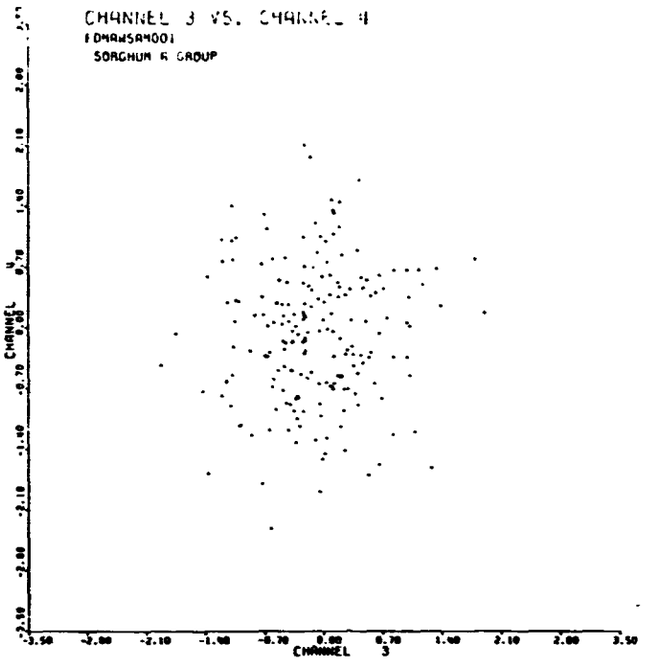
Use of 600-observation-target training sets began with the testing of the multivariate normal assumption with transformed data. Classification results obtained from the Bayesian Decision Function Program with eight components of transformed data entering into the multivariate normal probability density computation are shown in the top of Table 5-3, with truth tables for comparison in the center. The information divergences calculated also with the eight factor multivariate normal functions computed between the samples and whose classification results are presented are generally consistent with the results. Comparing the divergences in the bottom of Table 5-3 with the misclassification of data, it is noted that those samples for which there is the smallest divergence in bits are, in most cases, those for which misclassification is highest.

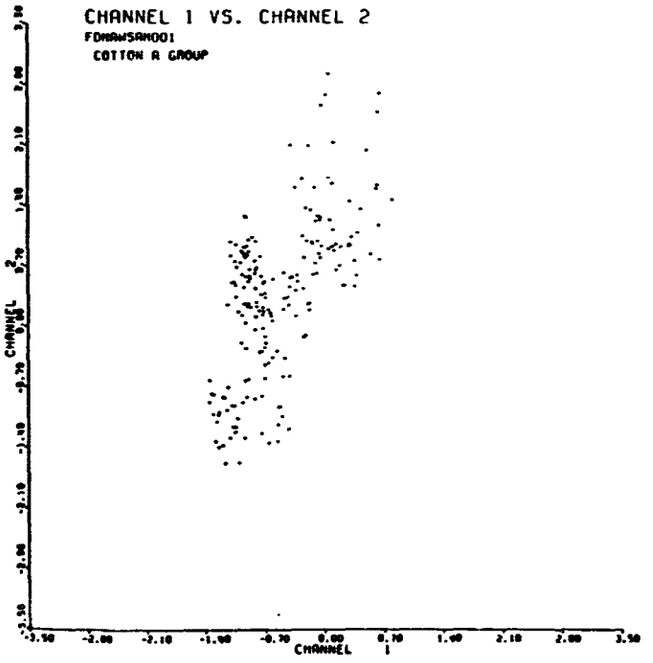
Any conclusion as to the usefulness of the common transformation for reduc.ion of dimensionality is premature if based upon the four-factor multivariate nc.rmal classification and divergence results following. In Table 5-4 those results are displayed. While three columns of the confusion table remain very close to those in Table 5-3, note that a reduction from eight factors to four factors entering into the multivariate normal probability density function reduces the classification accuracy for cotton from 66% to 41%. The cotton-sorghum divergence with eight components is 4.9 bits, while with four components the same is 2.8 bits. This is indicative of the degeneration of classification results which occurs between the same two crops. In the eight-factor table, 176 cotton samples are classified as sorghum, but in the four factor table 270 cotton are labeled as sorghum.

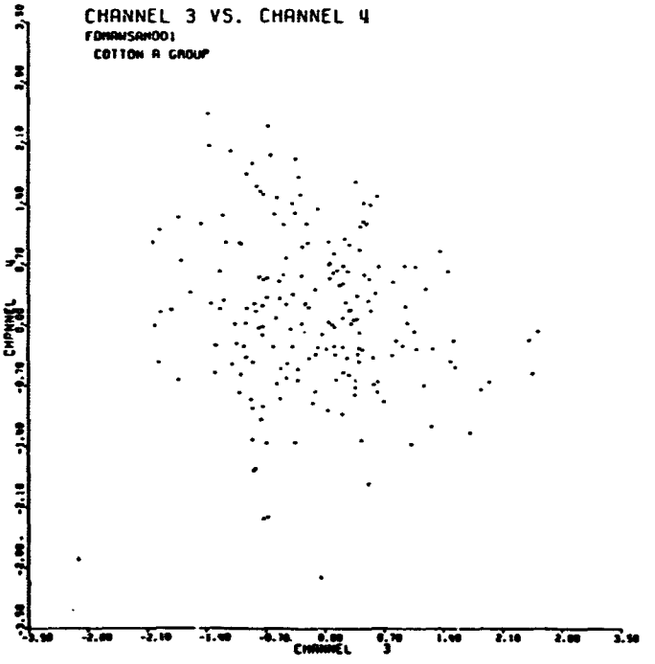
5.2.1.2 Critique

Classification success with a smoothing function for density computation can not be brought up to expectation without careful selection of smoothing parameters essential to correct representation of the target population. As presented in Section 4.3.2, the first step toward controlled smoothing is transformation of the data into a coordinate system in which the new variables are uncorrelated. Such a


CHANNEL 1 VS. CHANNEL 2
FDMAWSAM001
SORGHUM A GROUP
CHANNEL 2
CHANNEL 1


CHANNEL 3 VS. CHANNEL 4
FDMAWSAM001
SORGHUM A GROUP
CHANNEL 4
CHANNEL 3


CHANNEL 1 VS. CHANNEL 2
FDMAWSAM001
COTTON A GROUP
CHANNEL 2
CHANNEL 1


CHANNEL 3 VS. CHANNEL 4
FDMAWSAM001
COTTON A GROUP
CHANNEL 4
CHANNEL 3

8059-61

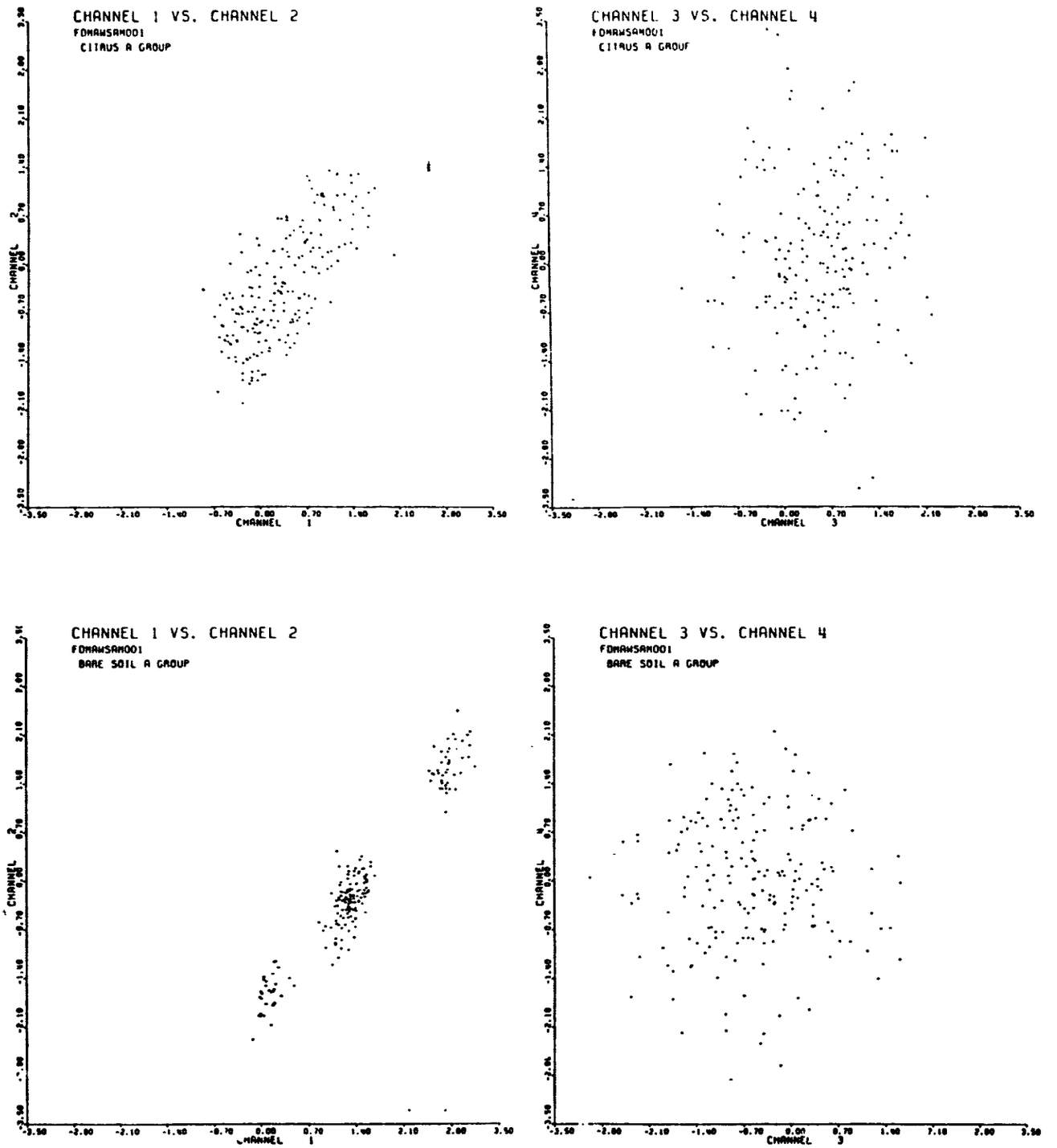


Figure 5-10 Scatter Diagrams of Training Sets

TABLE 5-3

EIGHT-FACTOR MULTIVARIATE NORMAL CLASSIFICATION RESULTS

Sample / Assignment	Cotton	Sorghum	Citrus	Bare Soil
Cotton	366	85	34	6
Sorghum	176	476	28	0
Citrus	48	38	517	10
Bare Soil	10	1	21	584

Truth Tables

Sample / Assignment	Cotton	Sorghum	Citrus	Bare Soil
Cotton	600	0	0	0
Sorghum	0	600	0	0
Citrus	0	0	600	0
Bare Soil	0	0	0	600

EIGHT-FACTOR MULTIVARIATE NORMAL DIVERGENCE RESULTS

Divergence	Cotton (bits)	Sorghum (bits)	Citrus (bits)	Bare Soil (bits)
Cotton	0.2	4.9	13	59
Sorghum	4.9	0.2	22	97
Citrus	13	22	0.2	26
Bare Soil	59	97	26	0.2

TABLE 5-4

FOUR-FACTOR MULTIVARIATE NORMAL CLASSIFICATION AND DIVERGENCE RESULTS

Sample / Assignment	Cotton	Sorghum	Citrus	Bare Soil
Cotton	247	87	38	6
Sorghum	270	475	19	0
Citrus	76	36	497	12
Bare Soil	7	2	46	582

Divergence	Cotton	Sorghum	Citrus	Bare Soil
Cotton	0.05	2.8	9.0	53
Sorghum	2.8	0.04	18	83
Citrus	9.0	18	0.06	17
Bare Soil	53	83	17	0.06

transformation removes some at the difficulties of specifying a smoothing matrix by calling for a diagonal matrix instead of a more generally symmetric one.

The common transformation space is not one in which data are uncorrelate as illustrated by the training set covariance matrices used in the multivariate normal computations above, Figure 5-11. The non-zero off-diagonal terms of the matrices are evidence of correlation, and make the common transformation space a poor frame of reference in which to represent targets with a smoothing function.

The costly exponentiation involved in the Specht function makes such impractical for further use with real-world data investigations. Further study of smoothing functions in various frames of reference make use of the more practical hyperpyramid function described in the parametric data studies, Section 4.3.

Cotton Covariance

1	2	3	4	5	6	7	8
0.33660	0.17548	0.08295	-0.06952	-0.18687	0.04450	0.03035	-0.12347
0.17548	0.99814	-0.09189	-0.00654	-0.00179	0.09388	-0.23815	-0.03941
0.08295	-0.09189	1.07436	0.04521	-0.11364	-0.10616	0.11070	-0.11211
-0.06952	-0.00654	0.04521	0.98425	0.05528	0.19353	-0.17920	0.12511
-0.18687	-0.00179	-0.11364	0.05528	1.15593	0.06192	-0.02071	0.24622
0.04450	0.09388	-0.10616	0.19853	0.06192	0.55837	0.02671	-0.01576
0.03035	-0.23815	0.11070	-0.17920	-0.02071	0.02671	0.98172	0.05772
-0.12347	-0.03941	-0.11211	0.12511	0.24622	-0.01575	0.05772	0.94531

Sorghum Covariance

1	2	3	4	5	6	7	8
0.14689	0.02527	0.01358	-0.09275	-0.06115	0.08750	0.09401	-0.08565
0.02527	1.22588	0.00613	-0.07564	-0.04515	-0.04477	0.18912	-0.03265
0.01358	0.00613	0.90508	0.19901	0.05925	-0.05114	0.03921	-0.10783
-0.09275	-0.07564	0.19901	0.79834	-0.04349	0.13158	-0.11413	0.06743
-0.06115	-0.04516	0.06926	-0.04349	0.72622	-0.01017	-0.06577	-0.01194
0.08750	-0.04477	-0.05114	0.13158	-0.01017	0.57393	0.10785	-0.04251
0.09401	0.18912	0.03921	-0.11413	-0.06577	0.10785	0.82091	0.05055
-0.08565	-0.03265	-0.10783	0.06743	-0.01194	-0.04251	0.05055	0.65083

Citrus Covariance

1	2	3	4	5	6	7	8
0.28500	0.21031	-0.02697	-0.20638	0.00151	0.14151	-0.10092	0.18451
0.21031	0.44625	-0.02518	-0.15781	0.03508	0.13104	-0.10539	0.24135
-0.02697	-0.02518	1.21897	0.30844	0.18841	-0.15855	0.04014	-0.06810
-0.20638	-0.15781	0.30844	1.24496	-0.13091	0.05172	0.16884	-0.20159
0.00151	0.03508	0.18841	-0.13091	0.92079	0.17780	-0.01150	0.10201
0.14151	0.13104	-0.15855	0.05172	0.17780	0.34512	0.03363	0.17045
-0.10092	-0.10539	0.04014	0.16884	-0.01150	0.03363	0.79242	-0.15746
0.18451	0.24135	-0.06810	-0.20159	0.10201	0.17045	-0.15746	1.27003

Bare Soil Covariance

1	2	3	4	5	6	7	8
0.52929	0.55144	0.08396	0.05340	0.14649	-0.03408	-0.04064	0.03780
0.55144	0.71598	0.07128	0.09749	0.22709	-0.09473	-0.01514	0.06148
0.08396	0.07128	0.70275	-0.07370	-0.09304	0.07845	-0.06689	0.01061
0.05340	0.09749	-0.07370	0.91611	0.07149	0.14448	0.18036	0.05367
0.14649	0.22709	-0.09304	0.07149	1.29839	-0.09577	0.16737	-0.15782
-0.03408	-0.09473	0.07845	0.14448	-0.09577	1.44499	-0.12629	-0.03691
-0.04064	-0.01514	-0.06689	0.18036	0.16737	-0.12629	1.31930	0.20910
0.03780	0.06148	0.01061	0.05367	-0.15782	-0.03691	0.20910	1.00228

8059-62

Figure 5-11 Covariance Matrices of Training Sets

5.2.2 Multiple Transformations

An orthogonal factor analysis, as presented in Section 4.3.2, produces uncorrelated factor scores, eliminating problems of directionality in the selection of smoothing parameters. It is noted there that the factor analysis eliminates the need for a smoothing matrix similar in concepts to the inverse covariance matrix of the multivariate normal probability density function. Smoothing of samples used to represent target populations is therefore performed in factor space for simplicity. This section will examine the utility of the approach of factor analyzing each target sample individually and subsequently transforming the unknown signatures into each of the target spaces. It is shown that dimensionality reduction obtained by an analysis of this type is useful; but a portion of the useful information in multispectral observations is lost by using principal components to reduce dimensionality.

This conclusion motivates the development of a transformation, discussed in Section 5.3, which is performed with specific directions to transform the original variables into a coordinate system in which the new variables are uncorrelated and which also allows dimensionality reduction with less loss of useful information.

The analysis techniques developed during the parametric studies are used to conduct this study of real world data. Much emphasis is placed upon careful choice of sample size for representation of the targets. The concept of information divergence is applied to two samples from one target; when this divergence becomes small it is an indication that the samples are large enough to represent the population. * This divergence can be made arbitrarily small by choosing large enough smoothing parameters. A small divergence obtained in this way is not acceptable. To avoid this result, marginal densities are computed and plotted as a function of one variable. A subjective requirement is imposed that the curve be smooth yet show the major population centers evident in a scatter diagram of the first two variables.

Probability density computation throughout this study is based upon the hyperpyramid smoothing function, which is faster and more practical than the exponential form of the smoothing function used earlier for the common transformation study. The hyperpyramid smoothing function is:

$$f(\vec{X}) = \frac{1}{2^{m-1} N(S_1 S_2 \cdots S_m)} \sum_{j=1}^{N} g(\vec{X}, \vec{Y}_j)$$

*See Section 4.3.1

where $$g(\vec{X}, \vec{Y}_j) = 1 - \max\left\{\left|\frac{X_i - Y_{ij}}{S_i}\right| , i = 1 \; m\right\} \text{ for}$$

$$\max\left\{\left|\frac{X_i - Y_{ij}}{S_i}\right| , i = 1, m\right\} \leq 1, \text{ and}$$

$$g(\vec{X}, \vec{Y}_j) = 0 \text{ for } \max\left\{\left|\frac{X_i - Y_{ij}}{S_i}\right| , i = 1, m\right\} > 1.*$$

The smoothing parameters, S_i, have the dimensions of the standard deviation and must be chosen after considerations of the number of points in the sample, the shape and number of modes in the distribution, and in general, the direction along which smoothing should take place.

The density divergence computation** is used to determine the divergence between two representations of the same population, and can be useful in choosing smoothing parameters:

$$D = \frac{1}{N} \sum_{j=1}^{N} \left(1 - \frac{f_2}{f_1}\right) \log \frac{f_1}{f_2}$$

where f_1 if one representation of the probability density and f_2 is a second.

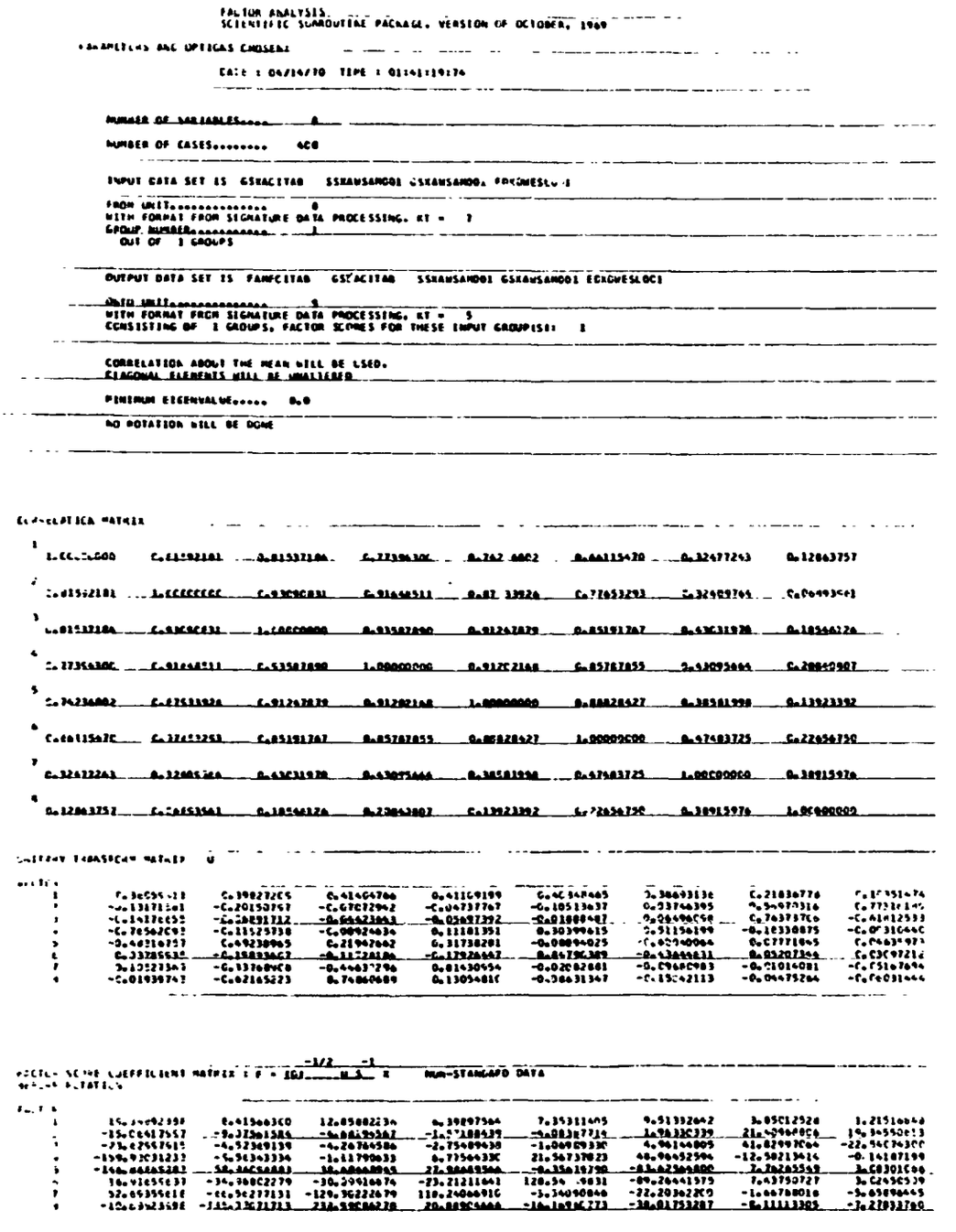
5.2.2.1 Initial Investigation

To test initially the hypothesis that a target is easily distinguishable from its background (the nontarget signatures) when all the signatures are represented in that target's factor space, specialized scatter diagrams are produced as follows. Four hundred signatures from each of the four target crops (cotton, sorghum, citrus, and bare soil) are factor analyzed separately.

Principal components are used exclusively throughout this study. Rotations are never performed on transformations obtained by diagonalizing the correlation matrix. Computer printouts showing a summary of the analysis performed and the final result for each of the categories are shown in Figures 5-12, 5-13, 5-14, and 5-15. The four resultant factor score coefficient matrixes are then used to

* See Section 4.3.

** See Section 4.3.1.



FACTOR ANALYSIS
SCIENTIFIC SUBROUTINE PACKAGE, VERSION OF OCTOBER, 1969

DATE = 06/14/70 TIME = 01:41:19.74

NUMBER OF VARIABLES..... 8

NUMBER OF CASES..... 400

INPUT DATA SET IS GSXACITAB SSXAWSAMDDI GSXAWSAMDDA FORMEST(0,1)

FROM UNIT..... 8
WITH FORMAT FROM SIGNATURE DATA PROCESSING. RT = 7
GROUP NUMBER..... 1
OUT OF 1 GROUPS

OUTPUT DATA SET IS FAMFCITAB GSXACITAB SSXAWSAMDDI GSXAWSAMDDI FCXGWESLOC1

UNIT UNIT..... 9
WITH FORMAT FROM SIGNATURE DATA PROCESSING. RT = 5
CONSISTING OF 1 GROUPS, FACTOR SCORES FOR THESE INPUT GROUP(S): 1

CORRELATION ABOUT THE MEAN WILL BE USED.
DIAGONAL ELEMENTS WILL BE UNALTERED

MINIMUM EIGENVALUE..... 0.0

NO ROTATION WILL BE DONE

CORRELATION MATRIX

1	1.000000000	0.811902181	0.815337184	0.773396306	0.742 66002	0.666115470	0.32477243	0.12863757
2	0.811902181	1.000000000	0.930900831	0.916668511	0.87 33926	0.776653293	0.329099764	0.090493401
3	0.815337184	0.930900831	1.000000000	0.935807890	0.912678879	0.851917867	0.436131930	0.185466126
4	0.773396306	0.916668511	0.935807890	1.000000000	0.912721868	0.857877855	0.430095644	0.208409907
5	0.742366002	0.873533926	0.912678879	0.912721868	1.000000000	0.888228527	0.385819980	0.139233922
6	0.666115470	0.776653293	0.851917867	0.857877855	0.888228527	1.000000000	0.474833725	0.226567301
7	0.324772431	0.329099764	0.436131930	0.430975444	0.385819980	0.474833725	1.000000000	0.389159760
8	0.128637572	0.090493401	0.185466126	0.208439807	0.139233922	0.226567301	0.389159760	1.000000000

INVERSE CORRELATION MATRIX

1	7.3e0595128	0.398272265	0.414464706	0.411691991	0.4C3648445	3.386931136	0.218366776	0.1[illegible]
2	-0.139711161	-0.201503767	-0.67672942	-0.04737767	-0.105133637	0.937446395	0.9[illegible]	0.7731[illegible]
3	-0.164176252	-0.382917112	-0.666421843	-0.059973392	-0.021888487	0.026498C58	0.763787C6	-0.41[illegible]
4	-0.765602C91	-0.115253738	-0.08924434	0.11181351	0.303398615	2.911561199	-0.103308875	-0.0[illegible]
5	-0.469316757	0.469389065	0.21947602	0.317382281	-0.008945025	-0.6[illegible]	0.677771845	0.[illegible]
6	0.33285532[illegible]	-0.158836C2	-0.312[illegible]	-0.[illegible]	0.847[illegible]	-0.43[illegible]	0.05207344	0.C3C97214
7	0.1352233[illegible]	-0.33760[illegible]	-0.44463[illegible]	0.81430954	-0.02C82881	-0.C9[illegible]	-0.21014081	-0.[illegible]
8	-0.01939747	-0.02165223	0.74860689	0.13054811	-0.38631347	-0.15[illegible]	-0.09475264	-0.[illegible]

FACTOR SCORE COEFFICIENT MATRIX [F = [G] [M] R] NON-STANDARD DATA

1	15.[illegible]02358	0.415e03C0	12.05882230	0.39897564	7.353111695	9.513926e2	3.05C12528	1.2151e044
2	-15.C[illegible]7557	-9.[illegible]584	-[illegible]	-1.[illegible]	-[illegible]	1.[illegible]	21.[illegible]	1[illegible]
3	-73.[illegible]615	-4.52369159	-4.2676458[illegible]	-2.75409[illegible]	-1.0[illegible]	4.961[illegible]	41.[illegible]	-22.[illegible]
4	-139.[illegible]1233	-[illegible]	-1.[illegible]	0.[illegible]	21.[illegible]	48.[illegible]	-12.[illegible]	-0.[illegible]
5	-[illegible]	[illegible]	[illegible]	[illegible]	-0.[illegible]	-83.[illegible]	7.[illegible]	3.[illegible]
6	36.[illegible]31	-34.[illegible]	-30.[illegible]	-23.[illegible]	128.[illegible]	-89.[illegible]	7.[illegible]	3.[illegible]
7	52.[illegible]	-[illegible]	-129.[illegible]	110.[illegible]	-3.[illegible]	-22.[illegible]	-1.[illegible]	-5.[illegible]
8	-[illegible]	-[illegible]	[illegible]	[illegible]	-[illegible]	-38.[illegible]	-[illegible]	-[illegible]

8069-32

Figure 5-12 Factor Analysis of Citrus

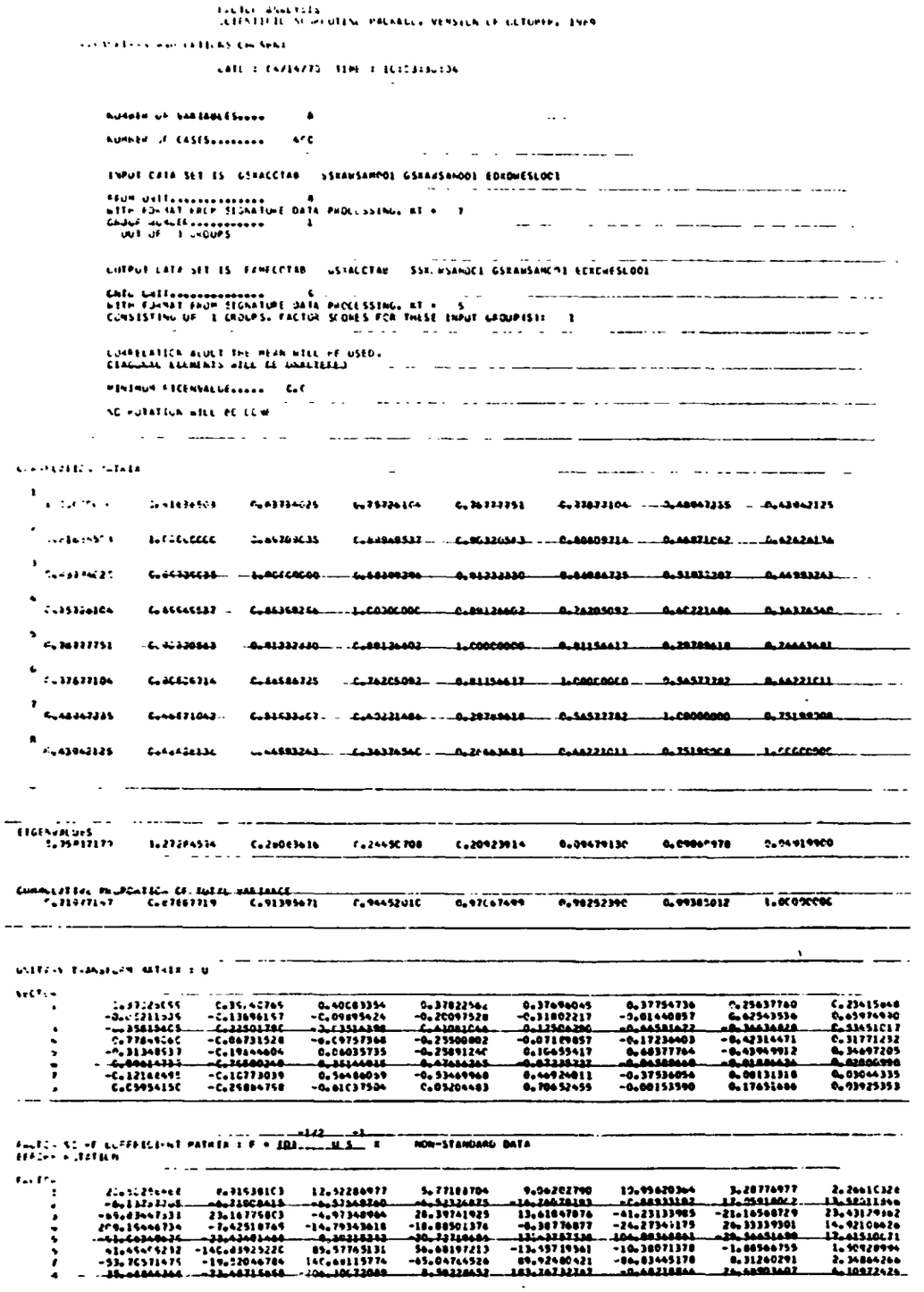
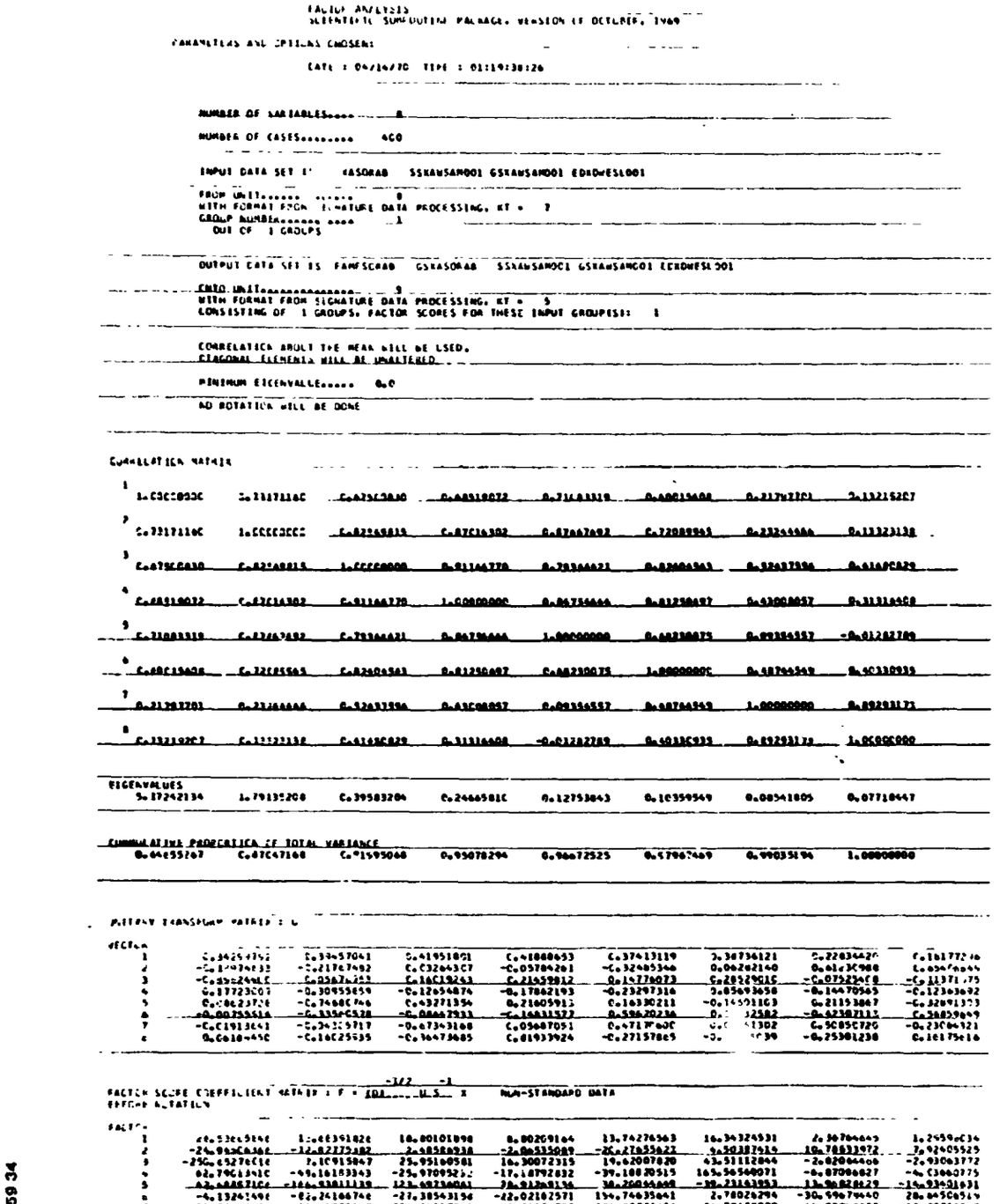


Figure 5-13 Factor Analysis of Cotton



```
                    FACTOR ANALYSIS
                    SCIENTIFIC SUBROUTINE PACKAGE, VERSION OF OCTOBER, 1969

PARAMETERS AND OPTIONS CHOSEN:

                    DATE : 04/16/70  TIME : 01:19:38:26

      NUMBER OF VARIABLES.....    8

      NUMBER OF CASES.........    400

      INPUT DATA SET IS    GASORAB   SSKANSAMOO1 GSKANSAMOO1 EDKDWESL001

      FROM UNIT.....  .....     8
      WITH FORMAT FROM SIGNATURE DATA PROCESSING, RT =   7
      GROUP NUMBER..... ....    1
        OUT OF   1 GROUPS

      OUTPUT DATA SET IS  FANFSCRAB   GSKASORAB   SSKANSAMOO1 GSKANSAMOO1 EDKDWESL001

      CARD UNIT.....  .....     9
      WITH FORMAT FROM SIGNATURE DATA PROCESSING, RT =   5
      CONSISTING OF   1 GROUPS. FACTOR SCORES FOR THESE INPUT GROUPS IS:   1

      CORRELATION ABOUT THE MEAN WILL BE USED.
      DIAGONAL ELEMENTS WILL BE UNALTERED.

      MINIMUM EIGENVALUE.....   0.0

      NO ROTATION WILL BE DONE

CORRELATION MATRIX

  1
    1.000000000   0.331171810   0.825650830   0.685918072   0.714183319   0.680154408   0.217492701   0.132152677

  2
    0.331171810   1.000000000   0.825650835   0.870716302   0.878687692   0.720889565   0.232644686   0.133323138

  3
    0.825650830   0.825650835   1.000000000   0.811446770   0.793644621   0.836684563   0.324837596   0.616490829

  4
    0.685918072   0.870716302   0.811446770   1.000000000   0.867356644   0.812758697   0.430068057   0.313184458

  5
    0.714183319   0.878687692   0.793644621   0.867356644   1.000000000   0.882380875   0.493564557  -0.012682789

  6
    0.680154408   0.720889565   0.836684563   0.812758697   0.882380875   1.000000000   0.487965549   0.450339359

  7
    0.217492701   0.232644686   0.324837596   0.430068057   0.493564557   0.487965549   1.000000000   0.892931723

  8
    0.132152677   0.133323138   0.616490829   0.313184458  -0.012682789   0.450339359   0.892931723   1.000000000

EIGENVALUES
    5.372421340   1.791352080   0.395832040   0.246658160   0.127538430   0.103595490   0.085418050   0.077184470

CUMULATIVE PROPORTION OF TOTAL VARIANCE
    0.671552675   0.895470850   0.945259500   0.976082500   0.991996250   0.995945625   0.999618125   1.000000000

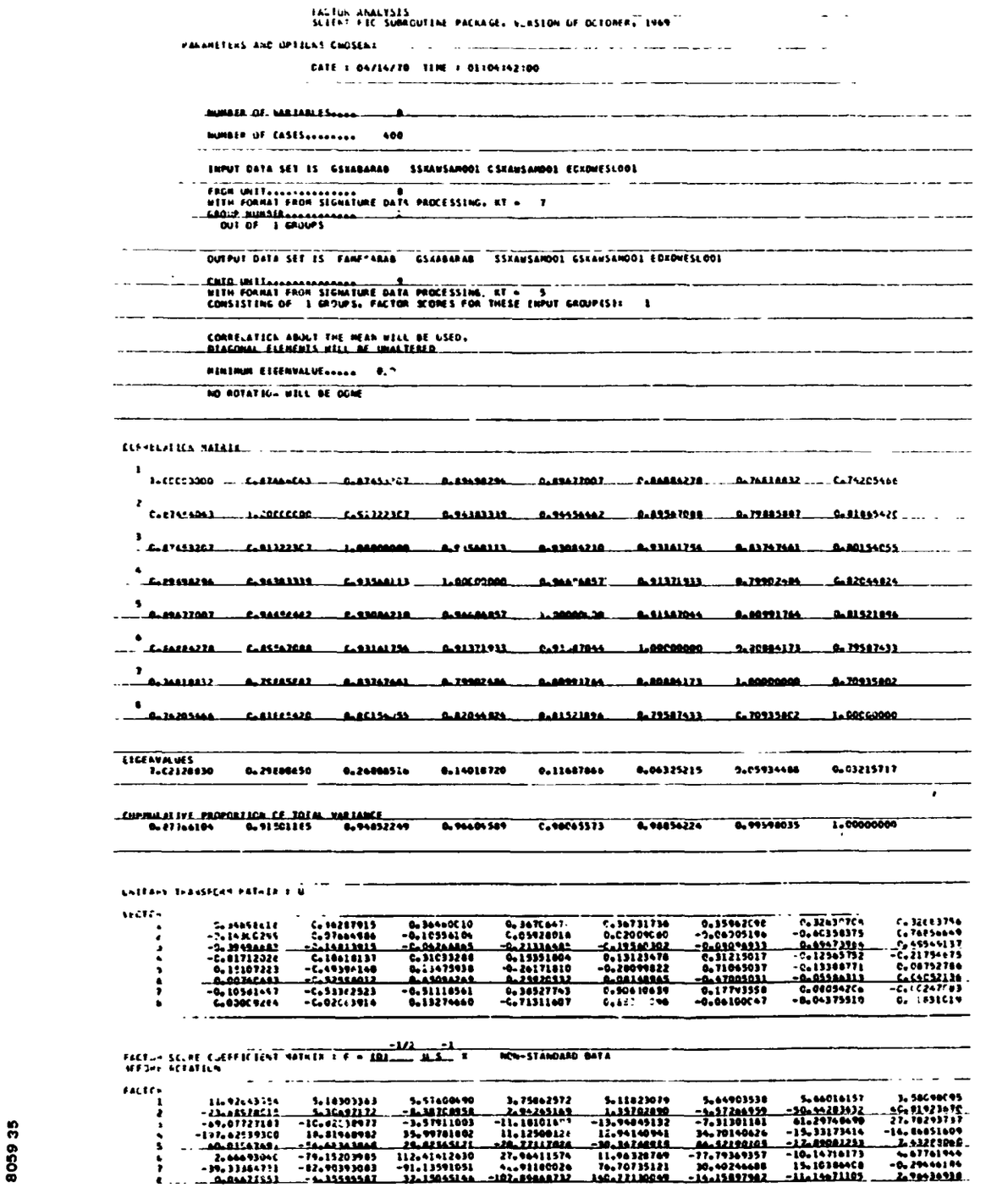
MATRIX TRANSFORMATION MATRIX :  0

VECTOR
     1        0.342547352   0.339457041   0.419518921   0.410880653   0.374131190   0.367361210   0.220344200   0.181772740
     2       -0.217674832  -0.217674832   0.032643307  -0.057842610  -0.324853460   0.066282140   0.612305980   0.656760440
     3       -0.455524440   0.455524440   0.056474220   0.164139263   0.215539812   0.167784073   0.285249815  -0.075234080  -0.113711750
     4        0.337233007  -0.309555859  -0.126548740  -0.178821930  -0.232977310   0.856993650  -0.144705650  -0.123036920
     5        0.000862372  -0.746881740   0.043271350   0.216059130   0.163302110  -0.145931160   0.211933840  -0.328913230
     6       -0.000755510  -0.339455320   0.088479310  -0.168115270   0.586202300   0.032582000  -0.925871150   0.568590490
     7       -0.619138410  -0.342159170  -0.047363180   0.056487050  -0.471190020   0.041302000   0.905857200  -0.230644210
     8        0.660804450  -0.166235350  -0.364738650   0.819339240  -0.271578060  -0.032582000  -0.253812380   0.181756160

FACTOR SCORE COEFFICIENT MATRIX : F = [D1]  -1/2  [R]  -1     NON-STANDARD DATA
EQUATION

FACTOR
     1      -49.532605840   1.000391820   18.801010980   0.802091640   13.742276560   16.343249530   0.367044440   1.259900230
     2      -24.988504360  -12.827793480   2.485884310  -2.004535080  -25.278559220   4.503187410   18.781331970   7.924055290
     3      -25.652766160   1.801915860   25.951605810   16.300723150   19.620078200   63.511128040  -2.020044000  -2.930631770
     4       62.790618340  -69.161833040  -25.970955210  -17.187927830  -39.188705150   165.565660070  -4.087006827  -4.013602770
     5       42.448671160  -14.538111130   123.487360040   78.913481540   38.200445690  -38.231613950   13.508484290  -14.933431630
     6       -4.132451490  -82.241867460  -27.385431960  -22.021025710   134.744835640   2.780262790  -30.596794400   28.845065140
     7      -11.525923710  -11.619531670  -234.053117540   9.299267150   134.832238540   26.794997700   41.029948822  -12.807322490
     8      33.144588240  -85.857858360  -133.811178570   140.995131590  -81.003572230  -18.228239850  -21.478986370   5.445084870
```

Figure 5-14 Factor Analysis of Sorghum



FACTOR ANALYSIS
SCIENTIFIC SUBROUTINE PACKAGE, VERSION OF OCTOBER, 1969

PARAMETERS AND OPTIONS CHOSEN:

DATE = 04/14/70 TIME = 01:04:42:00

NUMBER OF VARIABLES..... 8

NUMBER OF CASES..... 400

INPUT DATA SET IS GSXABARAB SSXAWSAM001 GSXAWSAM001 EDXDWESL001

FROM UNIT..... 8
WITH FORMAT FROM SIGNATURE DATA PROCESSING. RT = 7
GROUP NUMBER..... 1
OUT OF 1 GROUPS

OUTPUT DATA SET IS FAMF*ARAB GSXABARAB SSXAWSAM001 GSXAWSAM001 EDXDWESL001

CNTD UNIT..... 9
WITH FORMAT FROM SIGNATURE DATA PROCESSING. RT = 5
CONSISTING OF 1 GROUPS. FACTOR SCORES FOR THESE INPUT GROUP(S): 1

CORRELATION ABOUT THE MEAN WILL BE USED.
DIAGONAL ELEMENTS WILL BE UNALTERED

MINIMUM EIGENVALUE..... 0.7

NO ROTATION WILL BE DONE

CORRELATION MATRIX

1	1.00000000	0.87466643	0.87653267	0.89498296	0.89627007	0.88886278	0.76818832	0.74220566
2	0.87466643	1.00000000	0.95122367	0.94383319	0.94556662	0.89587088	0.79885887	0.81895425
3	0.87653267	0.95122367	1.00000000	0.95568113	0.93084218	0.93141756	0.83797661	0.80154655
4	0.89498296	0.94383319	0.95568113	1.00000000	0.96678857	0.91371933	0.79902496	0.82044824
5	0.89627007	0.94556662	0.93084218	0.96678857	1.00000000	0.91587044	0.80991766	0.81521896
6	0.88886278	0.89587088	0.93141756	0.91371933	0.91587044	1.00000000	0.80886173	0.79587433
7	0.76818832	0.79885887	0.83797661	0.79902496	0.80991766	0.80886173	1.00000000	0.70935802
8	0.74220566	0.81895425	0.80154655	0.82044824	0.81521896	0.79587433	0.70935802	1.00000000

EIGENVALUES

3.62128030	0.29808650	0.26088516	0.14018720	0.11687866	0.06325215	0.05934486	0.03215717

CUMULATIVE PROPORTION OF TOTAL VARIANCE

0.27766104	0.91501165	0.94852249	0.96604589	0.98065573	0.98856224	0.99598035	1.00000000

UNITARY TRANSFORMATION MATRIX :

VECTOR

1	0.34858618	0.36287915	0.36440C10	0.36706647	0.36731730	0.35962C0E	0.32637*7C9	0.32CE3756
2	-0.14JLC295	0.97664986	-0.1C556104	0.05928018	0.C2009C00	-0.C6705196	-0.0C358375	0.76856669
3	-0.39498887	-0.14813015	-0.06248885	-0.21338688	-0.19580302	-0.0190*8933	0.89673285	0.45541137
4	-0.81712026	0.18618137	0.31C33268	0.19391004	0.13125478	0.31215017	-0.12965792	-0.21794675
5	0.11107223	-0.49398148	0.13475938	-0.26171810	-0.20099822	0.71865037	-0.13368771	0.08752786
6	0.00346483	-0.23456612	0.45004420	0.24026532	0.08148845	-0.47805031	-0.05186111	0.64C52136
7	-0.10581447	-0.53362523	-0.51118561	0.36527703	0.50610838	0.177*3558	0.0005426	-0.CCC247*03
8	0.03CC9264	-0.02CC13014	0.13274660	-0.71311087	0.[illegible]	-0.06100C47	-0.04375519	0.18316219

FACTOR SCORE COEFFICIENT MATRIX : F = [R]^-1/2 [A] S = NON-STANDARD DATA
BEFORE ROTATION

FACTOR

1	11.92643554	9.18303363	9.57600490	3.75862572	5.18823079	5.64903538	5.66016157	3.58C90C95
2	-23.85780215	5.3CA97172	-8.38JC80958	2.94265169	1.35702890	-4.57266959	-50.44283632	-6.81923675
3	-69.07727183	-1C.02C38927	-3.59911003	-11.1010107*3	-13.99405132	-7.31301181	61.29746490	27.78249737
4	-197.62539300	10.81468982	35.99781002	11.12500126	12.94140941	34.70140626	-15.33173414	-14.86851609
5	80.81543A0	-5[illegible]	28.[illegible]	20.77117826	-30.36746818	44.52190265	-17.89081253	2.43259060
6	2.66693904	-70.15207985	112.41612630	27.06411574	11.96328789	-77.79369357	-10.14716173	4.67761944
7	-90.31866731	-82.90393083	-91.13591051	4[illegible]	76.70735121	30.60246488	15.1638640	-0.29946184
8	0.04427551	-4.1[illegible]	32.15045164	-107.898A8732	1[illegible]	-14.15897582	-11.1[illegible]	2.98926938

8059 35

Figure 5-15 Factor Analysis of Bare Soil

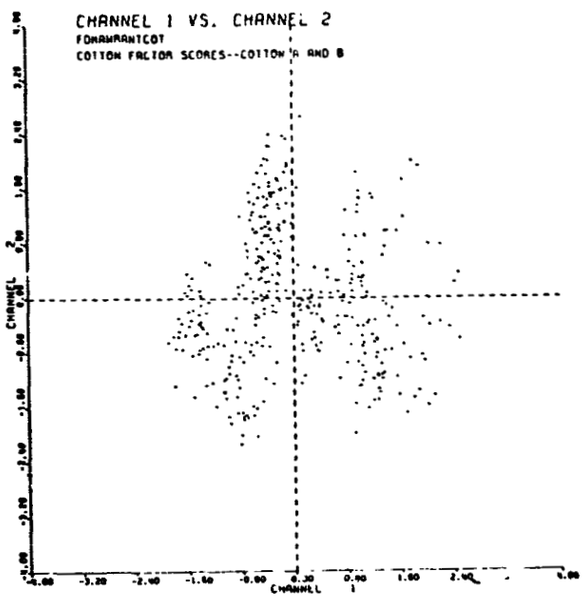
transform a data set composed of 1600 observations, 400 of each crop. The scatter diagrams representing all four pairs of cotton factor scores and non-cotton in cotton space, sorghum factor scores and non-sorghum in sorghum space, citrus factor scores and non-citrus in citrus space, and bare soil factor scores and non-bare soil in bare soil space, are seen in Figure 5-16. That each target crop clusters in a different region than its background under such transformations is most highly evident in the diagrams of citrus factor 1 vs. factor 2 and citrus background factor 1 vs. factor 2. In all cases, the background factor 1 vs. factor 2 and factor 3 vs. factor 4 scatters are skewed away from the corresponding target factor score plots and have a larger variance.

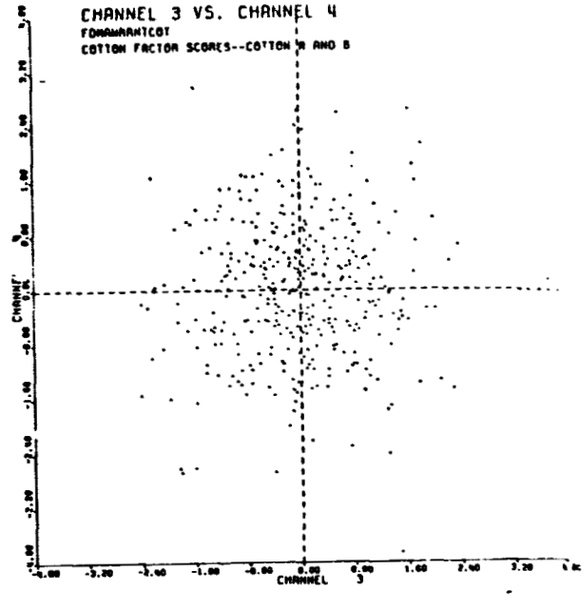
In many cases, it is also evident that the higher numbered factors, which account only for small amounts of the variance in each of the categories, still contain information useful for distinguishing the categories. In the scatter plots for factors 7 and 8 of bare soil and bare soil background, this is particularly clear. A linear combination of raw data variables, as any factor score is formed, that subtends only a small fraction of total variance in the sample, is a direction along which the raw data hardly ever change. Such a direction is of potentially greater interest than factors 1 and 2 which describe most of the variance for each sample. Factor analytical transformations are therefore less useful than they might be for reduction of dimensionality. Since the transformation is useful for other reasons, and since the use of all eight factor scores in all the probing studies is not feasible, the multiple transformation work of this section is conducted in large part using only the first four principal components, interchangeably referred to as factor scores, of each category.

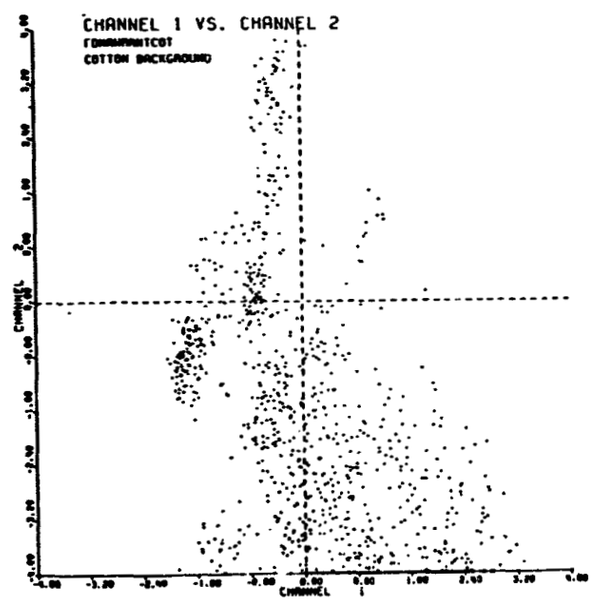
The importance of the information lost is estimated, but the basic deficiency is corrected by the approach described in Section 5.3.

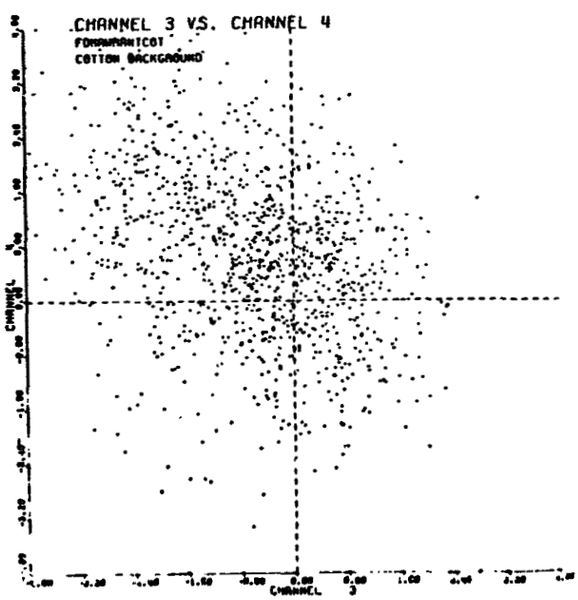
5.2.2.2 Sample Size and Smoothing

The task of first choosing appropriate training set sizes to ensure representation of the target populations is begun by collecting all of the ground truth signatures of the four crops and factor analyzing separately. From the sets of resultant factor scores, pairs of samples of sizes 200, 400, and 600 observations are randomly selected for each crop. Scatter diagrams displaying 2000 factor scores of each, Figure 5-17, are the basis for initial estimates of applicable smoothing parameters to be used in the hyperpyramid smoothing function for two-sample divergence. Analysis is limited to the first four factors out of eight While useful information is concentrated in these variables, it is clear that significant information loss is suffered by doing so.


CHANNEL 1 VS. CHANNEL 2
FONAWRANTCOT
COTTON FACTOR SCORES--COTTON A AND B
CHANNEL 2
CHANNEL 1


CHANNEL 3 VS. CHANNEL 4
FONAWRANTCOT
COTTON FACTOR SCORES--COTTON A AND B
CHANNEL 4
CHANNEL 3


CHANNEL 1 VS. CHANNEL 2
FONAWRANTCOT
COTTON BACKGROUND
CHANNEL 2
CHANNEL 1


CHANNEL 3 VS. CHANNEL 4
FONAWRANTCOT
COTTON BACKGROUND
CHANNEL 4
CHANNEL 3

8059-29

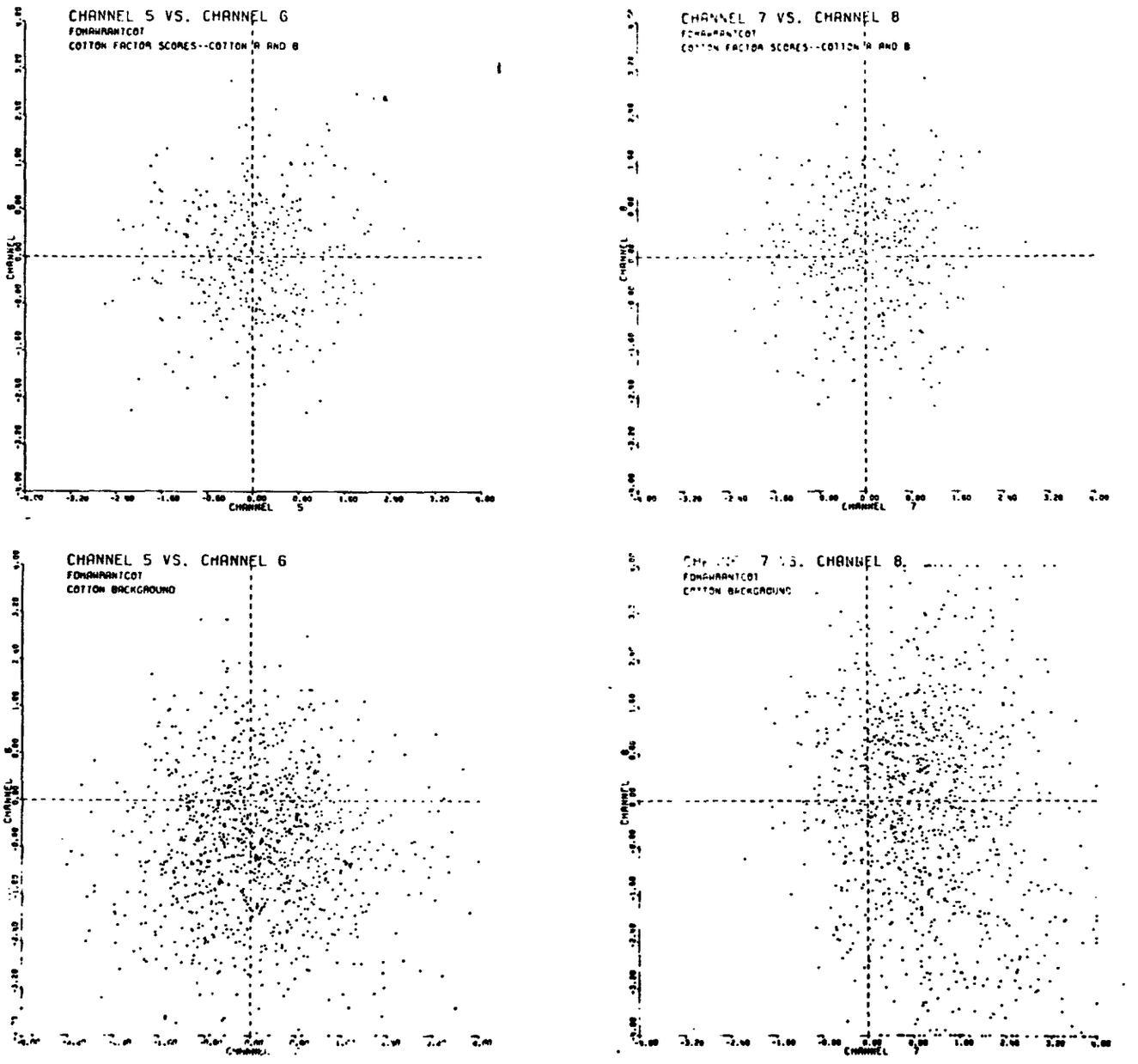
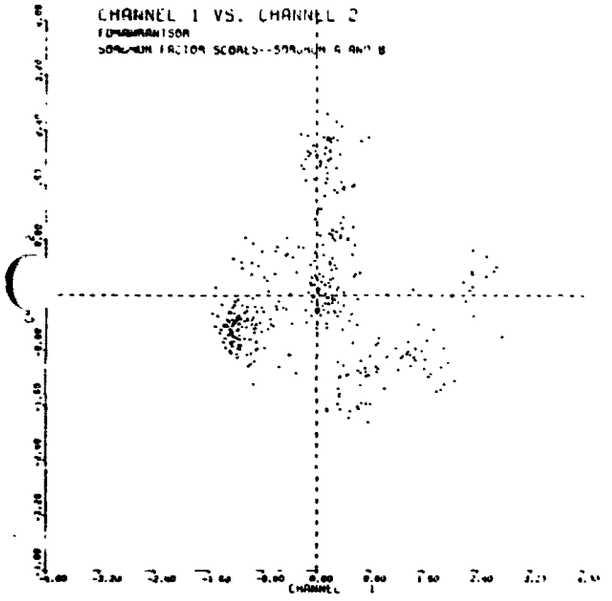
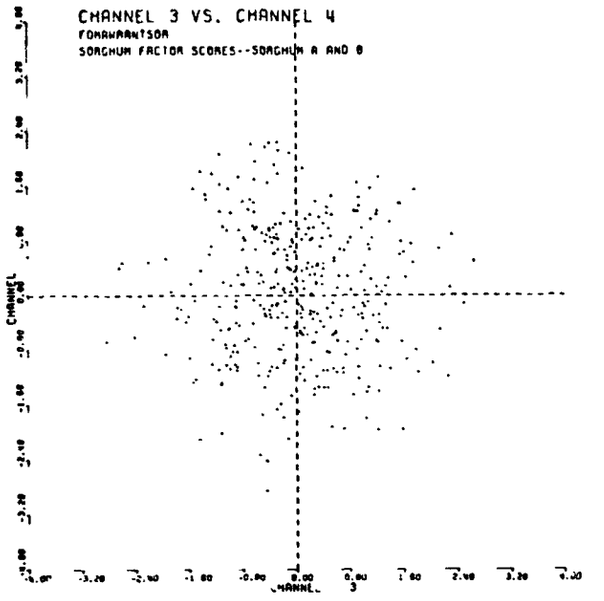
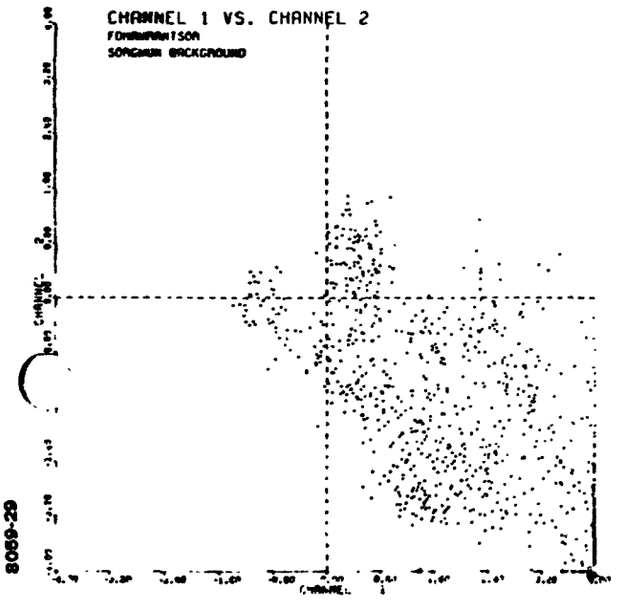
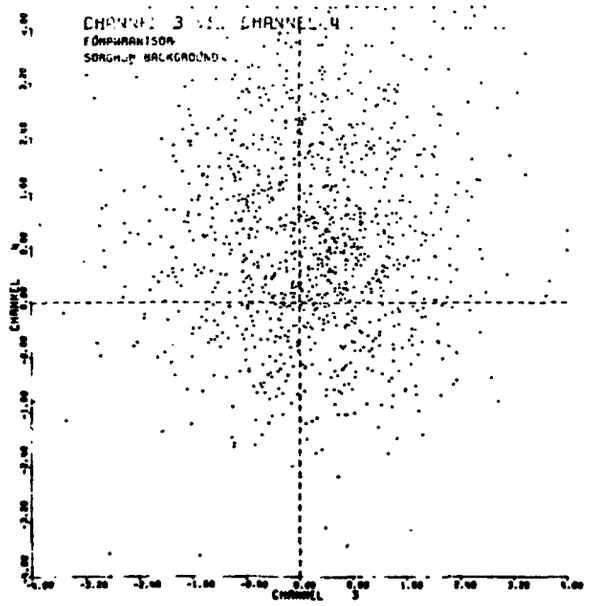


Figure 5-16 Scatter Diagrams of Crops and their Backgrounds


CHANNEL 1 VS. CHANNEL 2
FONAWANTSOR
SORGHUM FACTOR SCORES--SORGHUM A AND B


CHANNEL 3 VS. CHANNEL 4
FONAWANTSOR
SORGHUM FACTOR SCORES--SORGHUM A AND B


CHANNEL 1 VS. CHANNEL 2
FONAWANTSOR
SORGHUM BACKGROUND


CHANNEL 3 VS. CHANNEL 4
FONAWANTSOR
SORGHUM BACKGROUND

8059-29

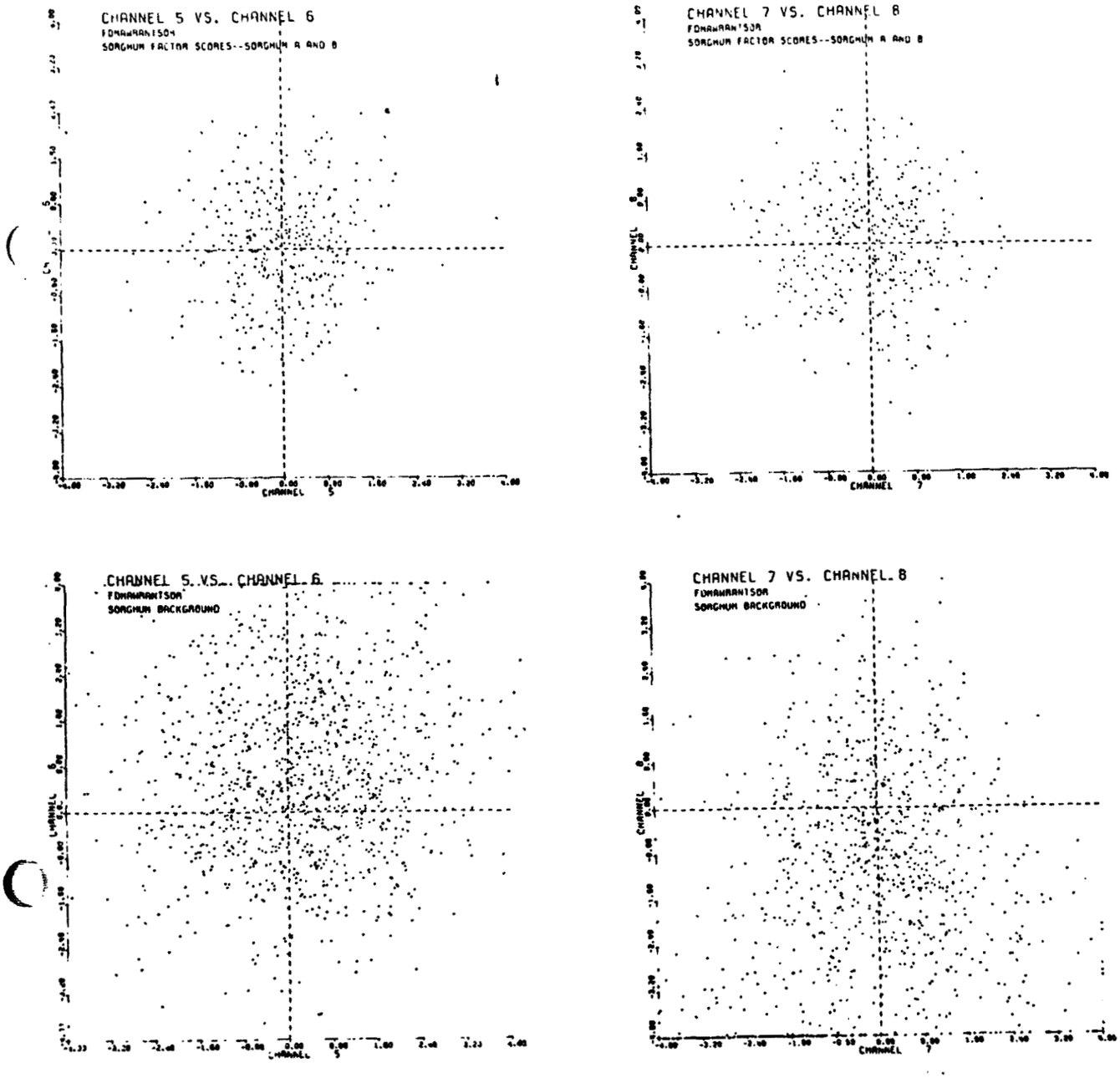
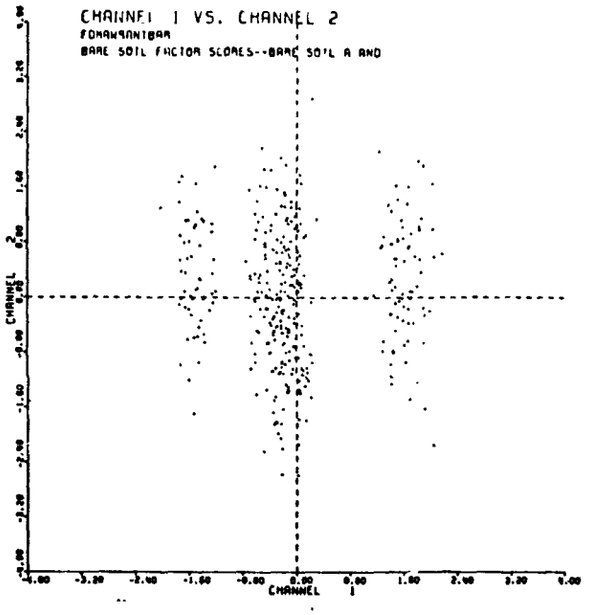
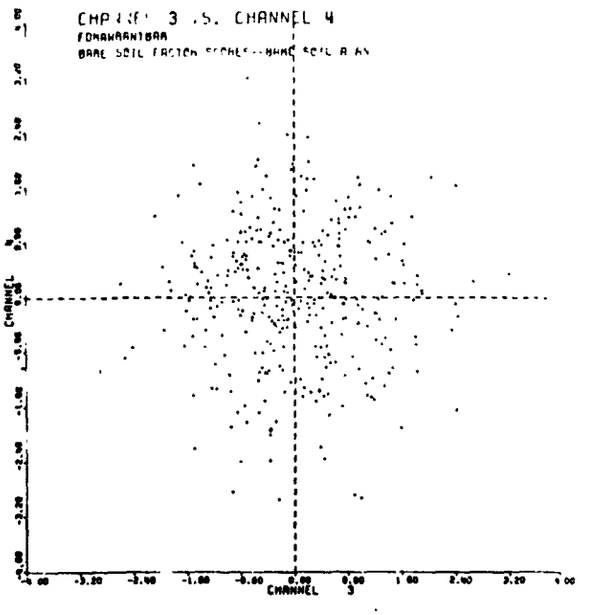
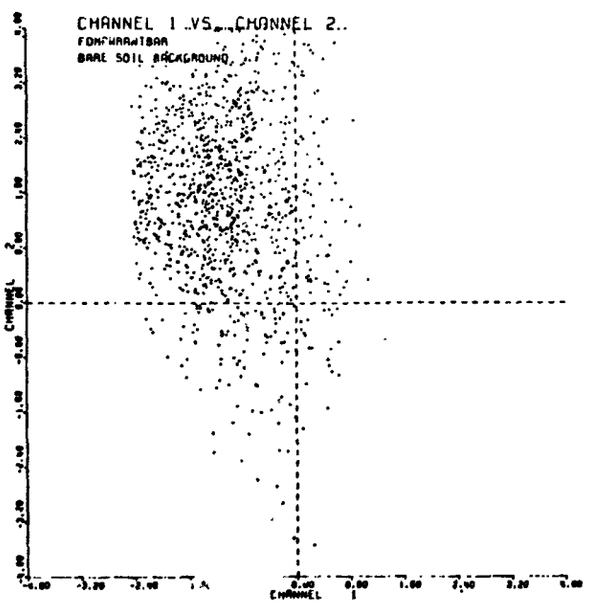
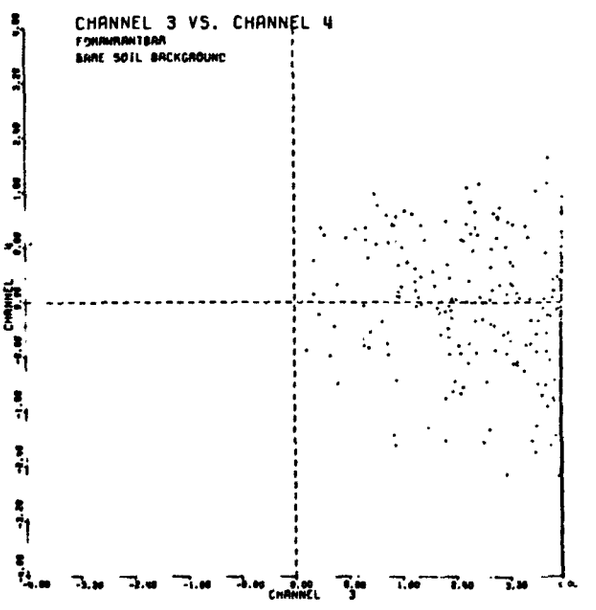


Figure 5-16 Scatter Diagrams of Crops and their Backgrounds


CHANNEL 1 VS. CHANNEL 2
FORAWRANTBAR
BARE SOIL FACTOR SCORES--BARE SOIL A AND
CHANNEL 2
CHANNEL 1


CHANNEL 3 VS. CHANNEL 4
FORAWRANTBAR
CHANNEL 3
CHANNEL 3


CHANNEL 1 VS. CHANNEL 2
FORAWRANTBAR
BARE SOIL BACKGROUND
CHANNEL 2
CHANNEL 1


CHANNEL 3 VS. CHANNEL 4
FORAWRANTBAR
BARE SOIL BACKGROUND
CHANNEL 4
CHANNEL 3

8059-29

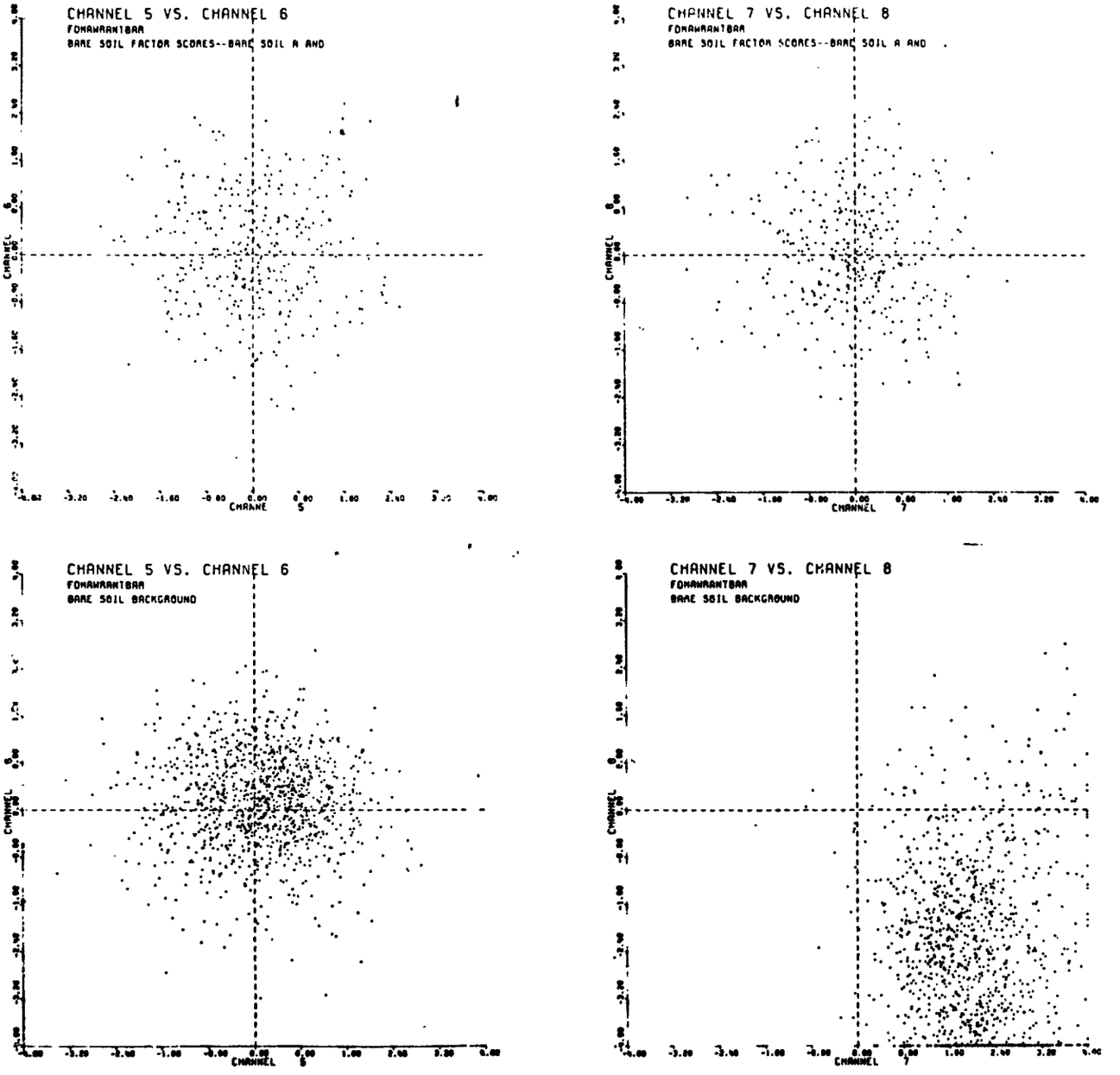
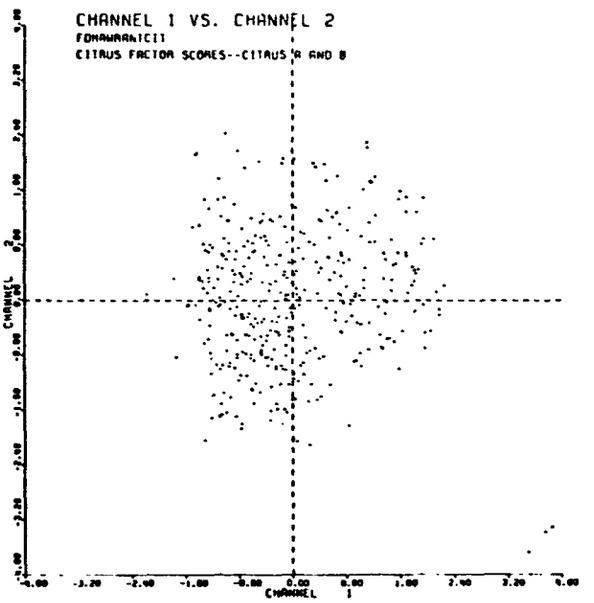
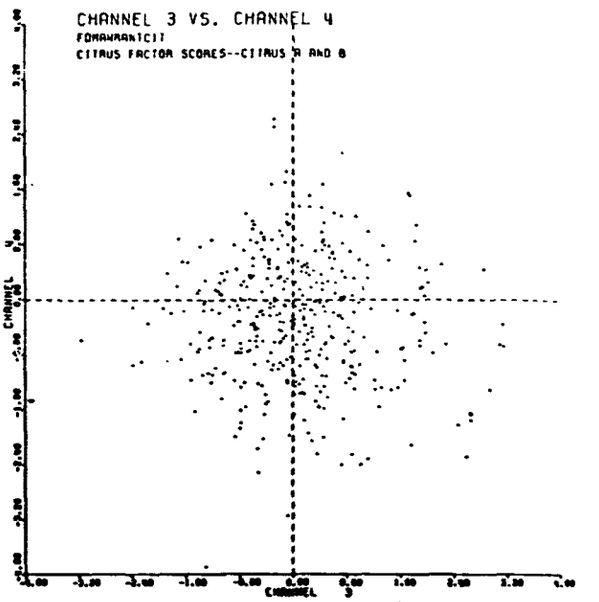
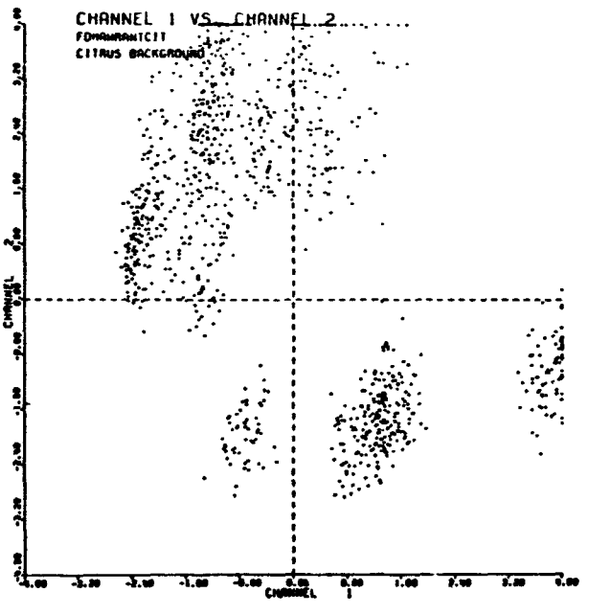
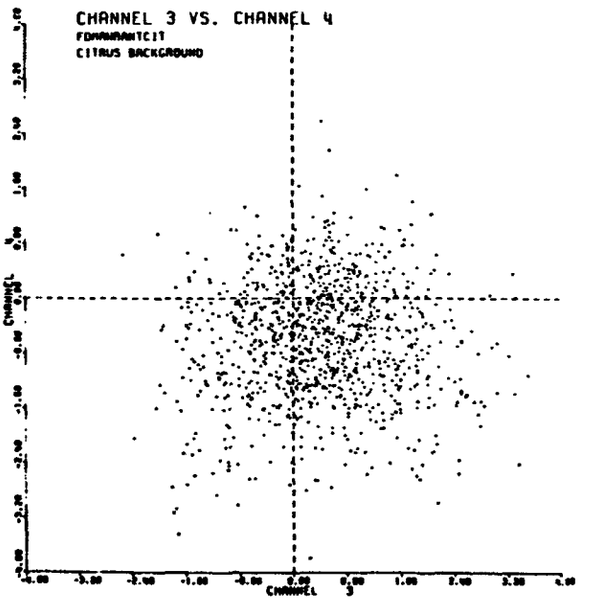


Figure 5-16 Scatter Diagrams of Crops and their Backgrounds


CHANNEL 1 VS. CHANNEL 2
FOMANRANTCIT
CITRUS FACTOR SCORES--CITRUS A AND B
CHANNEL 2
CHANNEL 1


CHANNEL 3 VS. CHANNEL 4
FOMANRANTCIT
CITRUS FACTOR SCORES--CITRUS A AND B
CHANNEL 4
CHANNEL 3


CHANNEL 1 VS. CHANNEL 2
FOMANRANTCIT
CITRUS BACKGROUND
CHANNEL 2
CHANNEL 1


CHANNEL 3 VS. CHANNEL 4
FOMANRANTCIT
CITRUS BACKGROUND
CHANNEL 4
CHANNEL 3

8059-29

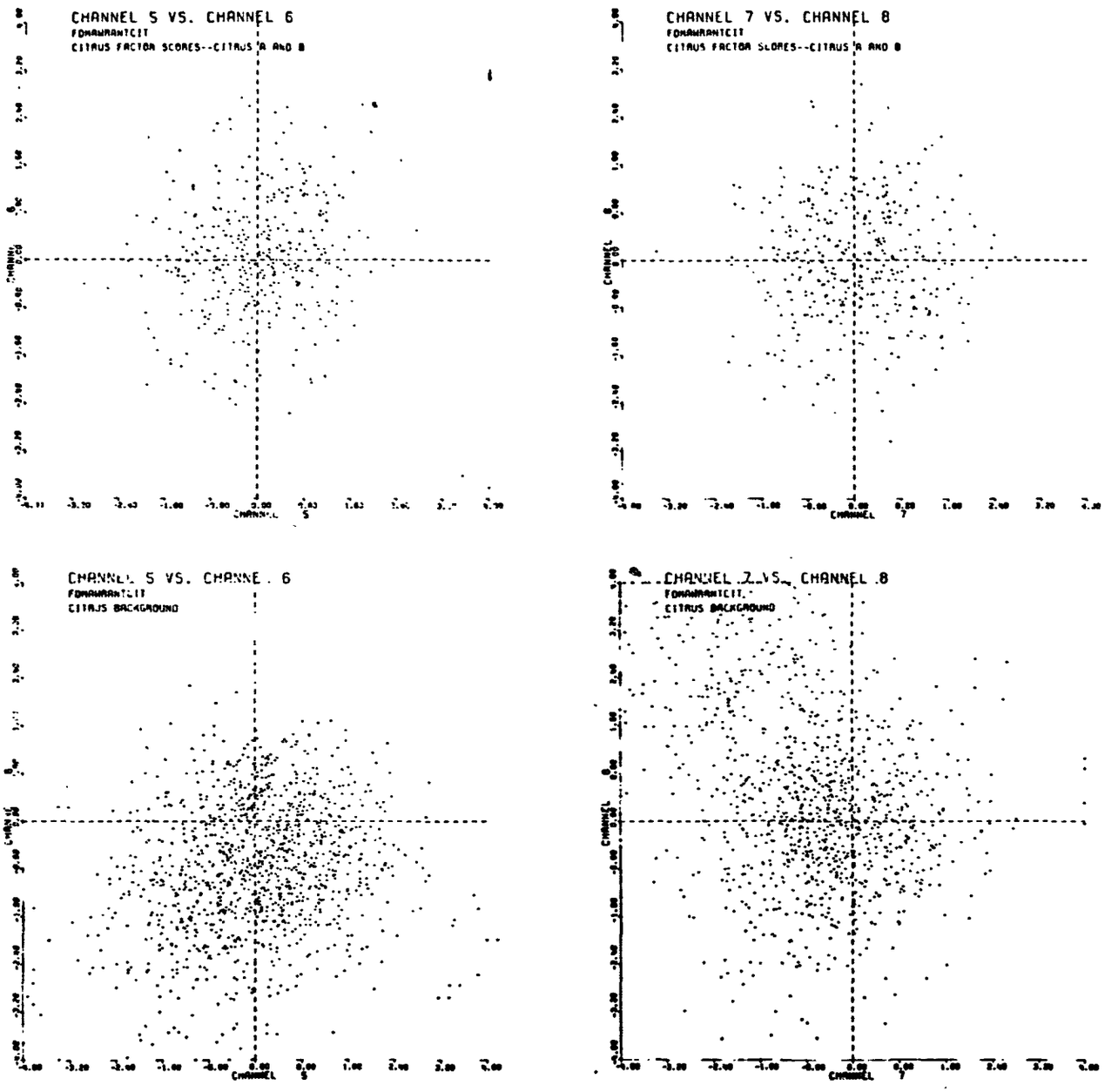


Figure 5-16 Scatter Diagrams of Crops and their Backgrounds

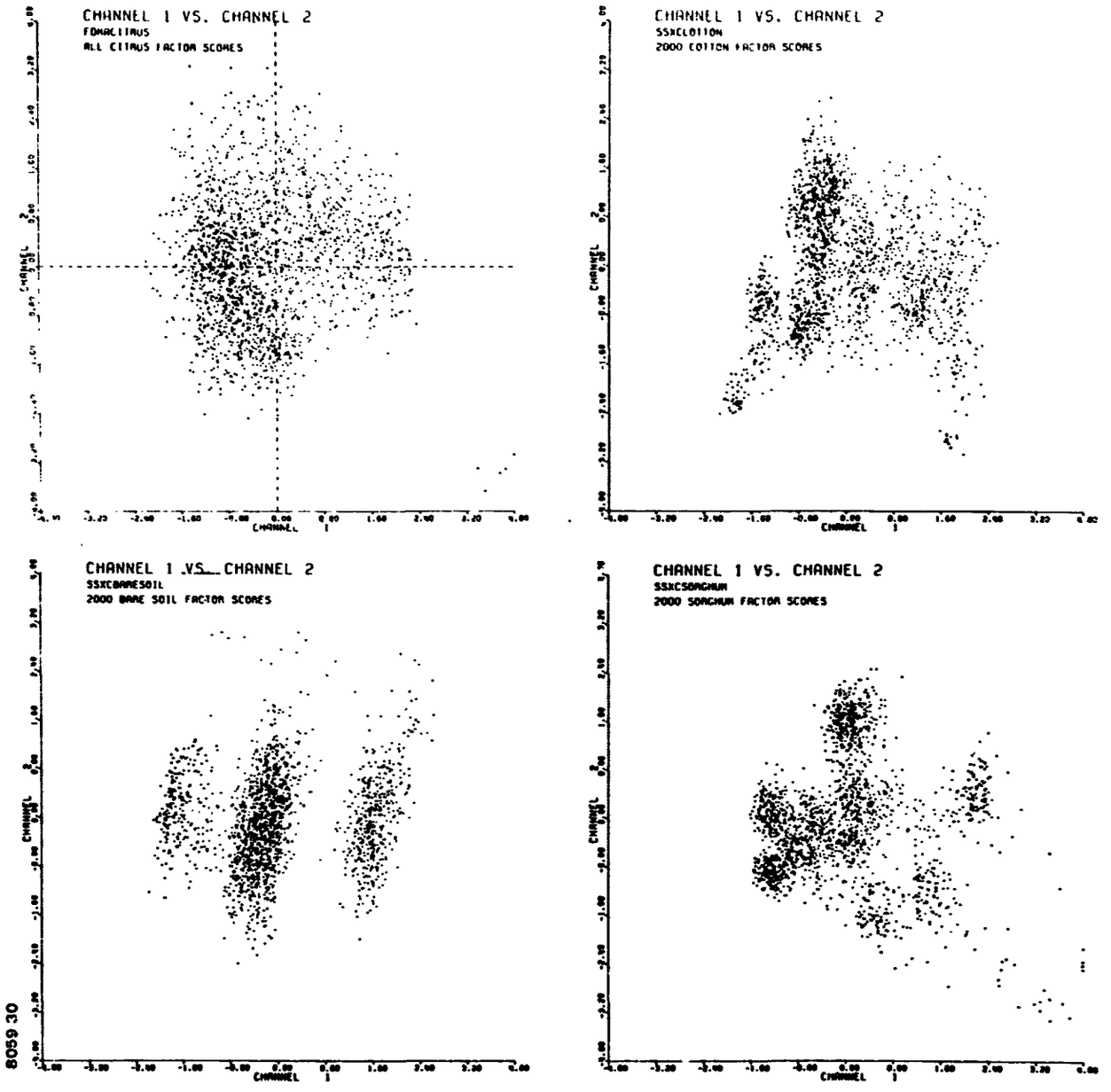


Figure 5-17 Scatter Diagrams 2000 Factor Scores of Four Crops 1 vs. 2

Figure 5-18 presents the results of two-sample divergence on the four crops with use of the visually chosen smoothing parameters. As expected on the basis of parametric study results, a minimum of 600 observations is required to ensure that the divergence between two samples from one target is in the 1-bit range. With smoothing of $\vec{S}$ = (1. , 1. , 1. , 1.) for citrus, $\vec{S}$ = (0.5, 0.5, 1. , 1.) for sorghum and cotton, and $\vec{S}$ = (0.5, 1. , 1. , 1.) for bare soil, citrus is represented by 600 observations with 0.72-bit divergence between two samples of that size, while cotton samples of that size are most divergent at 3.03 bits. An investigation of effects of variations in smoothing parameter choice is necessary before any further judgment on the presented values and sample sizes can be made.

To deal with the problem of smoothing parameter choice, marginal density plots are produced. Probability densities are evaluated for signatures projected into the factor 1 - factor 2 plane along the line factor 2=0, using 600 citrus factor scores for the training set. These densities are then plotted and compared with various choices of pyramid smoothing parameters. These densities should increase and decrease according to the densities represented along the horizontal axis of scatter diagrams in Figure 5-17. Two such density plots are shown in Figure 5-19 and Figure 5-20 for citrus. In comparing the smoothing parameters $\vec{S}$ = (1. , 1. , 1. , 1.) and $\vec{S}$ = (0.75, 0.75, 1. , 1.) applied to the citrus training set, both resemble the interpretation from the scatter diagram. The former smoothing when used in citrus two-sample divergence of 600 observations produced 0.72 bit between the two; the latter choice produced a divergence of 1.47 bits. The larger parameters, giving the smallest divergence, were selected for use in further immediate work.

The selection of parameter in any case is a matter of judgment. The judgment used to interpret the significance of divergence on the order of one bit between two samples from the same population is based partly on the study in Section 4. The density divergence between a parametrically describable population density and a sample from it was studied there. Divergence between the population and the sample of about one bit was obtained under conditions where the sample seemed to do an adequate job of describing the population. There is of course much that could be done to tighten the loose lines of reasoning connecting decisions. In the interest of proceeding far enough to examine the questions of real interest, however, questions of this kind were often given only enough attention to proceed without fear of gross error.

Two-sample divergences on various sizes of samples of sorghum with various smoothing parameters are summarized in Table 5-5 and Figure 5-21. The divergence of 7.9 bits between samples of 600 observations with $\vec{S}$ = (.25, .25, 1. , 1.), for example, is indicative of too little smoothing or too small a sample size.

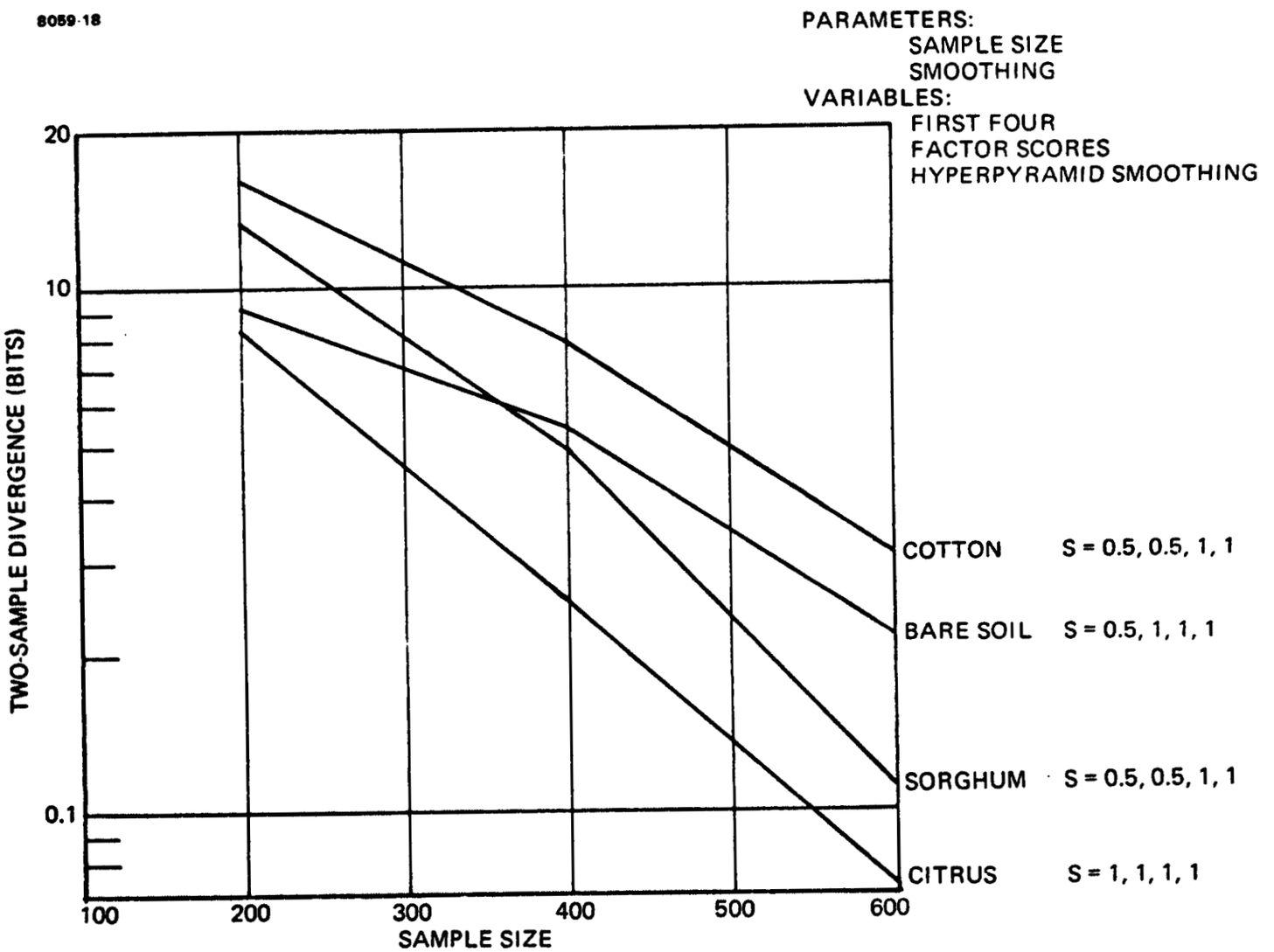


Figure 5-18 Sample Divergence of Similar Categories

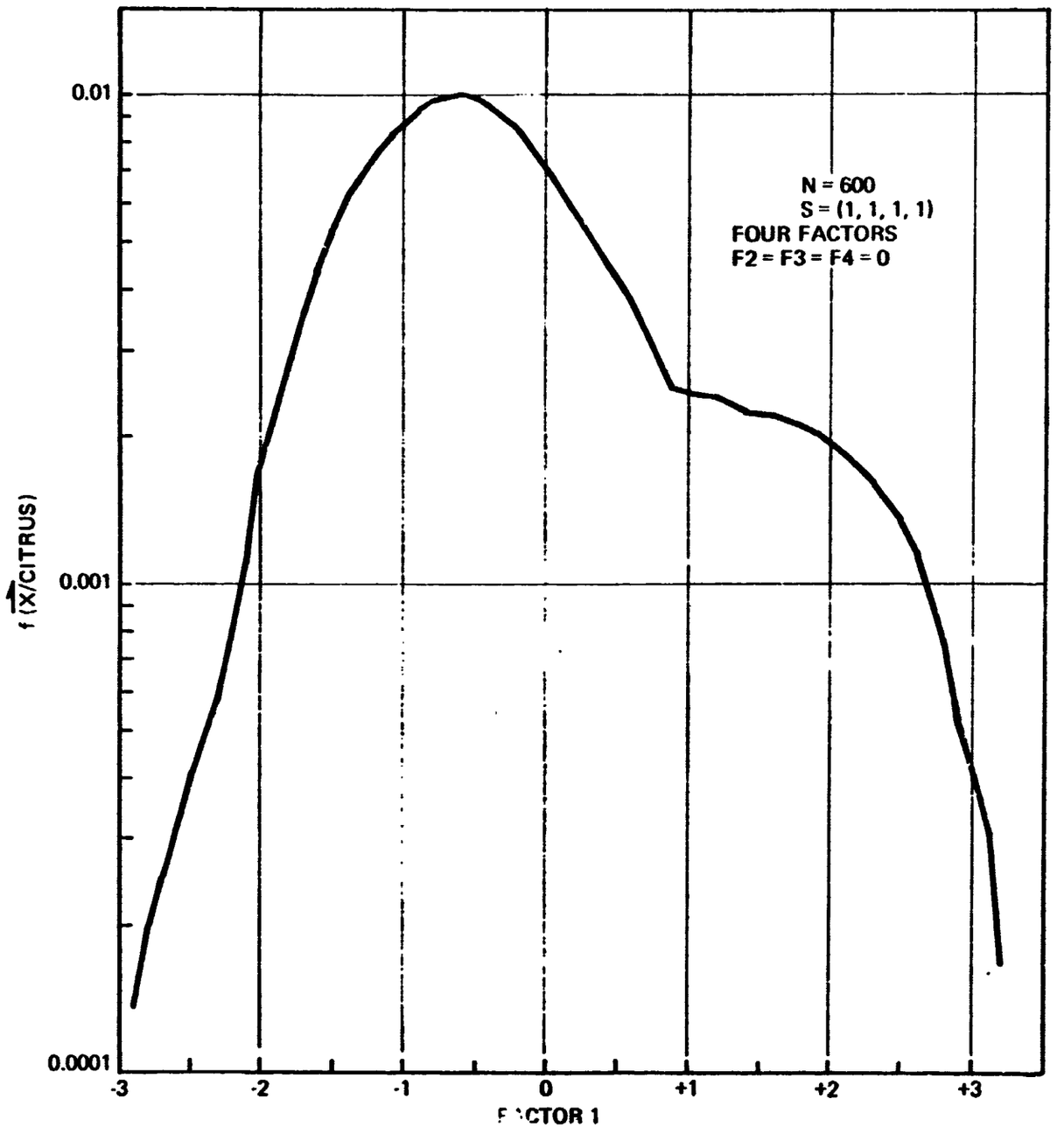
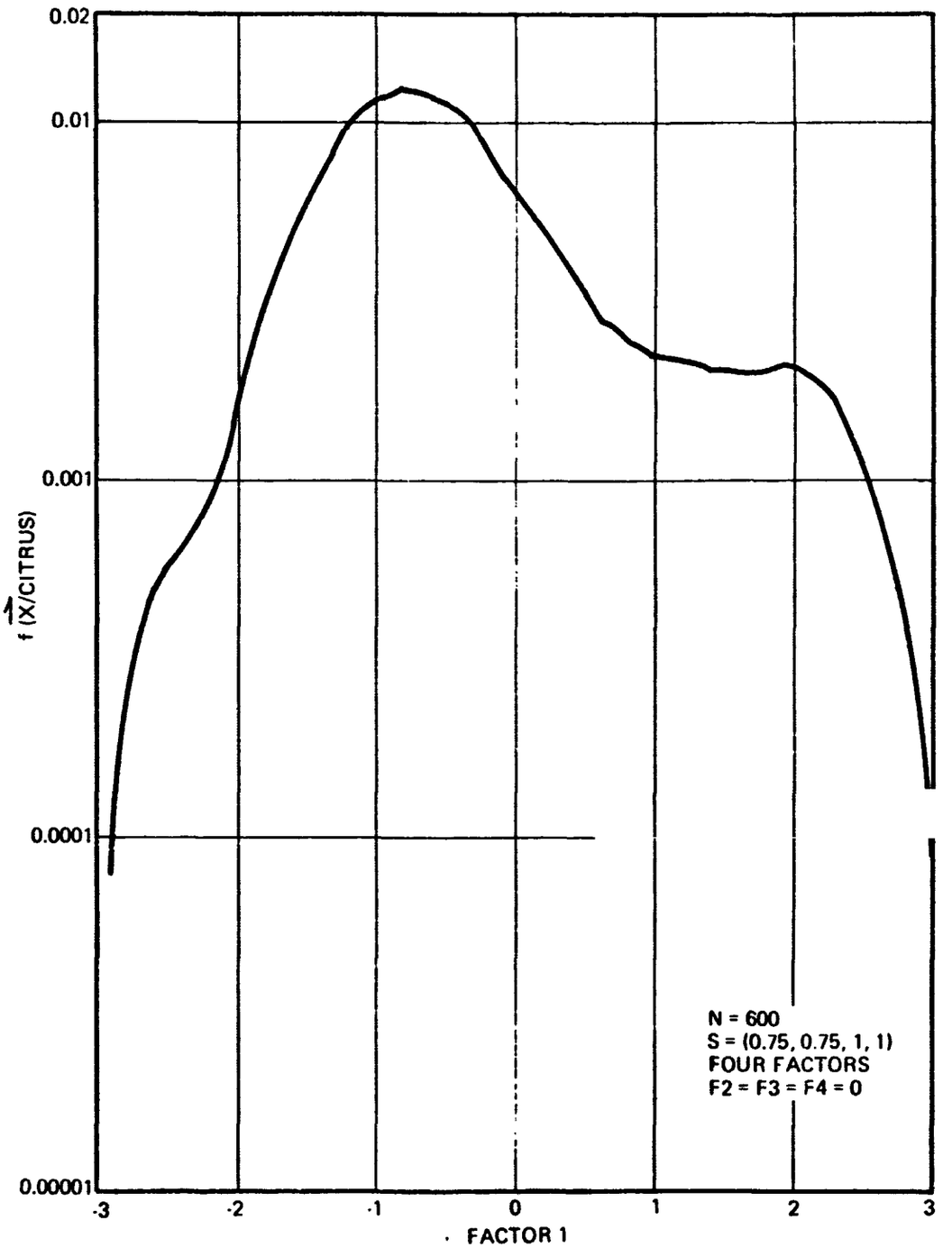


Figure 5-19 Marginal Density for Citrus



8059-20

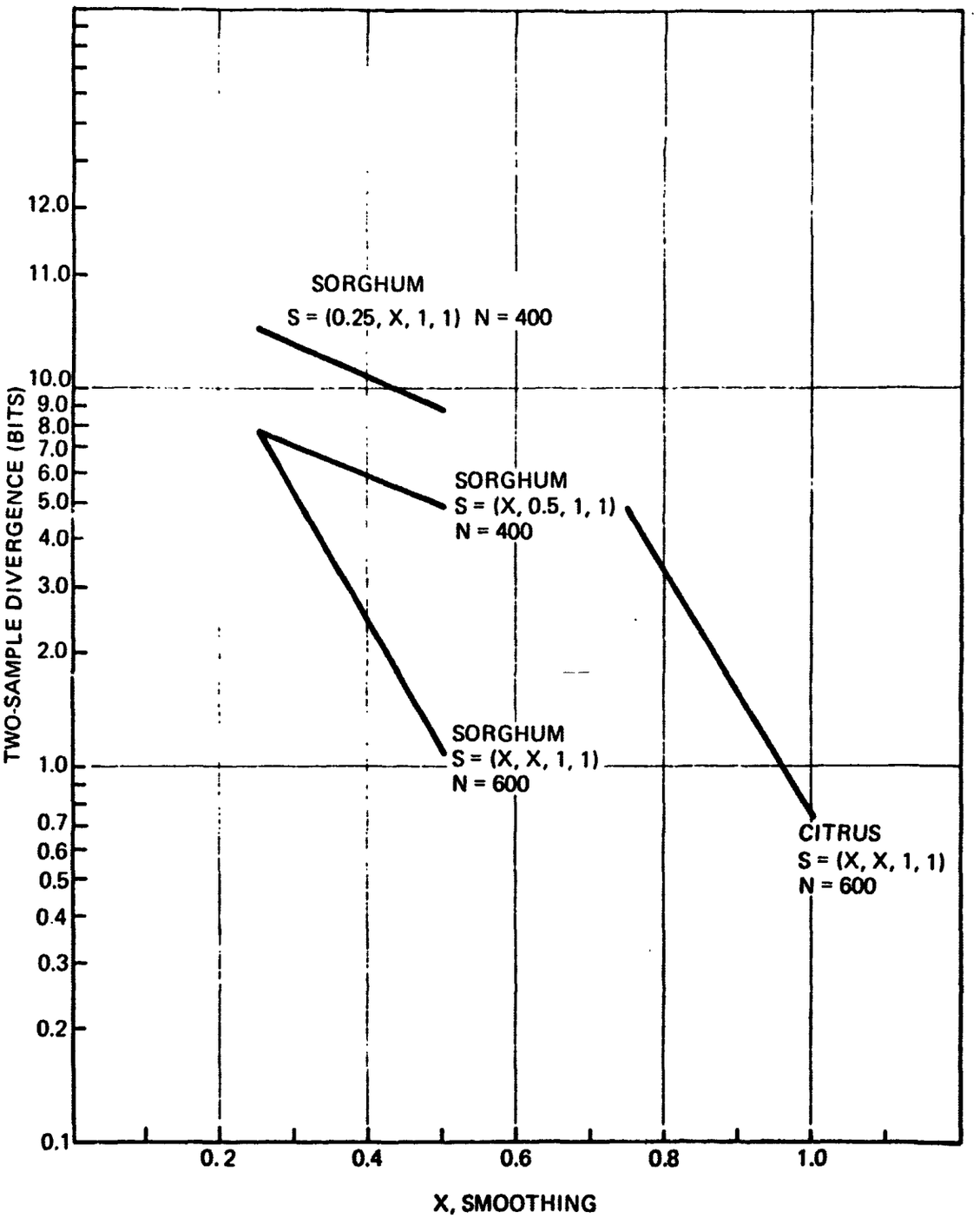
Figure 5-20 Marginal Density for Citrus

TABLE 5-5

TWO-SAMPLE DIVERGENCE FOR SORGHUM

Sample Size	SMOOTHING PARAMETERS EXPRESSED AS VECTOR $\vec{S} = (S_1, S_2, S_3, S_4)$			
	$\vec{S}$ =(0.25, 0.25, 1., 1.)	$\vec{S}$=(0.25, 0.5, 1., 1.)	$\vec{S}$ =(0.5, 0.5, 1., 1.)	$\vec{S}$ =(1., 1., 1., 1.)
200			13.4 Bits	
400	14.4 Bits	7.8 Bits	4.9 Bits	2.3 Bits
600	7.9 Bits		1.1 Bits	

BSR 2949



8059-31

Figure 5-21 Summary of Two-Sample Divergence Variations With Smoothing Parameters

Consider the two-sample divergences and marginal density plots of sorghum for $\vec{S}$ = (0.5, 0.5, 1., 1.) and $\vec{S}$ = (0.25, 0.5, 1., 1.). As with citrus, the marginal density plots are created using 600 factor scores of sorghum for the training set. Note that in the region factor 1 = -1.7 through factor 1 = +1.0 of Figure 5-22, the smoothing parameters $\vec{S}$ = (0.25, 0.5, 1., 1.) represent the sorghum densities smoothly. Yet, in the region factor 1 = +1.0 through factor 1 = +2.7, fluctuations in the probability density are occurring due to randomness in the sample rather than the true nature of the data. In comparison, $\vec{S}$ = (0.5, 0.5, 1., 1.) over-smoothes in the first region and represents the second region nicely, Figure 5-23. This would indicate that variable smoothing is needed for sorghum. However, in light of the small divergence of 1.10 bits between two 600-observation sorghum samples with $\vec{S}$ = (0.5, 0.5, 1., 1.) this set of smoothing parameters is chosen for further immediate work.

Bare soil and cotton marginal density plots, Figures 5-24 and 5-25, with $\vec{S}$ = (0.5, 1., 1., 1.) and $\vec{S}$ = (0.5, 0.5, 1., 1.), respectively, are pleasingly smooth representations. Even though the two-sample divergence for 600 bare soil observations is 2.12 bits with such smoothing and 3.04 bits for 600 cotton observations with the above smoothing (Figure 5-18), both of these combinations are regarded as suitable for use in some initial studies.

The selection of smoothing parameters is clearly a subjective task. The parameters used in the following work can be defended by noting that the probability densities computed from the samples using the parameters selected appear to describe the intended populations. The objective measures of information divergence and classification results are applied to these densities in the following sections.

The general problem of selecting smoothing parameters, however, is by no means solved. The use of factor score spaces in which to represent the data reduces the magnitude of the task by permitting the smoothing parameters to coincide in direction with the axes of the coordinate system. The simplification is enough to allow exploration of the effect of choosing the parameters in a fairly systematic, if not completely satisfactory, way. Fortunately, it is not too difficult to select reasonably good parameters by visual inspection of scatter diagrams. Objective methods permitting computer selection of parameters, is a far more desirable goal. An automatic procedure designed to select the parameters should also account for the necessity of changing the smoothing parameters as the local point density in the sample changes. Figures 5-22 and 5-23, showing the marginal densities for sorghum with two different selections of smoothing parameters by projecting points into the plane of factor one and factor two, and allowing factor one to take on all values while factor two equals zero, illustrate the requirement for variable smoothing. Compare these plots with the scatter for sorghum in Figure 5-17.

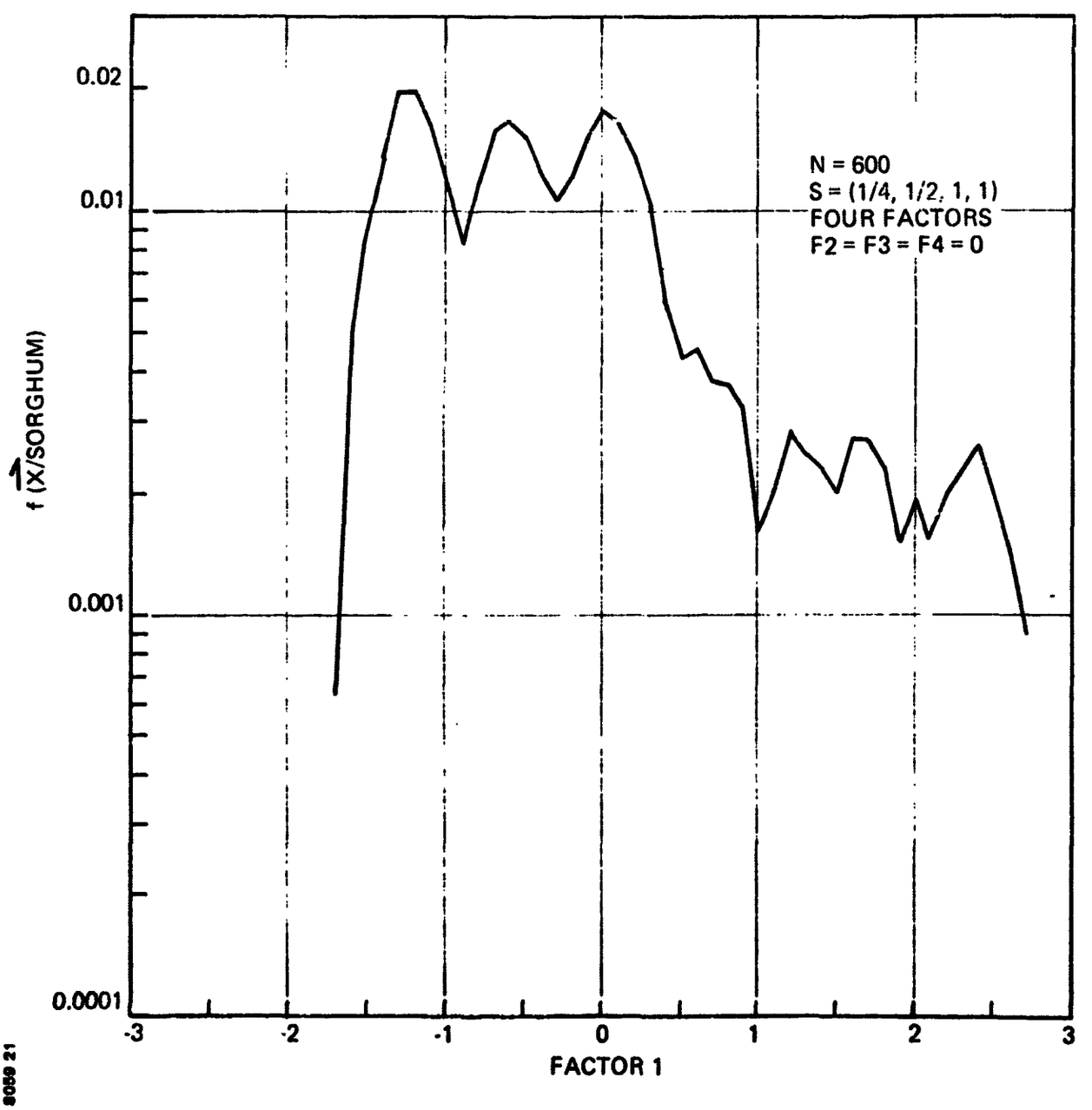
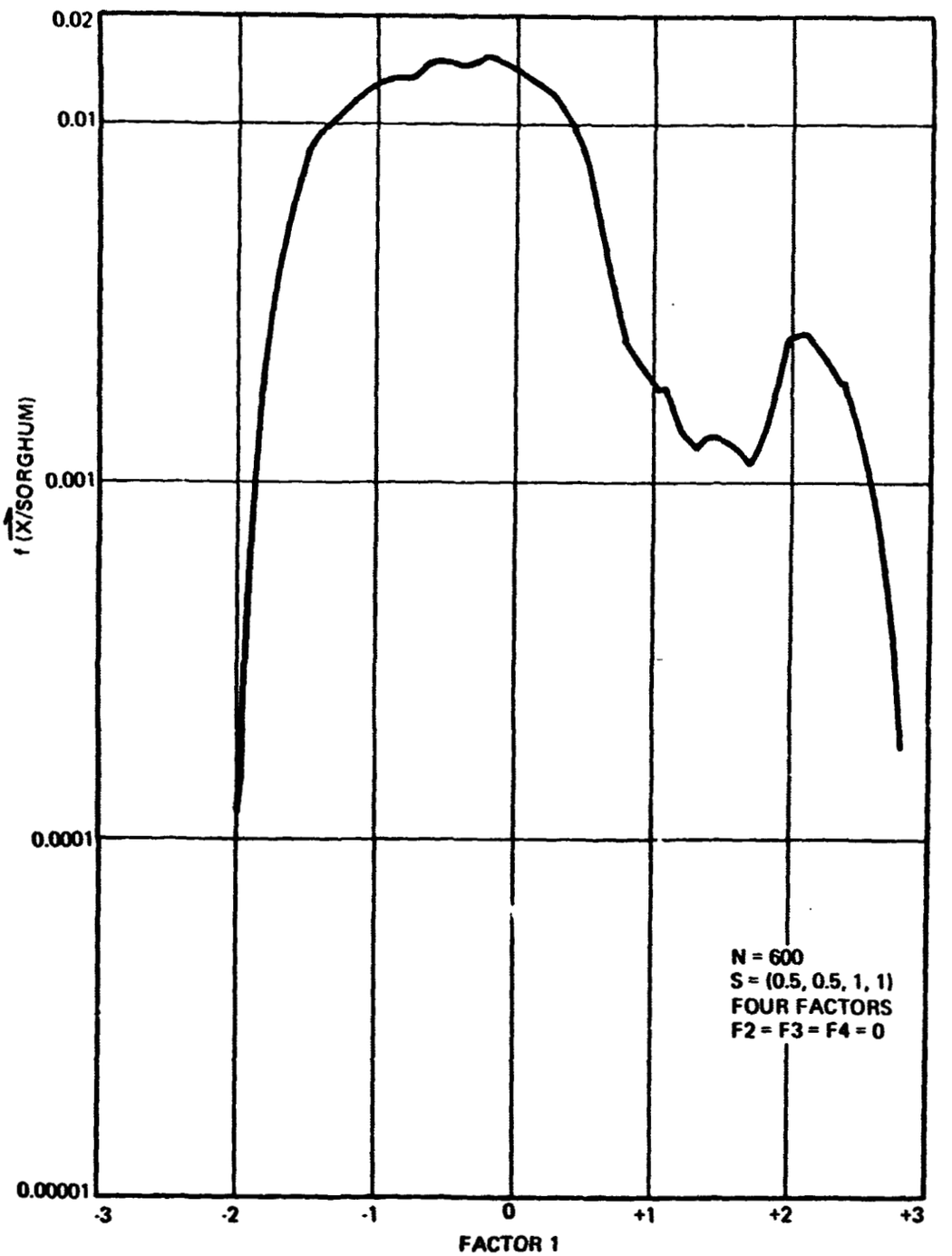


Figure 5-22 Marginal Density for Sorghum



8069-22

Figure 5-23 Marginal Density for Sorghum

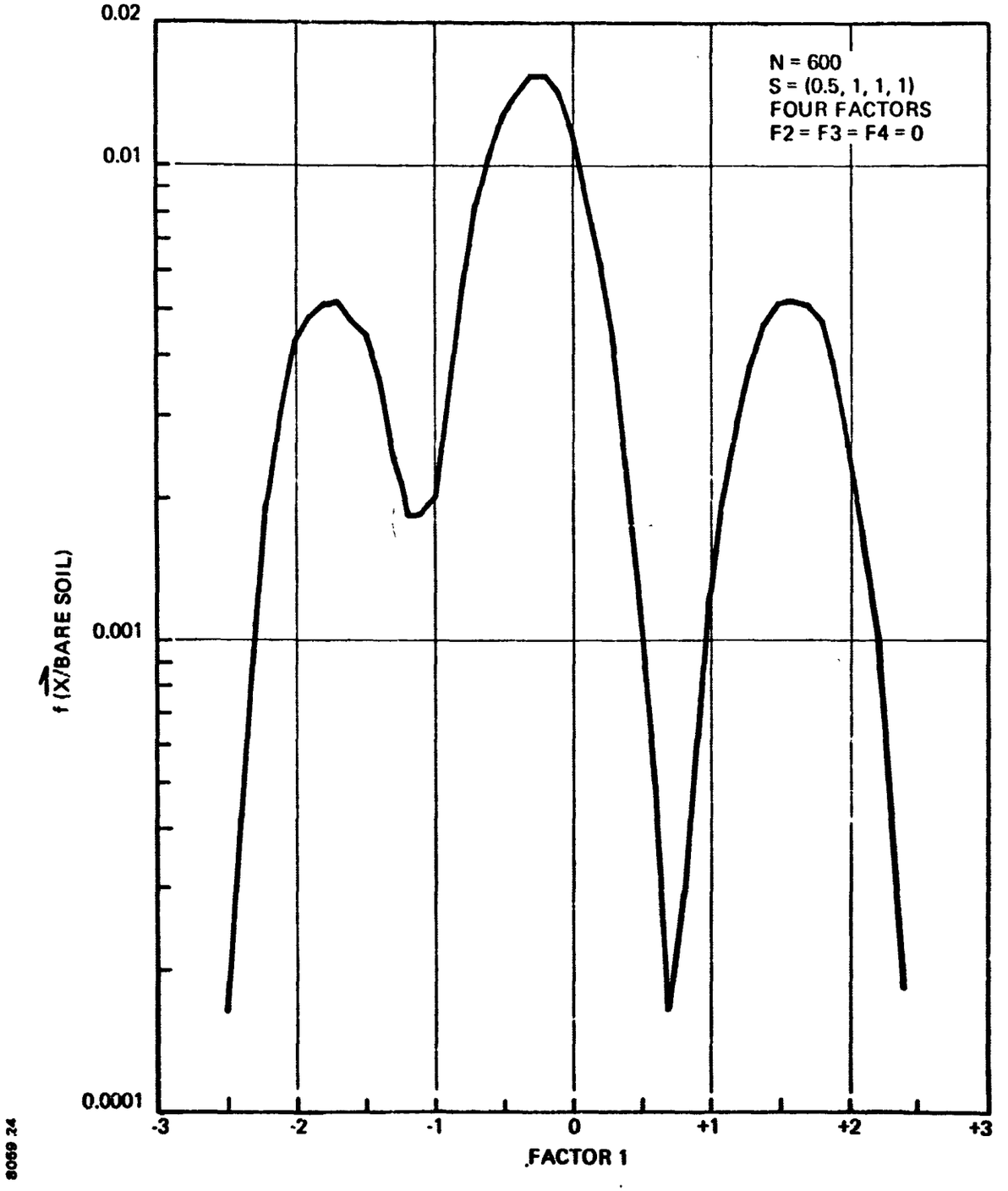
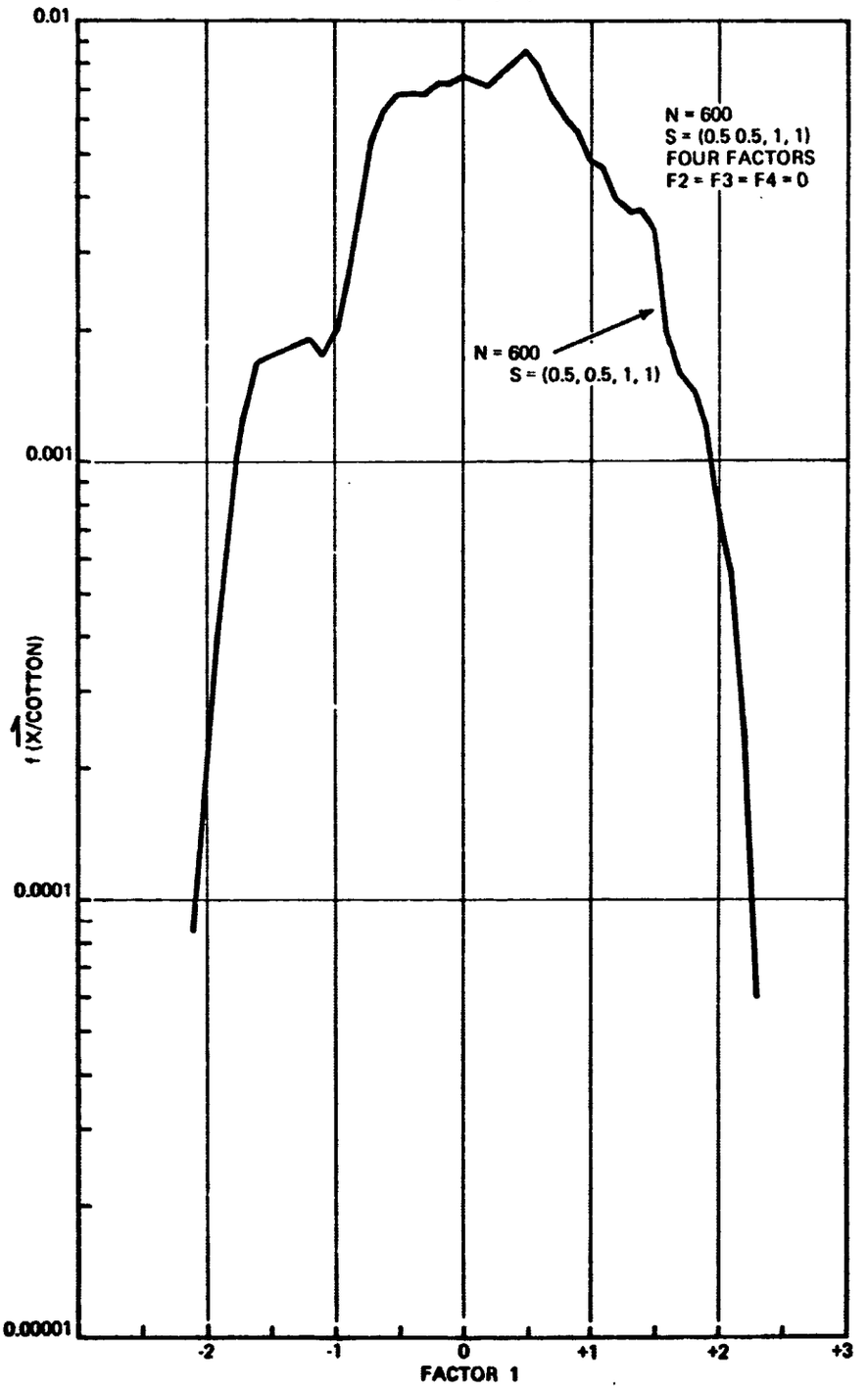


Figure 5-24 Marginal Density for Bare Soil



8059-23

Figure 5-25 Marginal Density for Cotton

The smoothing in Figure 5-22 appears satisfactory for the negative range of factor one up to about +0.5. The jagged curve for more positive values suggests under-smoothing. This part of the range is more smoothly represented in Figure 5-23 but at the expense of over-smoothing the negative and more highly populated portion of the distribution.

A variable smoothing scheme could be implemented with an Analog Table Look-Up Device (described in the following section) by allowing the amount of defocusing to be increased via a feedback loop until a sufficient amount of light passes through the transparency. The measure of probability density could then be taken from the focus voltage instead of the photomultiplier output.

5.2.2.3 Classification

Since computation of the probability densities is the most costly operation in the analysis, it is convenient to use a program which does the computation once and records the value for later use by other software. The Density Lister Program computes the hyperpyramid conditional probability density function for a sample signature; it is computed on the hypothesis of arising from a particular category. The value is appended to the signature on an external storage device as the (m+1)th component, where m is the number of channels in the signature.

Six-hundred raw data signatures of each crop type are sampled and organized into a data set to be transformed four times. The four resultant data sets then correspond to the four crops each transformed by the cotton factor score coefficient matrix, the four crops likewise transformed into sorghum space, the four crops into citrus space, and the four crops into bare soil space. Using the Density Lister Program, pyramid conditional probability densities are computed and appended to the signatures in each data set on the hypothesis that each actually came from a population similar to that of the training set corresponding to the transformation space in which they are expressed. Smoothing of the training sets of 600 factor scores of each crop is accomplished using the parameters selected in Section 5.2.2.2.

Histograms of the four sets of probability densities created by the Density Lister identify ranges of density values which typify each crop in each of the spaces. When a sample signature in truth belongs to a target on which the probability density computation is based and whose transformation was applied to the sample, its probability density tends to be larger than that of a non-target sample. For example, in Figure 5-26, more sorghum sample points have densities whose logs lie in the range -2.6 to -1.6, the extreme right of the scale, than do non-sorghum signatures, when the densities are computed with each sample transformed to sorghum space and attributed to the sorghum training set. This range of densities then better

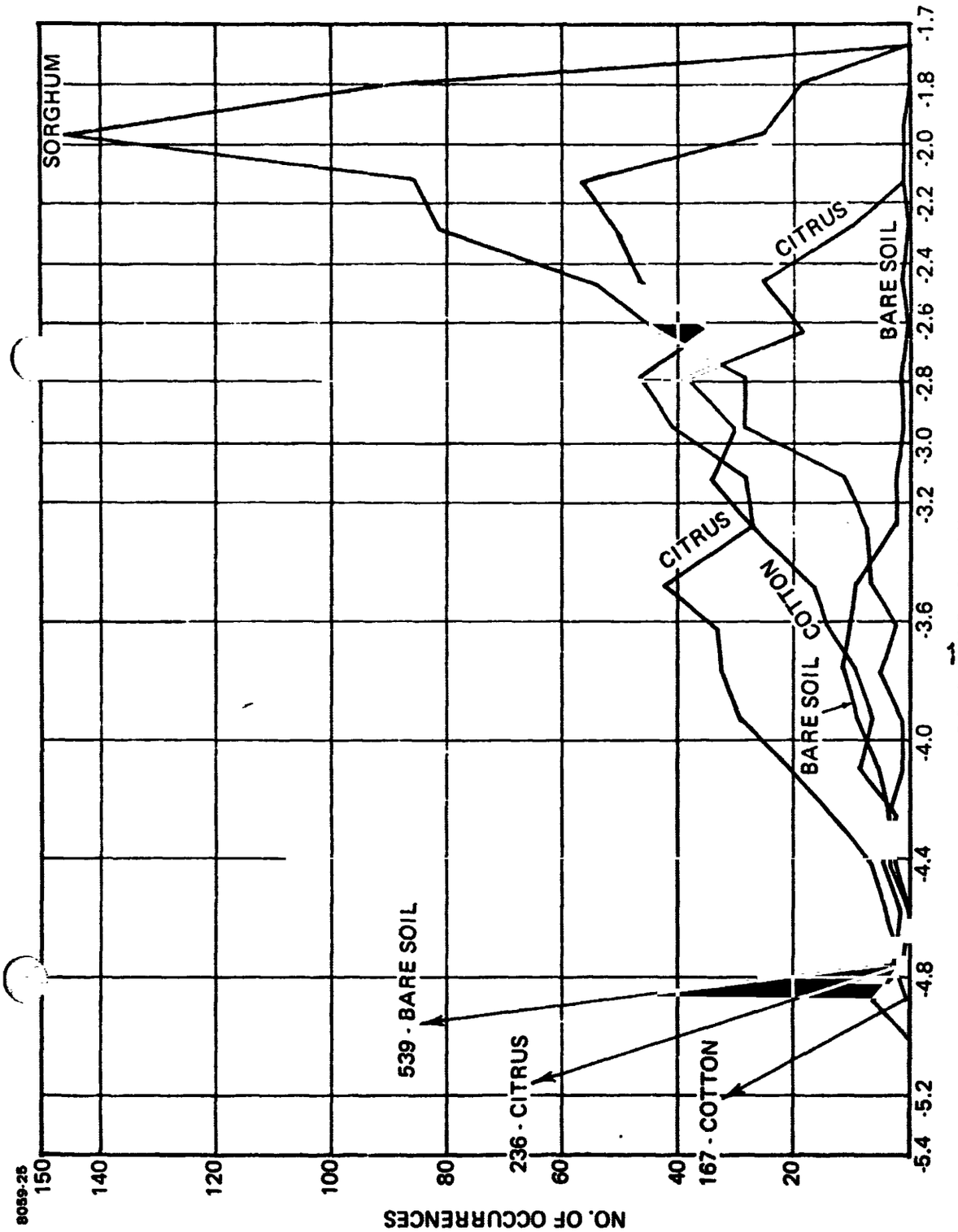


Figure 5-26 Distribution of $f(\vec{x}/\text{Sorghum})$ for Samples of N = 600

typifies sorghum than the other crops, although several citrus and cotton signatures have densities in the same range. True citrus signature densities based upon the citrus training set in citrus space fall into a range of values with about the same clarity as sorghum mentioned above. Note in Figure 5-27, that most of the citrus densities have logs lying in the range -3.2 to -2.0, although the other crops have several occurrences of densities in the same range. Bare soil in bare soil space, Figure 5-28 is most easily typified by a range of densities with non-soil density logs seldom entering the range -3.0 to -1.8. Thus, very little confusion of bare soil with cotton, sorghum, or citrus would occur with bare soil identification based solely upon probability density in bare soil space with a bare soil training set. In contrast, cotton is the most poorly typified crop on the basis of its probability densities in its factor space as compared to those of other targets in the same space. Note especially in Figure 5-29 the overlap of sorghum and cotton densities in cotton space. If the horizontal axis of Figure 5-29 were used as the gray-level scale on a film recorder with -6.6 representing clear and -1.8 black, much of what would appear as dark gray would not be cotton but sorghum, but decisions are not based on the conditional density for a single category. A decision is properly made by computing the conditional probability density for all categories for a single observation, and assigning that observation to the category which gives the largest conditional probability density, other considerations being equal. This is the method of maximum likelihood.

A classification program has only to read the appended density values from the four data sets produced by the Density Lister Program above. Category assignment of a sample signature is made to the target which gave the largest probability density when the sample signature was transformed into that target space and conditionally attributed to it. Classification results on the basis of such a rule are presented in Table 5-6 with the true classification indicated in Table 5-7. Citrus is classified with 75% accuracy, being most often misclassified as cotton. Cotton is classified with the least accuracy, 70%, with sorghum most often the misclassification. Sorghum and bare soil are correctly classified 82% and 94% of the time, respectively.

It is seen in Table 5-8 that the divergence results are completely consistent with the classification results obtained. Column 1 represents the divergence between the four crops in citrus space, Column 2 in cotton space, Column 3 in sorghum space, and Column 4 in bare soil space. Comparing each column of Table 5-8 with each column of Tables 5-6 and 5-7, it is noted that the off-diagonal matrix positions with the smallest divergences, indicating those pairs of crops for which the least information is present for separation, correspond to those matrix positions containing the largest misclassifications in the table of confusion, presented again in Table 5-8. For example, Column 1 of the divergence results indicates that the hardest distinction to be made is citrus from cotton, and the classification results show that citrus was most often mistaken for cotton.

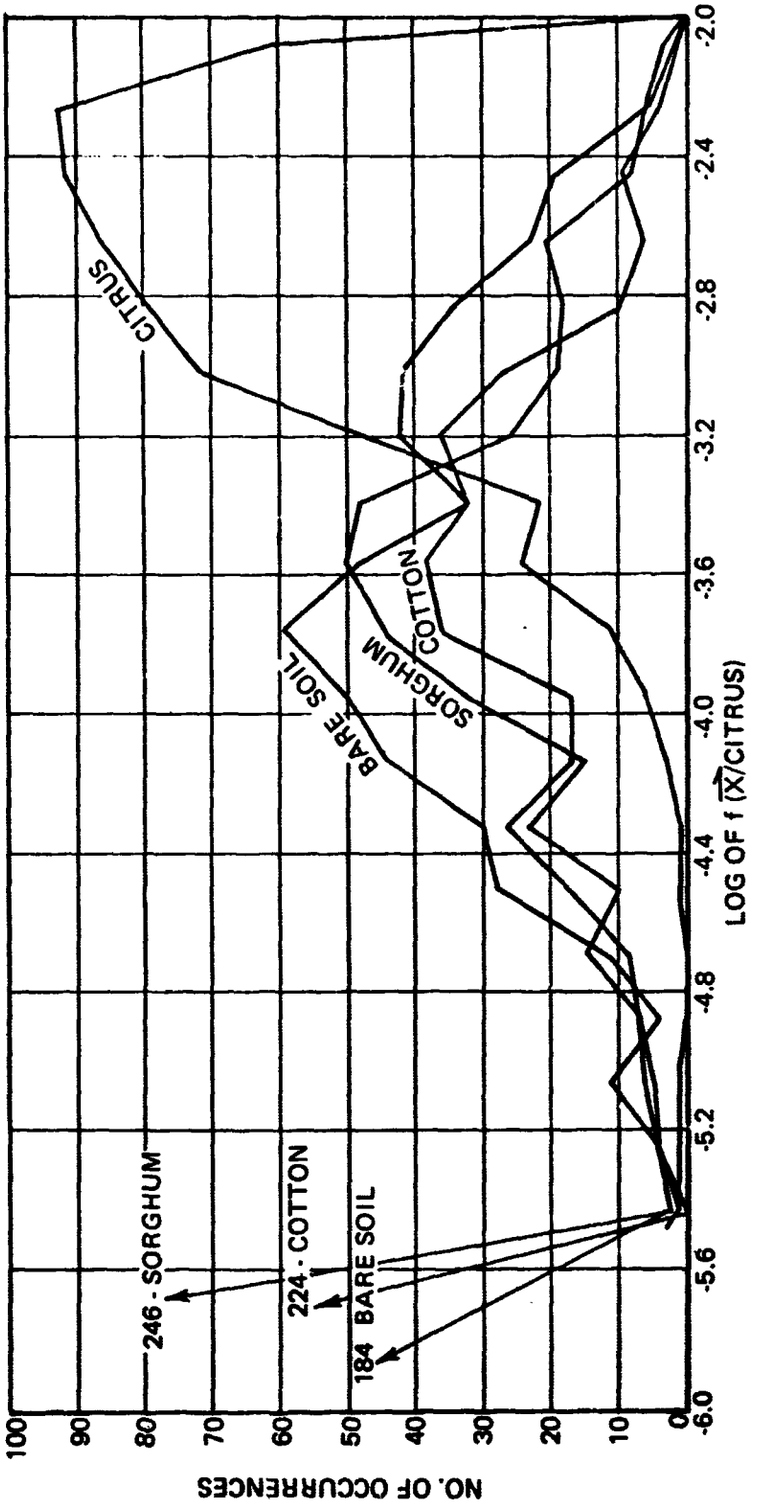
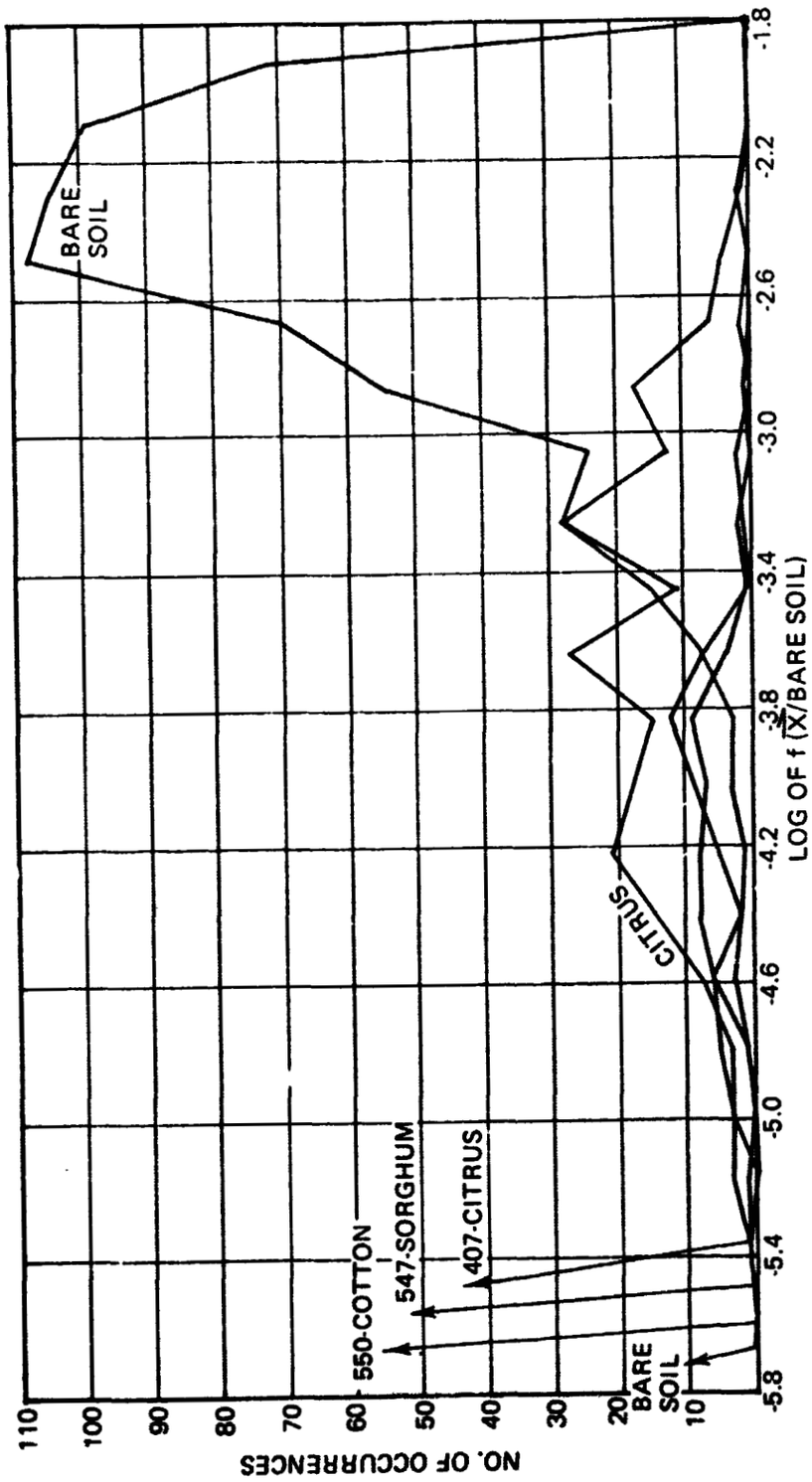


Figure 5-27 Distribution of $f(\bar{x}/citrus)$ for Samples of N = 600

8059-26



8069-27

Figure 5-28 Distribution of $f(\vec{x}/\text{Bare Soil})$ for Samples of N = 600

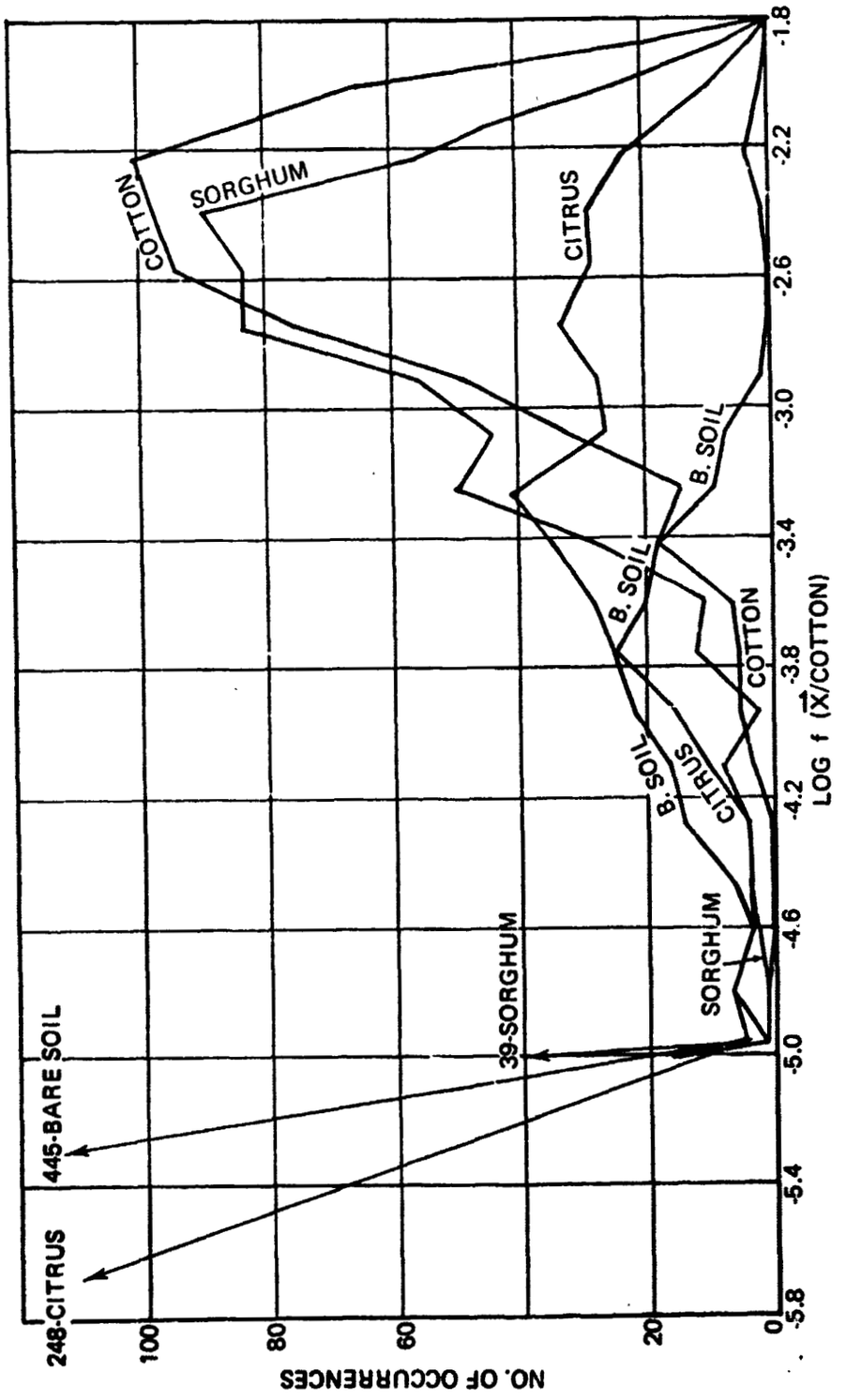


Figure 5-29 Distribution of $f(\vec{x}/\text{Cotton})$ for Samples of N = 600

TABLE 5-6

CLASSIFICATION RESULTS

Samples Known to be:

Target Assignment	Citrus	Cotton	Sorghum	Bare Soil
Citrus	449	32	21	29
Cotton	115	418	88	4
Sorghum	21	144	489	4
Bare Soil	14	2	1	562
Total	599	596 (4 not classified)	599	599 (1 not classified)
Accuracy	75%	70%	82%	94%

TABLE 5-7

EXPECTED CLASSIFICATION RESULTS

Samples Known to be:

Target Assignment	Citrus	Cotton	Sorghum	Bare Soil
Citrus	600	0	0	0
Cotton	0	600	0	0
Sorghum	0	0	600	0
Bare Soil	0	0	0	600
Total	600	600	600	600
Accuracy	100%	100%	100%	100%

TABLE 5-8

INFORMATION DIVERGENCE - SAMPLES OF 600 BITS OF INFORMATION

SPACE / SAMPLE	CITRUS	COTTON	SORGHUM	BARE SOIL
Citrus	(1.5) 0.72	50.0	56.8	85.8
Cotton	(48.2) 36.7	3.1	35.1	99.8
Sorghum	(49.4) 38.8	33.4	1.2	93.6
Bare Soil	(76.5) 62.2	95.1	93.5	2.3
SOURCE / ASSIGNMENT	Table of Confusion-Classification Results; 600 Samples			
	CITRUS	COTTON	SORGHUM	BARE SOIL
Citrus	(450) 449	32	(23) 21	(30) 29
Cotton	(113) 115	(417) 418	(86) 88	4
Sorghum	(22) 21	144	489	4
Bare Soil	(13) 14	2	1	(561)562
Total	(598) 600	(595) 596	599	599
	75%	70%	82%	94%

It must be noted that smoothing parameters used in calculation of the probability densities which led to the results discussed above are those chosen through the process described in Section 5.2.2.2, for citrus $\vec{S}$ = (0.5, 0.5, 1, 1) and for bare soil $\vec{S}$ = (0.5, 1., 1., 1.). The divergence matrix is not symmetric due to the use of two different spaces containing different amounts of information to compute the symmetrically located elements and to the use of different smoothing in different spaces. The divergence matrix might be expected to be more nearly symmetric, however, than that in Table 5-8, since an information measure should be an inherent property independent of the observations. This is another indication of the consequences of using a transformation for reduction of dimensionability without assurance that the variables used are high in information content and not just variance. It is, however, possible that the smaller divergences in citrus space may be caused by over-smoothing of the citrus training set instead of a basically information deficient set of variables. Referring once again to Figure 5-20, note the complete acceptability of the marginal density plot of citrus with $\vec{S}$ = (0.75 0.75, 1, 1). In an attempt to bring Column 1 of Table 5-8 in closer agreement with Row 1 of the table, and to bring the two-sample citrus divergence into the range of 1.0 to 4.0 bits found with the other diagonal elements of the table, the Column 1 divergences are recomputed with new smoothing, $\vec{S}$ = (0.75, 0.75, 1., 1.). The parenthetical results in Column 1 of Table 5-8 are those obtained with this smoothing. The result is a more nearly symmetric matrix as expected.

When the classification is repeated using conditional densities for citrus with this smoothing throughout, the table of confusion is affected scarcely at all. (See the parenthetical entries in the Table of Confusion.) The explanation that there is a genuinely smaller amount of information for the Column 1 distinctions is therefore tentatively accepted. Section 5.2.2.4 describes the attempt to measure the total amount of information available and to determine the best results obtainable by using the total amount of information.

5.2.2.4 Total Information Content

The tentative conclusions, to be tested by work described in this section, are that: (1) smaller than ordinary entries in the matrix of information divergence are significant measures of the relative difficulty of performing classifications, and (2) useful information is lost by using factor analysis for dimensionality reduction. If the conclusion is supported, then the divergence computation becomes a more reliable and useful measure of information because, when smoothing parameters are selected carefully enough that they are about right, the result of the divergence computation is more sensitive to information content than to adjustments in smoothing parameters or sample size.

Refer to Table 5-8 once more. The parenthetical entries for citrus with smaller smoothing parameters are larger than the originals in every case, but they are still smaller than the symmetrically located elements obtained by computing divergence in the other three spaces. Table 5-9 is obtained by dividing every element of Table 5-8 by the value of the diagonal element contained in the same column. The information divergence which was apparently improved for citrus by the use of smaller smoothing parameters is shown in Table 5-9 to have actually degraded the information-to-clutter ratio.

The interpretation of Table 5-9 as an array of information-to-clutter of S/N ratios treats the divergence on the diagonals of Table 5-8 as noise occurring naturally in each of the categories. Such noise, insofar as the processing methods are concerned, cannot be reduced once the training sample and the variables used to describe it are selected. Proper selection of samples and variables can reduce this noise, but for the current problem it is fixed.

Little comfort can be taken from the larger measures of information obtained for citrus by changing the smoothing parameters. Such a result is further support for the tentative conclusion that arrays such as Table 5-8, interpreted with the help of one such as Table 5-9 are more sensitive to information content than to peripheral influences such as the selection of smoothing parameters.

The two smallest off-diagonal elements of Table 5-9 are the symmetrically located cotton-sorghum and sorghum-cotton elements. Cotton emerges as the category most difficult to identify (only 70% correct) and also as the category for which both citrus and sorghum are most easily mistaken. Figures 5-26, 5-27, 5-28, and 5-29 show the distribution of the conditional probability densities for each of the categories according to the actual category showing the same thing.

None of the measures -- confusion in actual assignment, information divergence, or overlap in the distribution of conditional densities -- is symmetric. Compare, for example, Figures 5-26 and 5-29: cotton in cotton space is more easily mistaken for sorghum than sorghum in sorghum space for cotton.

Recall that only the first four principal components for each category have been used thus far. It could be that the asymmetry in the ease with which cotton and sorghum are mistaken for each other is because the principal components of sorghum are, fortuitously, useful for distinguishing the two while the principal components of cotton are useful mainly for describing the variance in cotton.

TABLE 5-9

INFORMATION-TO-CLUTTER RATIOS

	Citrus	Cotton	Sorghum	Bare Soil
Citrus	1	16	47	37
Cotton	(32) 51	1	29	43
Sorghum	(33) 54	11	1	41
Bare Soil	(51) 86	31	78	1

In the table of confusion, Table 5-8, cotton is more often mistaken for sorghum than sorghum for cotton - perhaps for this reason. If this is the case, then the asymmetry in the cotton-sorghum identification problem should be eliminated or made less so by the use of all information available.

To determine the value of the information contained in the last four principal components, several additional operations are performed. The last four principal components in the space of each category considered are used to repeat the studies performed thus far with the first four principal components. The operations performed are:

1. Using the last four components only, and the density lister program as in Section 5.2.2.3, compute hyperpyramid conditional probability densities for each category.

2. Compute the matrix of information divergence as in Table 5-8 for the last four only. This matrix is presented in Table 5-10.

3. Compute and plot histograms for each of the categories, showing the distribution of probability densities by category . These are shown in Figures 5-30 through 5-33.

4. Apply the decision rule (maximum liklihood) used in Section 5.2.2.3 to the densities obtained from the last four principal components. The resulting table of confusion is presented in Table 5-10, below the matrix of information divergence.

5. On the assumption of statistical independence of the first four and last four principal components, compute the conditional density as, for example:

$$f(\vec{X} \mid \text{citrus}) = f_1(\vec{X}_1 \mid \text{citrus})\, f_2(\vec{X}_2 \mid \text{citrus})$$

where

$\vec{X}$ = the vector formed by all principal components

$\vec{X}_1$ = the vector formed by the first four

$\vec{X}_2$ = the vector formed by the last four

$f(\vec{X} \mid \text{citrus})$ = general form of conditional density

$f_1(\vec{X}_1 \mid \text{citrus})$ = conditional density of first four principal components

TABLE 5-10

COMPUTATIONS WITH LAST FOUR PRINCIPAL COMPONENTS ONLY

Information Divergence - Samples of 600-Bits of Information

Space / Sample	Citrus	Cotton	Sorghum	Bare Soil
Citrus	2.0	6.0	5.6	12.4
Cotton	2.3	1.3	7.5	19.4
Sorghum	3.5	3.6	1.7	13.8
Bare Soil	23.3	21.4	50.0	1.6

Table of Confusion - Classification Results - 600 SAMPLES

Source / Assignment	Citrus	Cotton	Sorghum	Bare Soil
Citrus	331	151	160	34
Cotton	64	293	113	49
Sorghum	68	102	215	8
Bare Soil	135	52	110	504
Total	598	598	598	595
Percent Correct	55%	49%	36%	84%

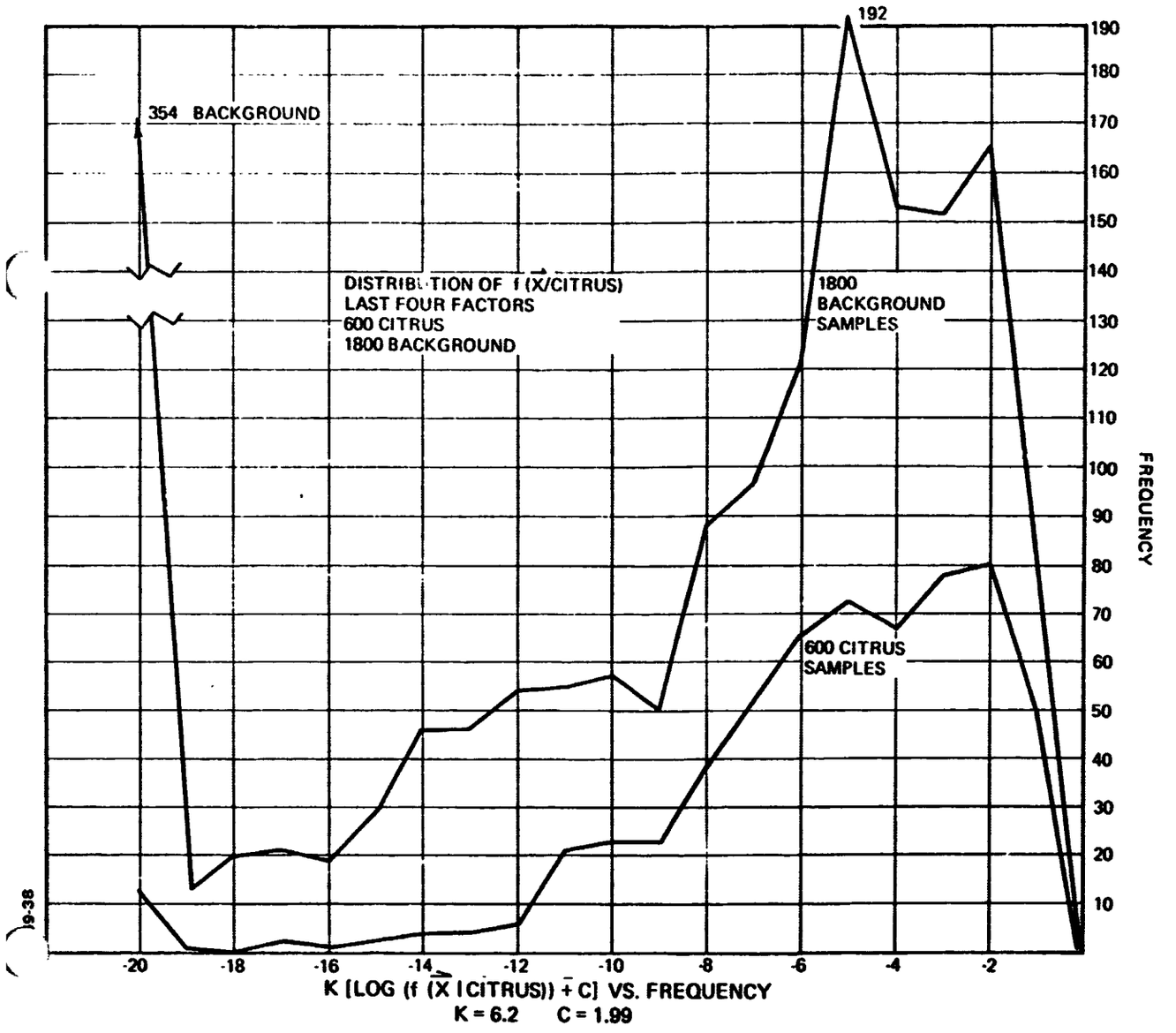


Figure 5-30 Distribution of Citrus Densities

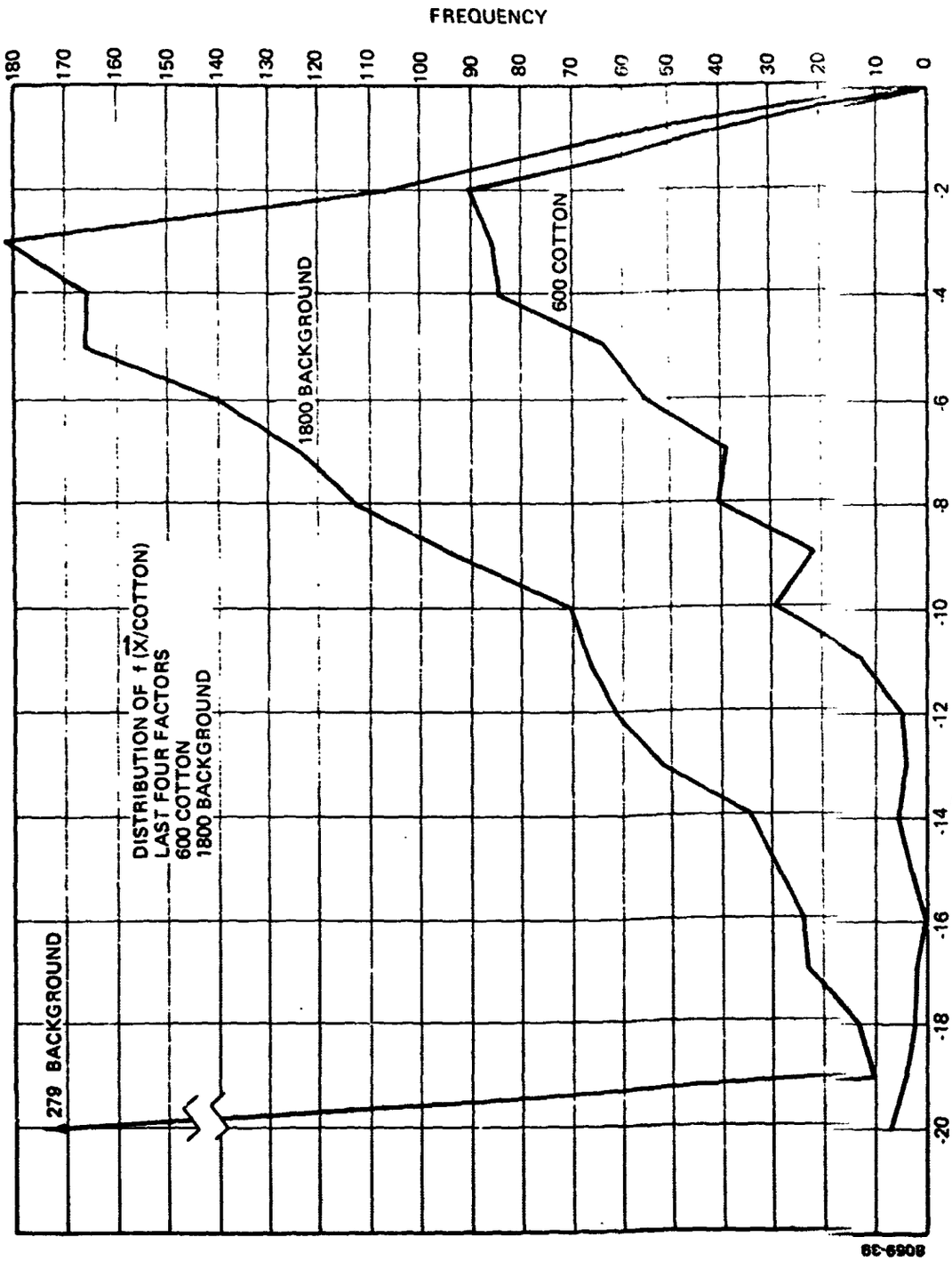


Figure 5-31 Distribution of Cotton Densities

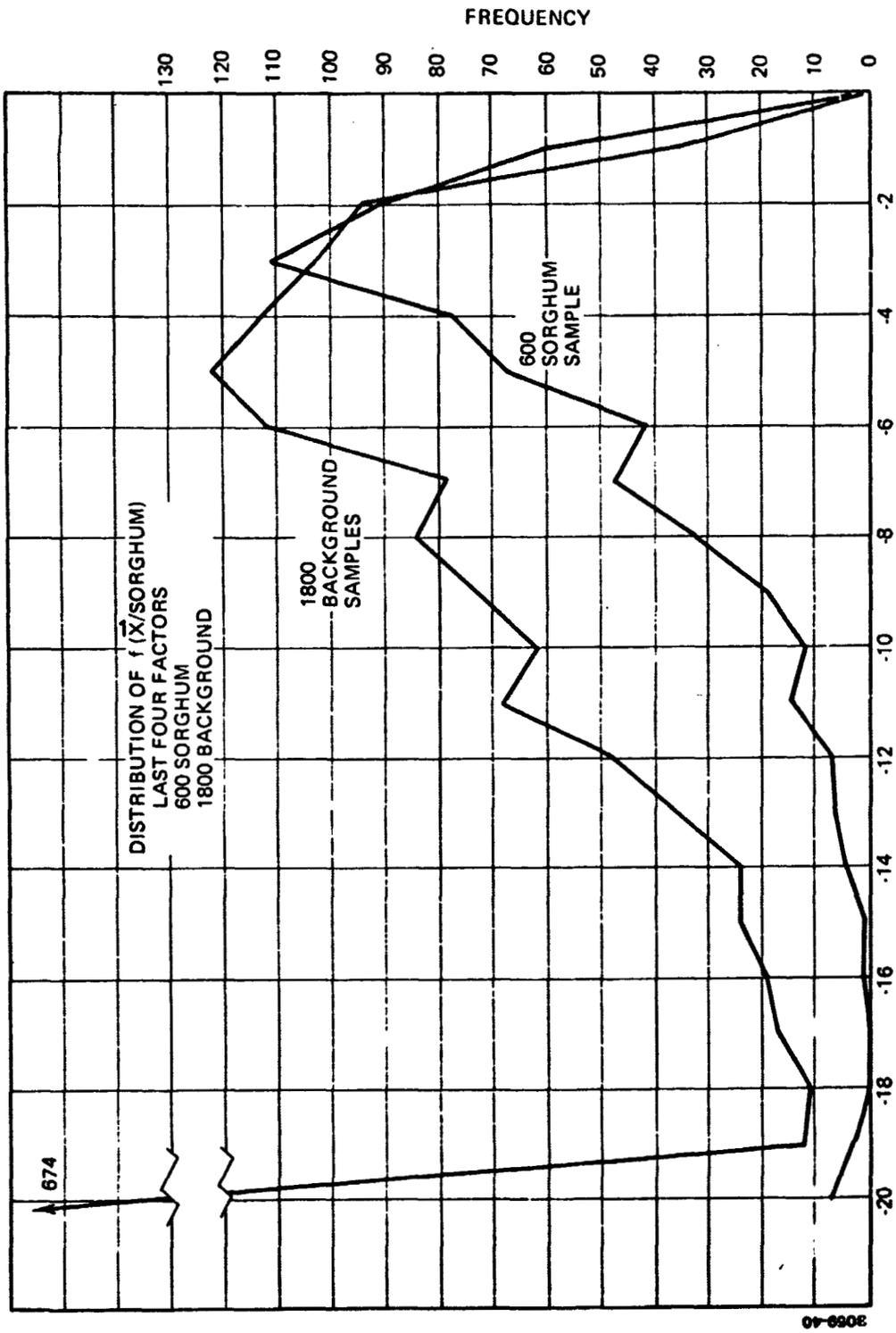


Figure 5-32 Distribution of Sorghum Densities

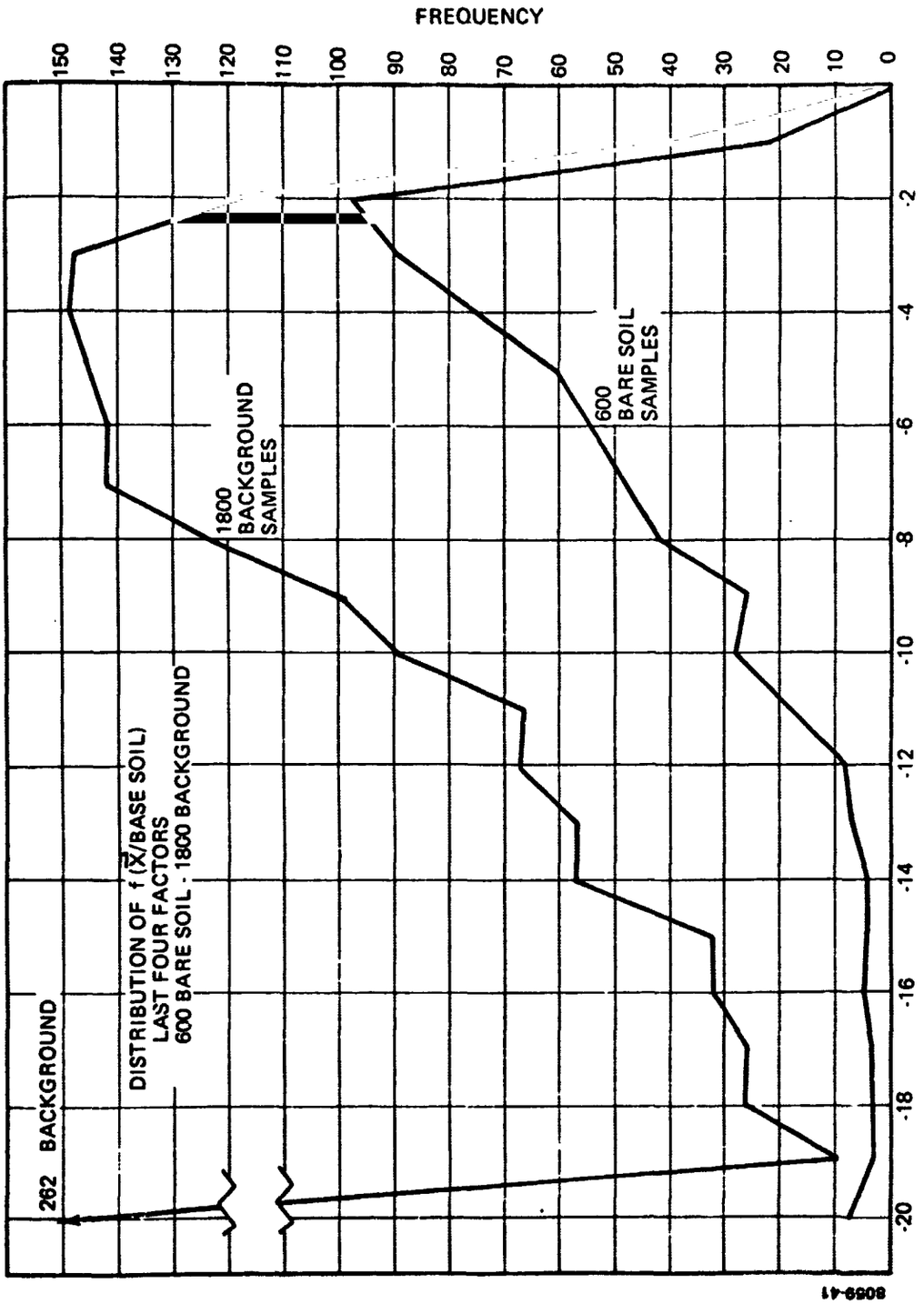


Figure 5-33 Distribution of Bare Soil Densities

$f_2(\vec{X}_2 \mid \text{citrus})$ = conditional density of last four principal components.

6. Compute histograms showing distribution of $f(\vec{X} \mid \ldots)$ by category and background. These are presented in Figures 5-34 through 5-37.

7. Determine the matrix of information divergence and the table of confusion using all available information. These results are presented in Table 5-11.

Consider first the direct measures of the information contained in the last four principal components. The matrix of information divergence (upper part of Table 5-10) shows the expected smaller amounts of information for these variables. The information-to-clutter-ratios, Table 5-10, still exceed one in every case. Classification results obtained applying the maximum likelihood rule to conditional probability densities of the last four principal components (lower part of Table 5-10) are not particularly good, but are clearly better than expected by chance.

The reference samples used to compute these densities are the same as used earlier for the first four components. The smoothing parameters used on the last four are, in every case, equal to 1.0. The scatter diagrams presented in Figure 5-16 show no evidence of multiple modes as clustering in factors 5, 6, 7, or 8. They have a variance of 1.0 by definition. The smoothing parameters of 1.0 for these factors are therefore used without further investigation based on the experience described earlier.

The histograms of Figures 5-30 through 5-33 provide another subjective look at the information content of these variables. The three non-target categories in each case are lumped together and classified as background; the backgrounds were plotted separately in Figures 5-26 through 5-29. Consequently, the background curve in each case now shows the distribution of three times as many samples as does the target curve. Remember that no decision rule is based on the value or distribution of any single function such as $f(\vec{X} \mid \text{citrus})$ shown in Figure 5-30. The decision rule requires computation of all four conditional densities $f(\vec{X} \mid \text{citrus})$, $f(\vec{X} \mid \text{sorghum})$, $f(\vec{X} \mid \text{cotton})$, and $f(\vec{X} \mid \text{bare soil})$, at a single point $\vec{X}$. The observation $\vec{X}$ is then assigned to the category given the largest conditional density.

However, such curves do suggest a direct interpretation for the probability density imagery presented in Section 6. The use of a single flying spot scanner to produce such imagery is comparable to mapping all observations $\vec{X}$ into a gray scale

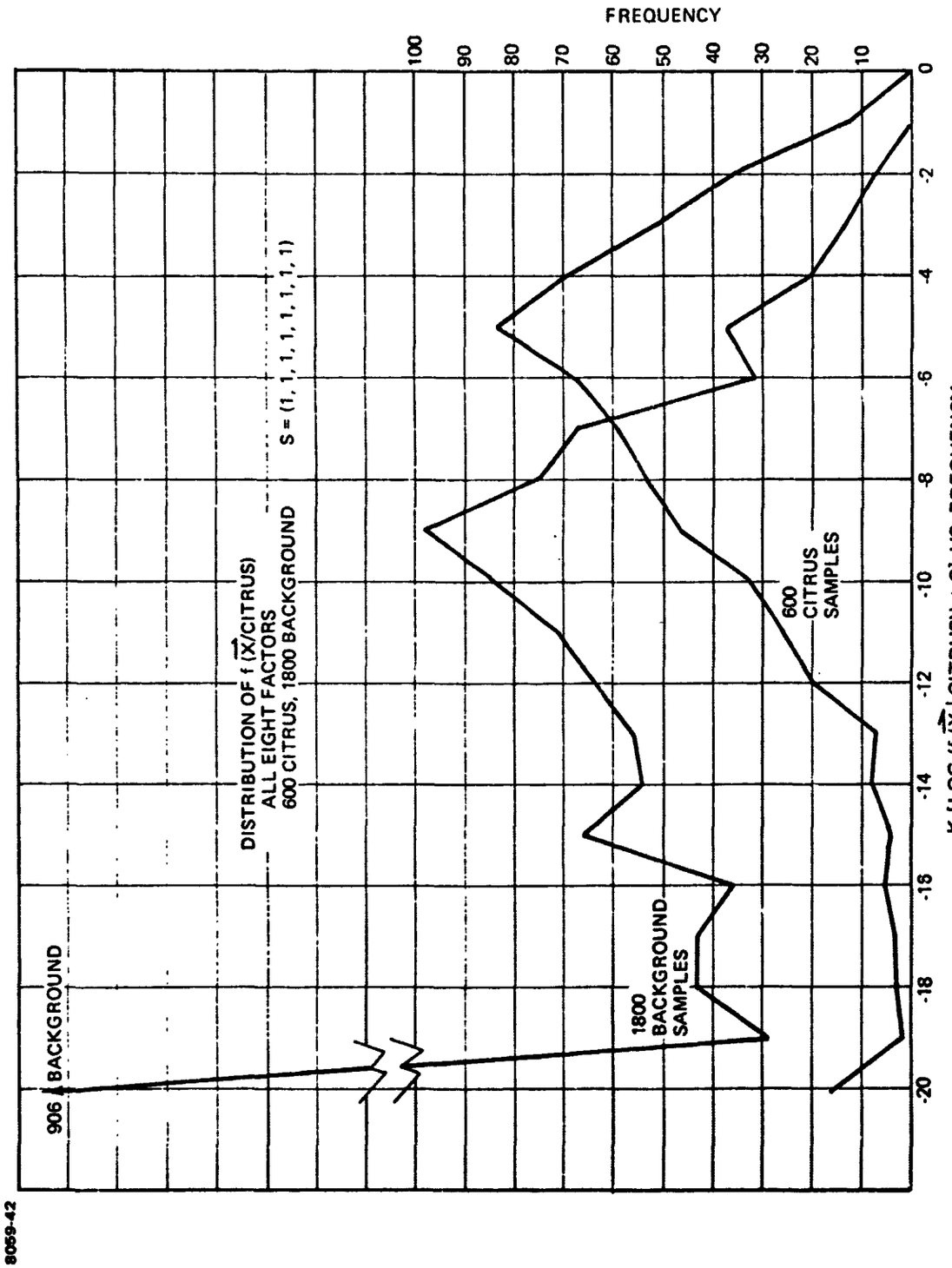
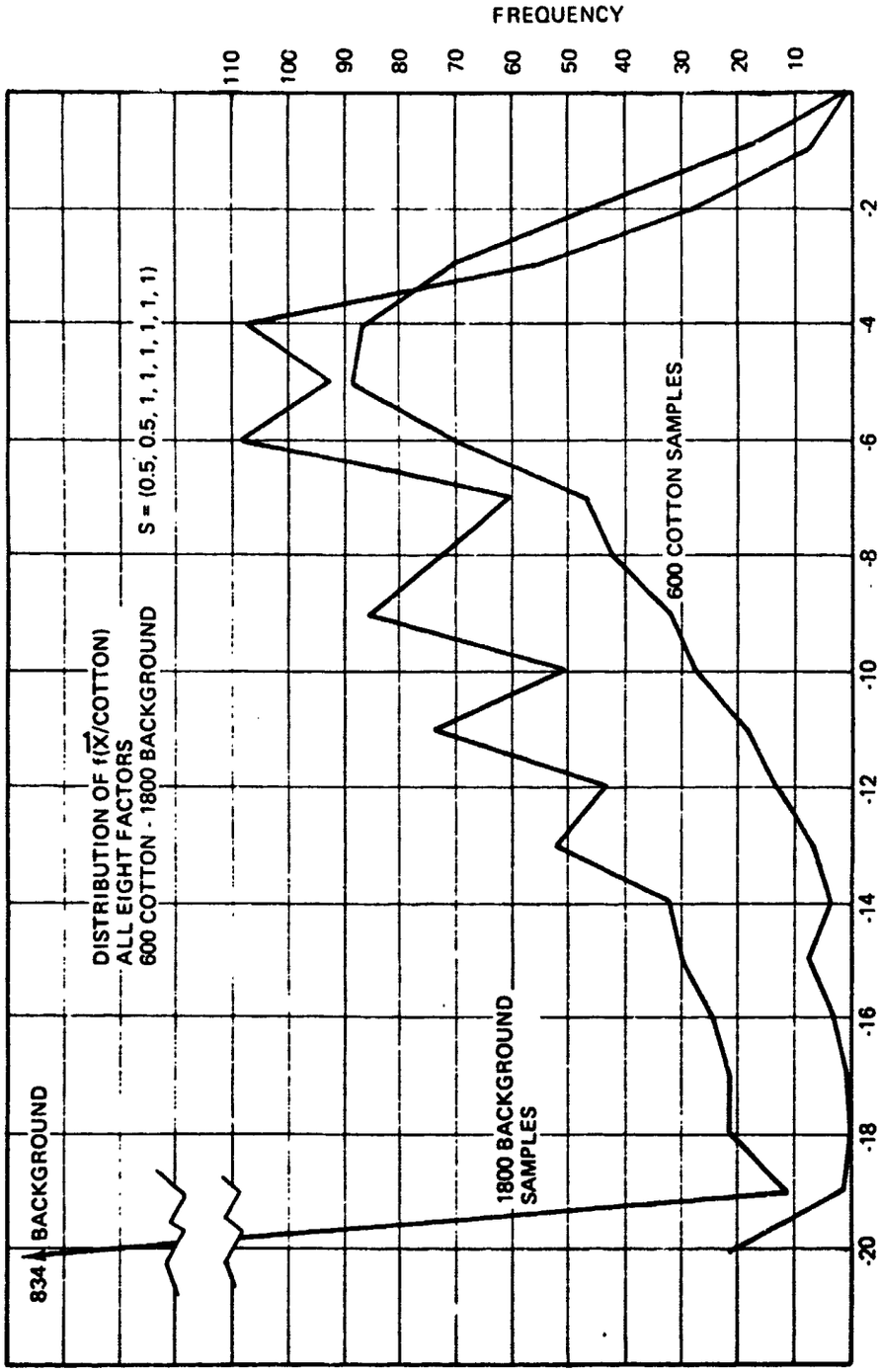


Figure 5-34 Distribution of Citrus Densities; All Eight Factors



8059-43

Figure 5-35 Distribution of Cotton Densities; All Eight Factors

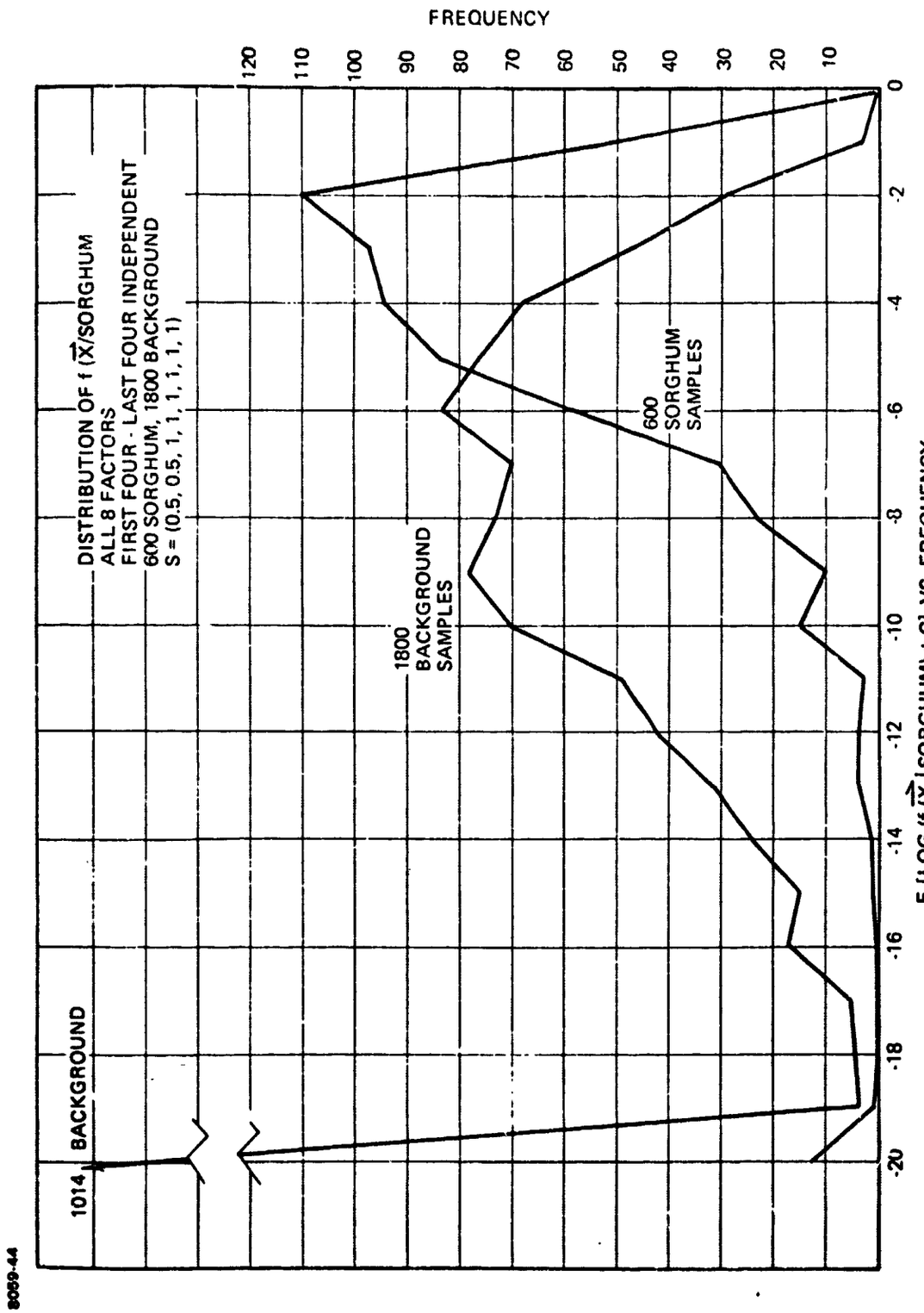


Figure 5-36 Distribution of Sorghum Densities; All Eight Factors

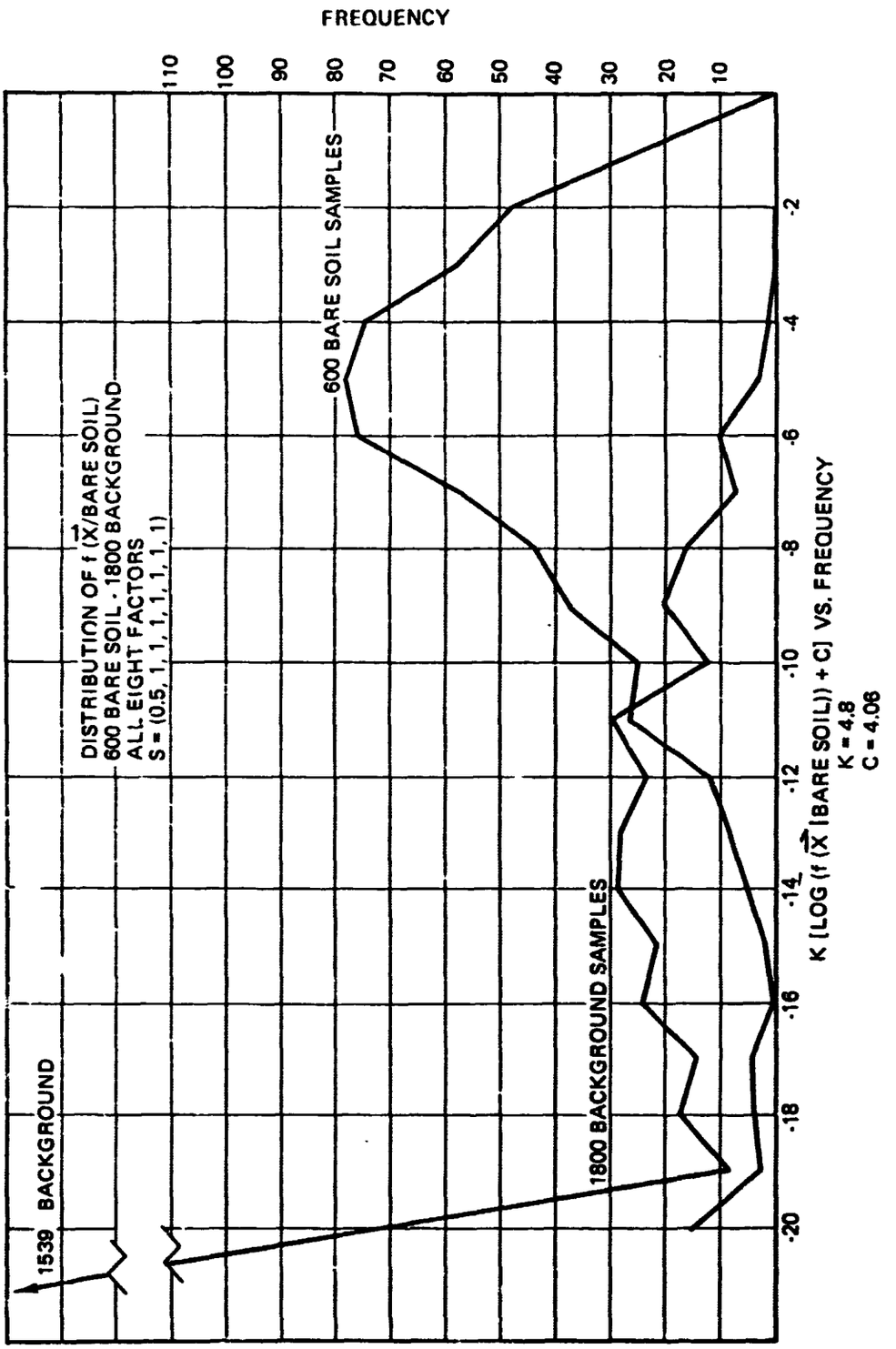


Figure 5-57 Distribution of Bare Soil Densities; All Eight Factors

8059-45

TABLE 5-11

COMPUTATIONS WITH ALL EIGHT PRINCIPAL COMPONENTS

Information Divergence - Samples of 600
- Sums of Information Bits
from first and last four factors

Space / Sample	Citrus	Cotton	Sorghum	Bare Soil
Citrus	2.7	56.0	62.4	98.2
Cotton	45.0	4.4	42.6	119.2
Sorghum	42.3	37.0	2.9	107.4
Bare Soil	85.5	116.5	143.5	3.9

Table of Confusion - Classification Results; 600 Samples

Source / Assignment	Citrus	Cotton	Sorghum	Bare Soil
Citrus	492	27	27	13
Cotton	73	445	104	6
Sorghum	17	113	464	2
Bare Soil	10	6	1	570
Total	592	591	596	591
Percent Correct	82%	75%	77%	95%

on film in which the film density is proportional to, for example, $f(\vec{X}|$ citrus). The frequency with which a high film density accurately identifies the desired target category is indicated by a histogram such as Figure 5-27 or Figure 5-30. This is discussed further in Section 6.

The fifth operation, described above, is then carried out for each of the target categories. The maximum likelihood decision rule is applied to the conditional probability densities which result. The matrix of information divergence is computed directly by adding the two presented in Tables 5-8 and 5-10. Note that information contained in the two factors of a general probability density is additive because of

$$f = f_1 f_2$$

and

$$g = g_1 g_2$$

then

$$\log \frac{f}{g} = \log \frac{f_1}{g_1} + \log \frac{f_2}{g_2} .$$

The results obtained by computing a hyperpyramidally smoothed conditional probability density for the first and last four principal components separately, assuming statistical independence of the first and last four variables, and forming the product of the two to give an approximation of unknown goodness to the general multivariate density in 8 variables, are presented in Table 5-11.

More correct identifications were made in each category but sorghum. To conclude, for this reason, that the information contained in the last four principal components of sorghum is not useful is unfair. It is possible that the information contained there (tabulated and illustrated in Table 5-10) is not properly used here, if for example, the last four and the first four principal components are not statistically independent. This can be tested but has not been accomplished within the scope of this study. Consider as an alternative explanation, the possibility that the extent to which cotton and sorghum signature distributions actually overlap and become indistinguishable was not properly represented in the first four principal components alone. In this context chance, and not skillful handling of information, can produce correct results for the wrong reasons.

Note further that a price is paid for correct classifications obtained for the wrong reasons. It would be not at all difficult to devise a conditional density function for sorghum that would properly classify 100% of the sorghum. The price would be the improper classification of many non-sorghum observations as sorghum. Such a price can be found in terms of cotton classified as sorghum in Table 5-8. In fact, the percentage by which cotton classification improves in Table 5-11 is very nearly equal to the percentage by which sorghum classification is degraded.

The asymmetry in the cotton-sorghum classification problem has actually been eliminated by applying the information in the last four principal components.

Therefore, it is concluded that the use of principal component analysis or factor analysis for dimensionality reduction is done at the risk of losing information of real value. It appears to serve reasonably well as a device to allow density functions to be factored, thereby reducing the amount of data that must be stored to store a given amount of information. Information divergence appears to serve well as at least a relative measure of information content and problem difficulty.

The problem of selecting smoothing parameters and providing large enough samples for multivariate analysis is by no means solved. Further manipulation of these parameters would doubtless permit still further improvements in ability to represent density functions and to classify observations in the best possible way. If it were economically feasible to use reference samples sufficiently large, the problem of wisely selecting smoothing parameters would be considerably attenuated. It might become possible to neglect the directionality of smoothing parameters. It has not been possible to do more than begin these investigations in this study.

Tables 5-12 and 5-13 are similar to Table 5-9 showing the information-to-clutter ratio matrix derived from the information divergence matrix. Table 5-9 showed the result obtained using the first four principal components; Tables 5-12 and 5-13 show the result obtained using the last four and all eight, respectively. The ratios apparently do not provide an unambiguous picture of the relative value of the three coordinate system choices. The poor ratios in the last four do provide useful information to assist in the classification problem when used with the first four; the information is added. The ratios which result for all eight are rendered poorer than those for the first four. However, the reference sample sizes could be increased until the diagonal information divergences become small. By adjusting smoothing parameters and increasing sample sizes, taking care to avoid spurious oversmoothing to obtain small diagonal elements, it could be determined whether good representations of probability density also have the unsatisfying property of allowing better results to be obtained with a poorer information-to-clutter ratio.

TABLE 5-12

INFORMATION-TO-CLUTTER RATIOS —
LAST FOUR PRINCIPAL COMPONENTS

	Citrus	Cotton	Sorghum	Bare Soil
Citrus	1	4.6	2.9	7.8
Cotton	4.1	1	4.4	12.1
Sorghum	1.8	2.0	1	8.6
Bare Soil	11.6	16.5	29.4	1

TABLE 5-13

INFORMATION-TO-CLUTTER RATIO —
ALL EIGHT PRINCIPAL COMPONENTS

	Citrus	Cotton	Sorghum	Bare Soil
Citrus	1	12.7	21.5	25.2
Cotton	16.7	1	14.7	30.6
Sorghum	15.8	8.4	1	27.6
Bare Soil	31.7	26.5	49.5	1

However, the objective of finding a linear transformation suitable for dimensionality reduction has suggested the development of canonical analysis discussed in the next section.

5.3 CANONICAL ANALYSIS

A linear transformation used for dimensionality reduction must be contrived to concentrate on the problem of interest. All variance of observations in a hyperspace is not of equal interest. When principal component analysis is performed on observations from a single category, the transformation which results is intended only to separate the underlying phenomena contributing to variance within the category. Pronounced clusters observed in the plane of the first two factors as principal components attest to the multiple influences which cause variance in a single category. Other factors, which account for less variance in the category of interest, can remain important for the purpose of distinguishing the category of interest from other categories.

For classification purposes, the variance within the category of interest can be regarded as noise and the variance between the category of interest and all others as signal. The problem of determining a line or transformation which maximizes the signal—good variance—with respect to the noise—bad varaince—can be expressed as an eigenvalue problem.

A linear transformation which should prove useful for dimensionality reduction is suggested by the eigenvalue problem:

$$GW = BWd \qquad (5.3\text{-}1)$$

since the direction associated with the larger eigenvalues are those in which the background variance is maximized relative to the target variance. The notation used is:

Symbol

G	=	covariance matrix of background observations (good variance)
B	=	covaraince matrix of target observations (bad variance)
W	=	matrix whose columns are the eigenvectors of G relative to B
d	=	diagonal matrix of the eigenvalues associated with W
U	=	unitary matrix whose columns are the eigenvectors of G
g	=	diagonal matrix of the eignevalues of G
H	=	unitary matrix whose columns are the eignevectors of B

b = diagonal matrix of the eigenvalues of B

I = the identity matrix

Matrices G and B may be factored readily by use of the unitary transformations U and H:

$$G = UgU^T \tag{5.3-2}$$

$$B = HbH^T \tag{5.3-3}$$

Note that:

$$B = Hb^{1/2}b^{1/2}H^T \tag{5.3-4}$$

and a matrix useful in the sequel can be defined:

$$B^{1/2} = b^{1/2}H^T \tag{5.3-5}$$

If B is non-singular:

$$B^{-1/2} = Hb^{-1/2} \tag{5.3-6}$$

Since the symmetry of G is preserved by the indicated row and column transformations, the transformed matrix can be factored by use of the unitary matrix M:

$$b^{-1/2}H^TGHb^{-1/2} = MdM^T \tag{5.3-7}$$

$$GHb^{-1/2} = Hb^{1/2} = Hb^{1/2}MdM^T \tag{5.3-8}$$

$$GHb^{-1/2}M = Hb^{1/2}Md \tag{5.3-9}$$

Let:

$$W = Hb^{-1/2}M \tag{5.3-10}$$

$$BW = HbH^THb^{-1/2}M \tag{5.3-11}$$

Equation 5.3-1 can then be obtained by combining Equations 5.3-9, -10, and -11 demonstrating that the matrix W as defined by Equation 5.3-10 is the matrix of eigenvectors of G relative to B, and that matrix d as determined in Equation 5.3-7 is composed of the associated eigenvalues.

The canonical variables resulting from the W transformation will be of unit variance and uncorrelated for observations with covariance matrix B, as shown by:

$$W^T B W = M^T b^{-1/2} H^T H b H^T H b^{-1/2} M \qquad (5.3\text{-}12)$$

$$= M^T M$$

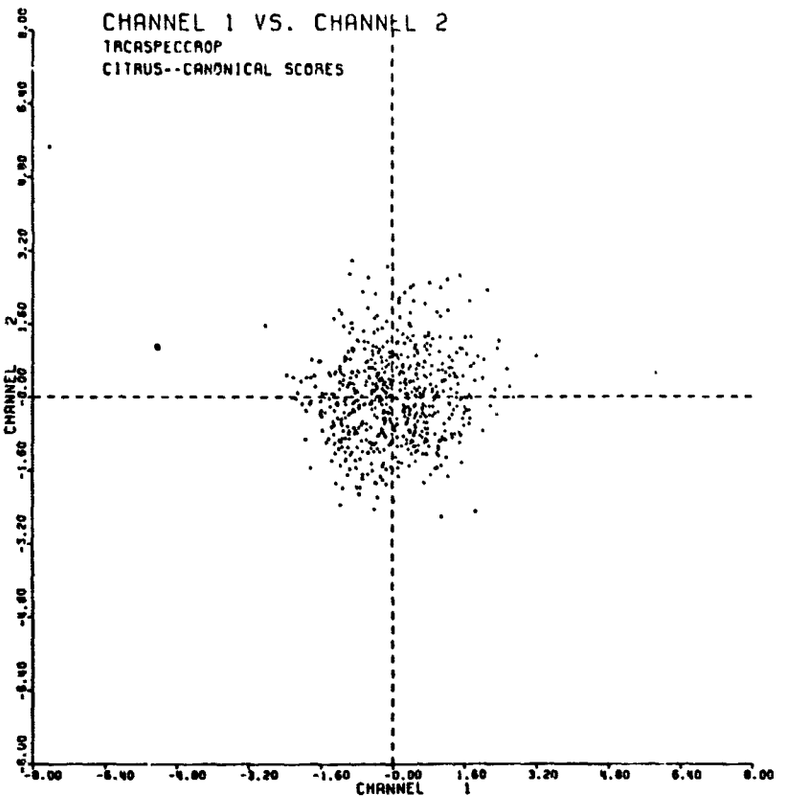
$$= I$$

The two covariance matrices, G and B, serve to distinguish noiselike variance, found within the category of interest, and signal variance. The transformation produced by solution of this eigenvalue problem is specific for the single category whose covariance is described by B. Just as for the multiple principal component transformations, a preferred space is used for computation of the condition probability densities for each category; i.e., cotton space for f (x|cotton), etc.

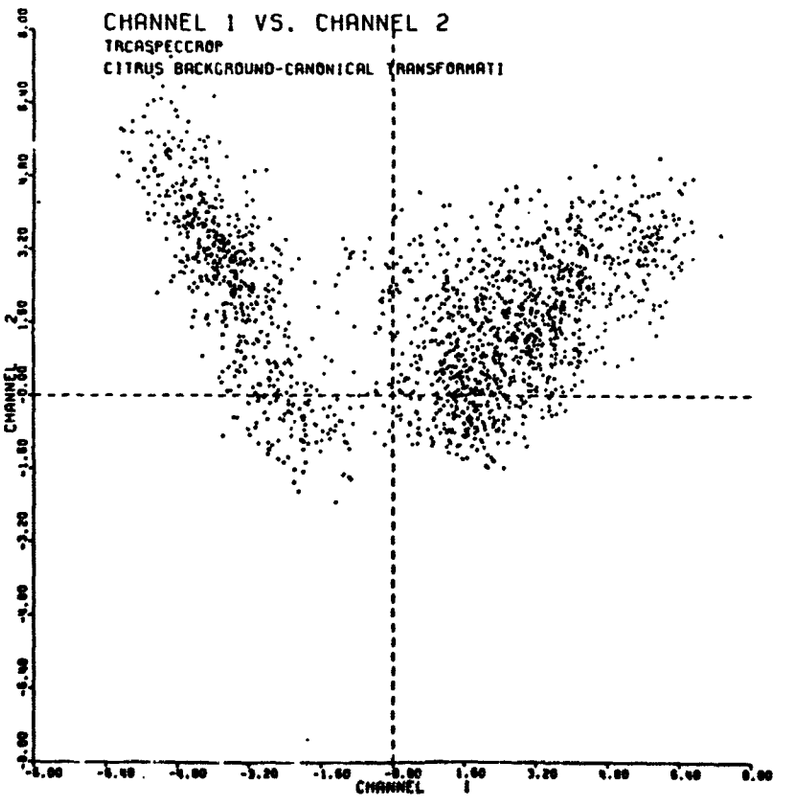
A crude computer program has been improvised to compute the linear transformation indicated. The four crop types described earlier are treated, each in turn, as the target category. The covariance matrix of the target category is used for B, defined above, and the covariance matrix of the background non-target categories, taken about the mean values of the target category, is used for G.

Scatter diagrams are then prepared of the first two pairs of canonical variables. Each target category and its background is observed in the plane of canonical scores 1 and 2 and of canonical scores 3 and 4 in the form of scatter diagrams. This series of scatter diagrams is presented in Figures 5-38 through 5-41 for citrus, cotton, sorghum, and bare soil, respectively.

When working in the space of canonical variables as defined, there is no evidence of clusters or multiple modes within a single category. This is the clearest evidence of the different objectives of principal component and canonical analysis. Even bare soil—a category which shows three distinct modes in every other space considered including untransformed channel space—appears in the planes of pairs of canonical variables as a distribution that looks bivariate normal. Bare soil scatter diagrams for the last four canonical variables are also presented in Figure 5-42. No multiple clusters are observed in either of these planes. The interpretation placed on the eigenvalues, d, associated with a canonical transformation, is illustrated by these scatter diagrams. Just as in principal component analysis, the largest of the eigenvalues is associated with the eigenvector which defines the direction of the first canonical variable. Its value, in principal component analysis, is proportional to the fraction of total variance found in the direction of the first principal component. In canonical analysis, the direction of the first canonical variable is that with the largest ratio of good variance to bad, as signal to noise. The largest eigenvalue is proportional to this ratio.



CRAWFORD



CRAWFORD

8059-4F

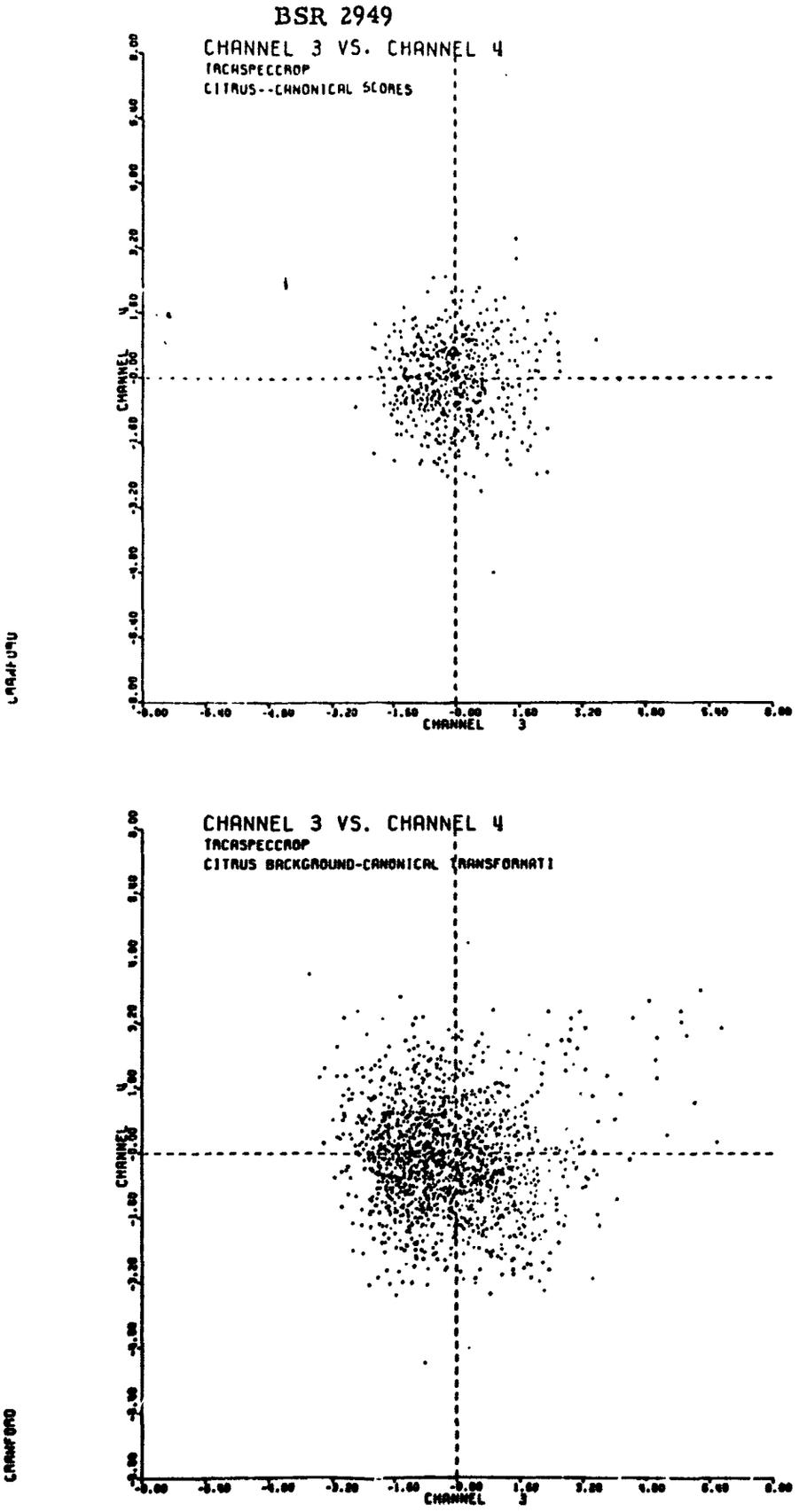
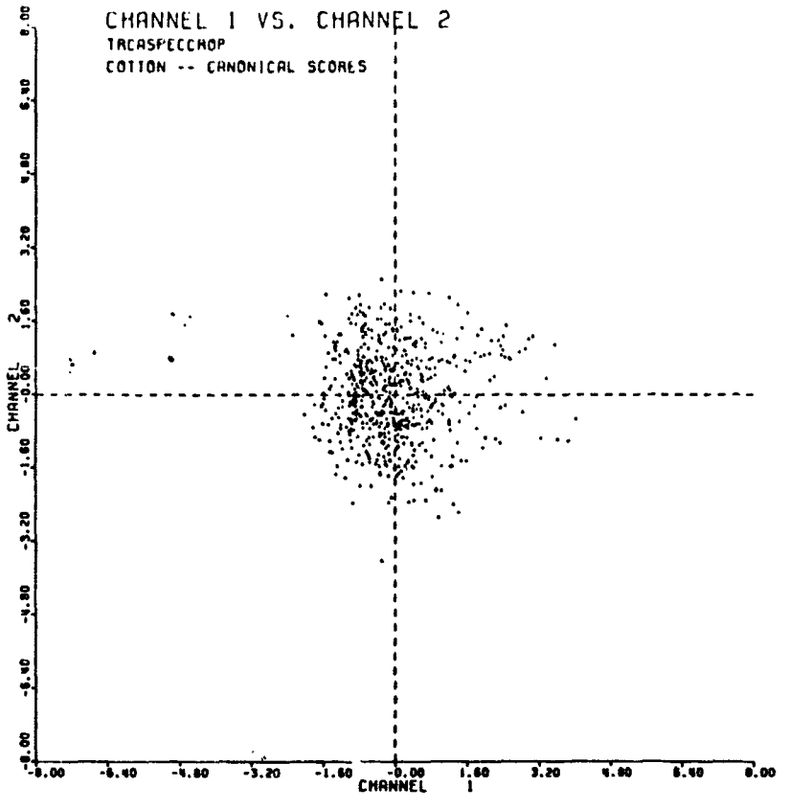
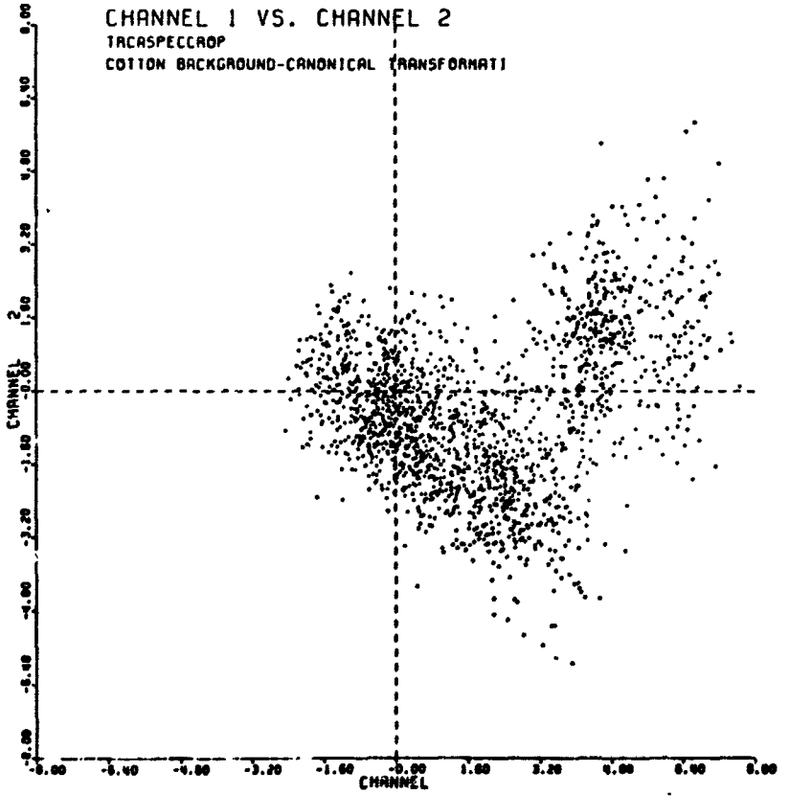


Figure 5-38 First Four Canonical Variables; Citrus



CRAWFORD



CRAWFORD

8080-47

BSR 2949

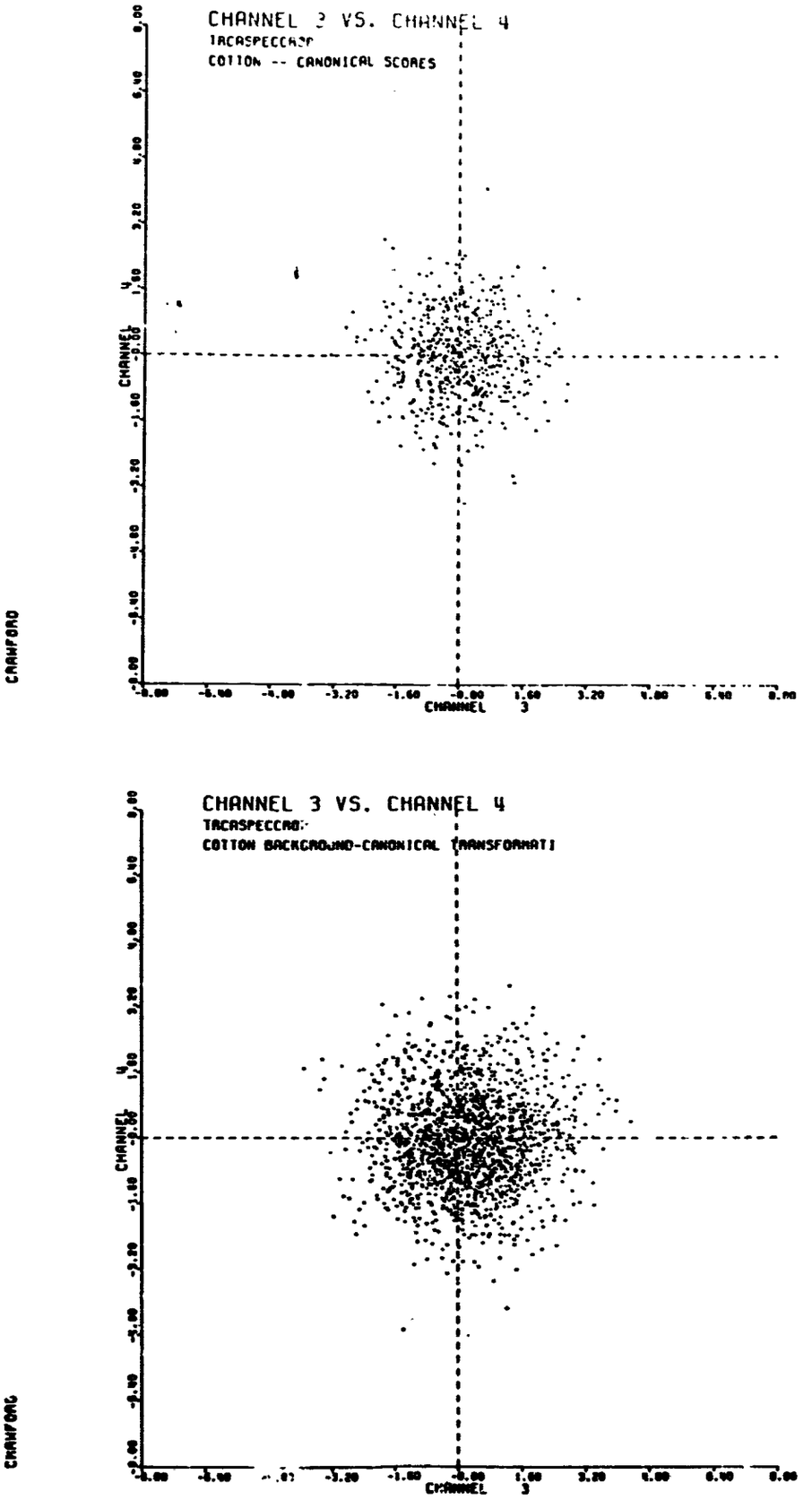
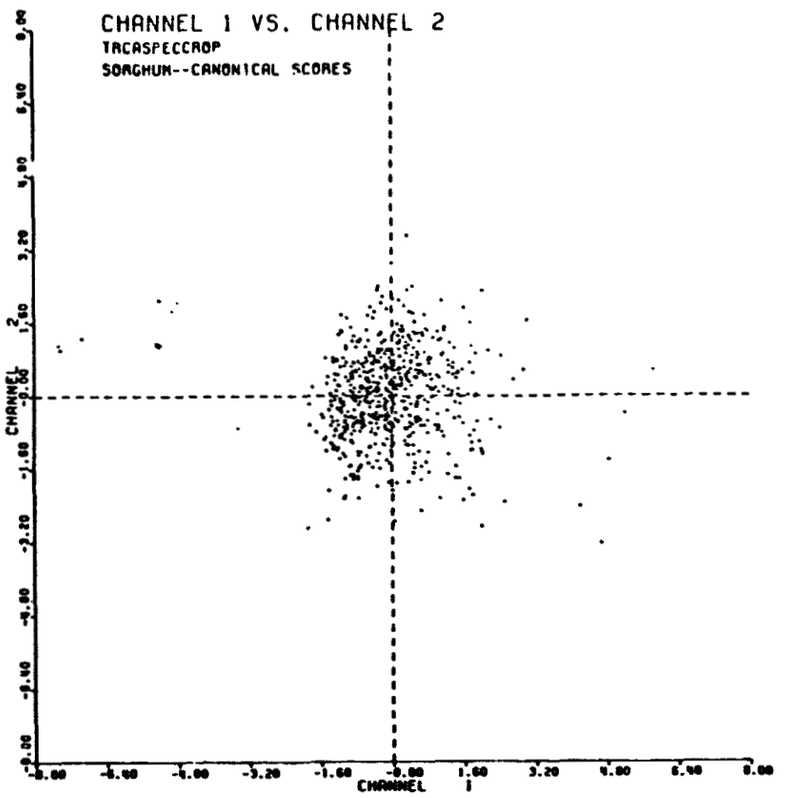
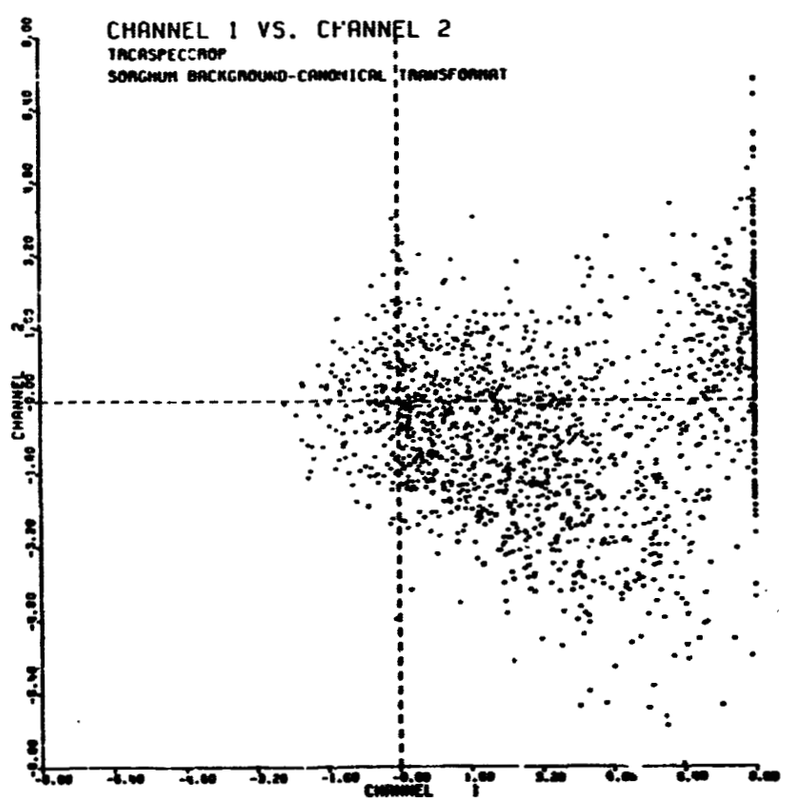


Figure 5-39 First Four Canonical Variables; Cotton



CRAWFORD



CRAWFORD

80060-46

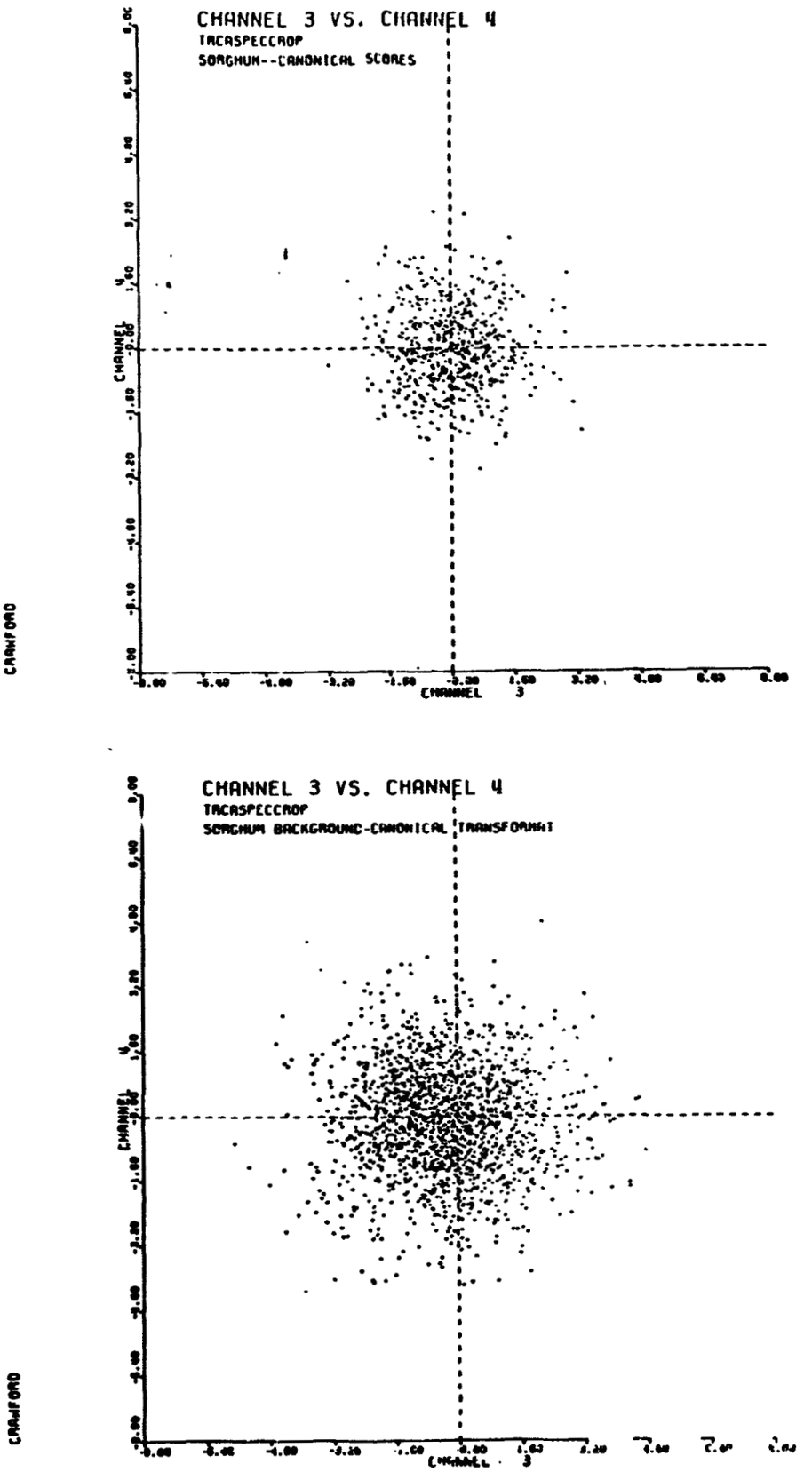
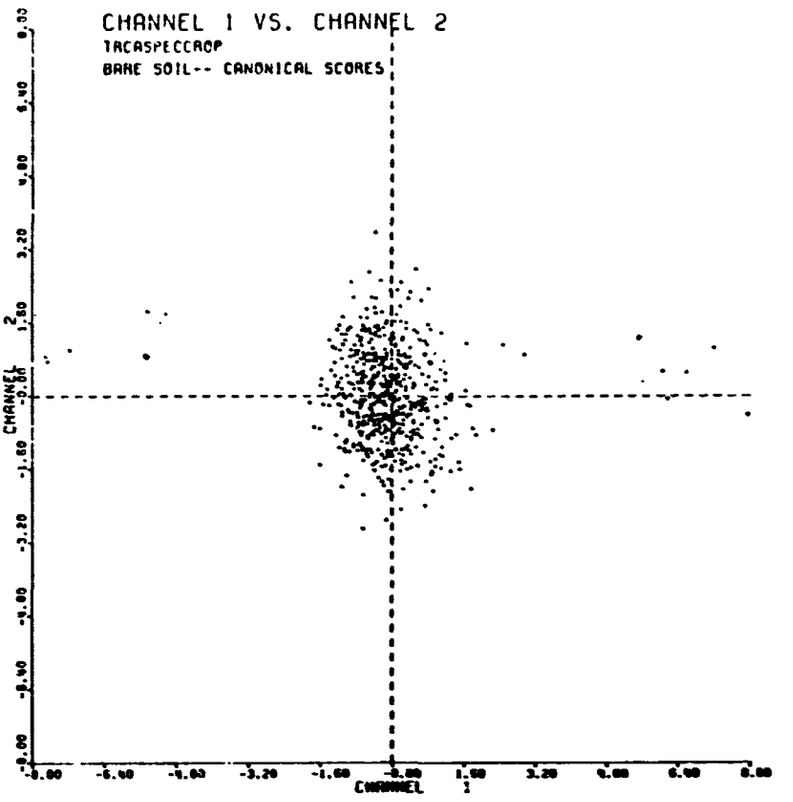
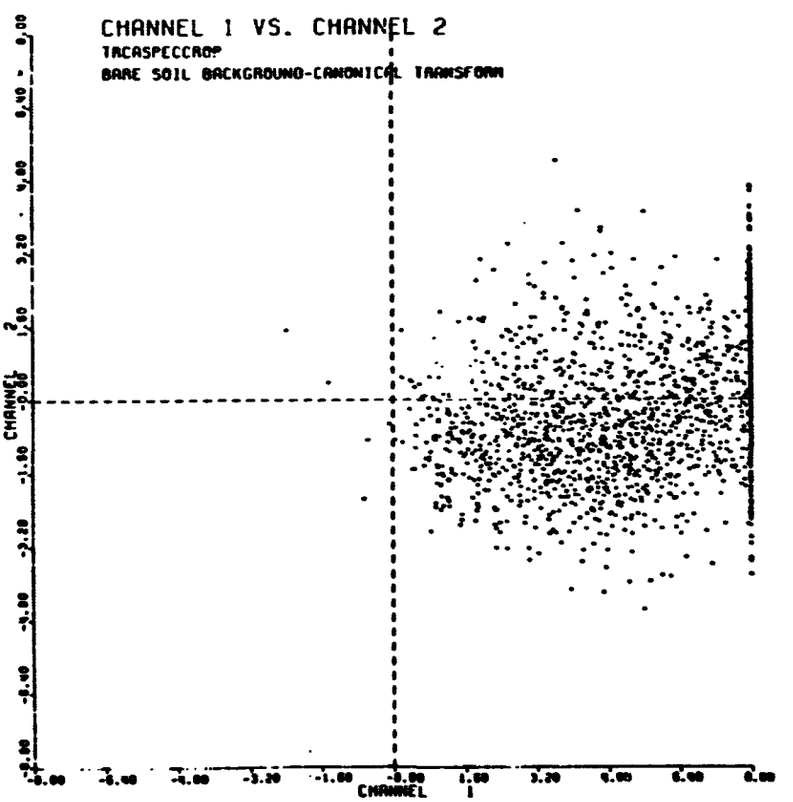


Figure 5-40 First Four Canonical Variables; Sorghum


CHANNEL 1 VS. CHANNEL 2
TACASPECCROP
BARE SOIL-- CANONICAL SCORES
CHANNEL 2
CHANNEL 1


CHANNEL 1 VS. CHANNEL 2
TACASPECCROP
BARE SOIL BACKGROUND-CANONICAL TRANSFORM
CHANNEL 2
CHANNEL 1

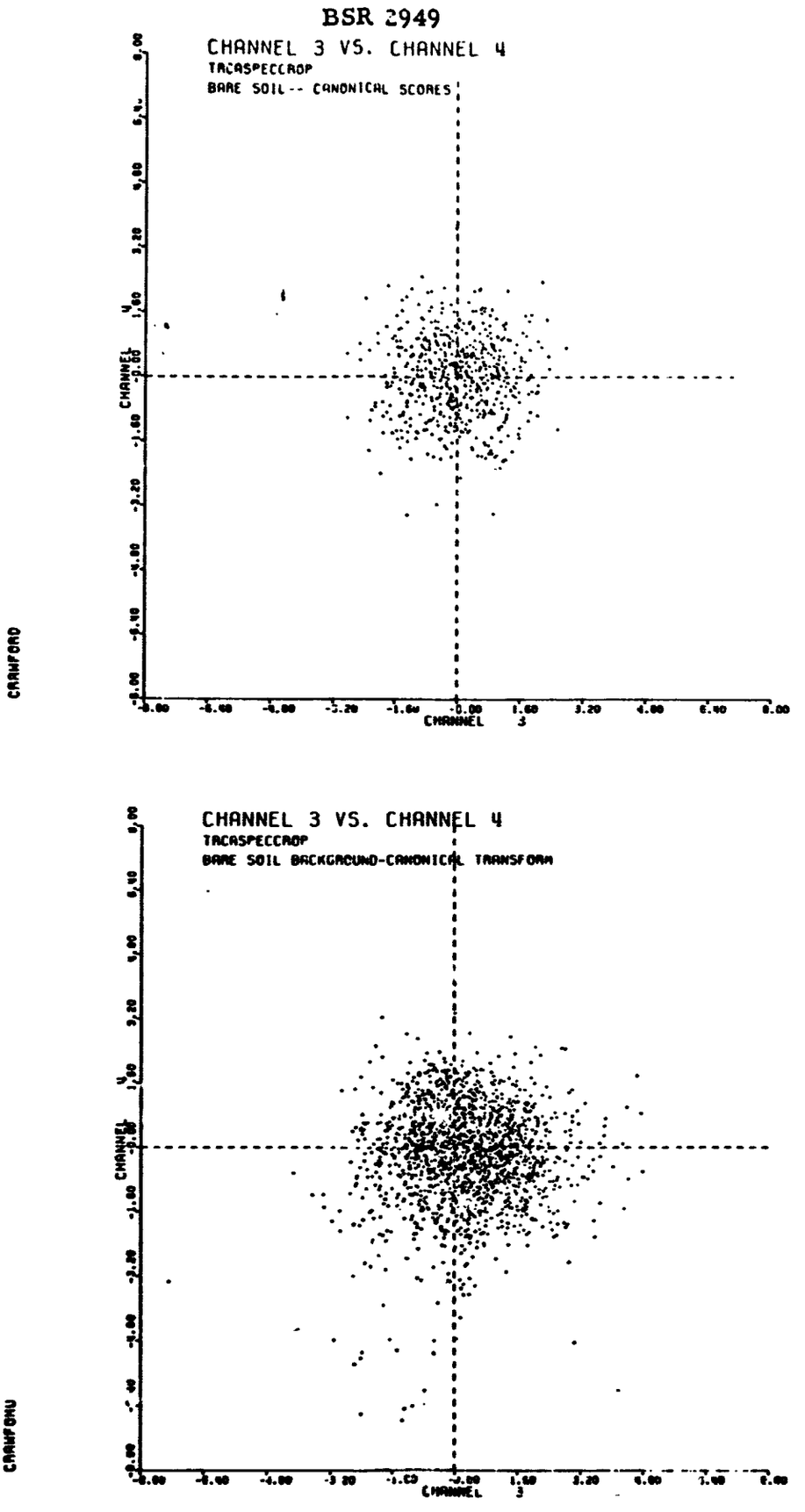


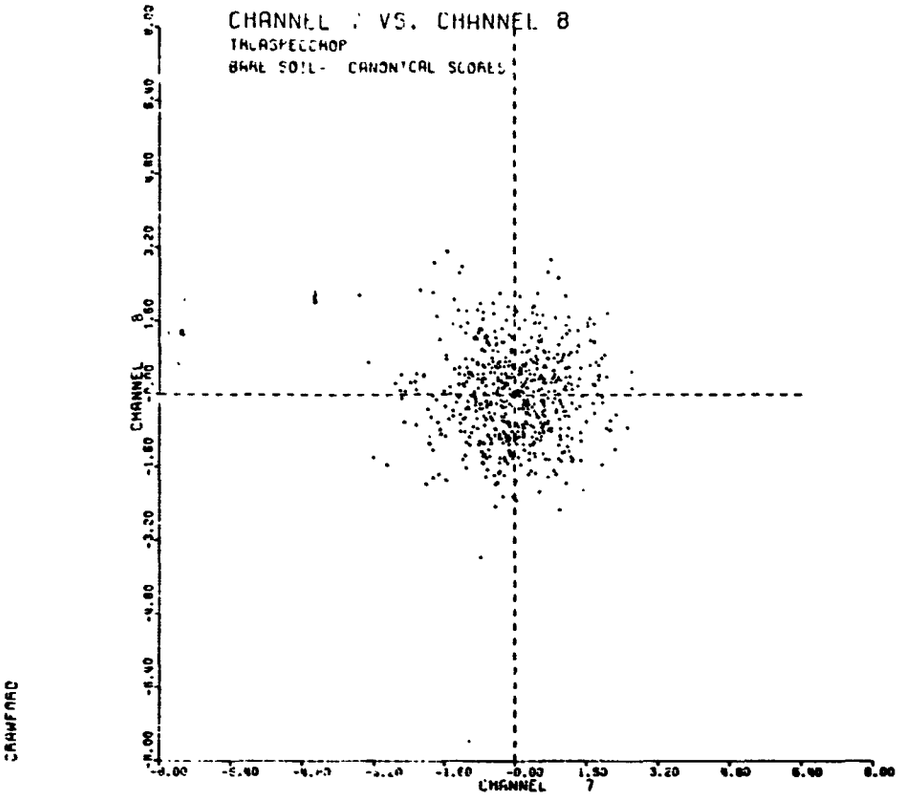
Figure 5-41 First Four Canonical Variables; Bare Soil

CRAWFORD

CRAWFORD

8080-80

BSR 2949



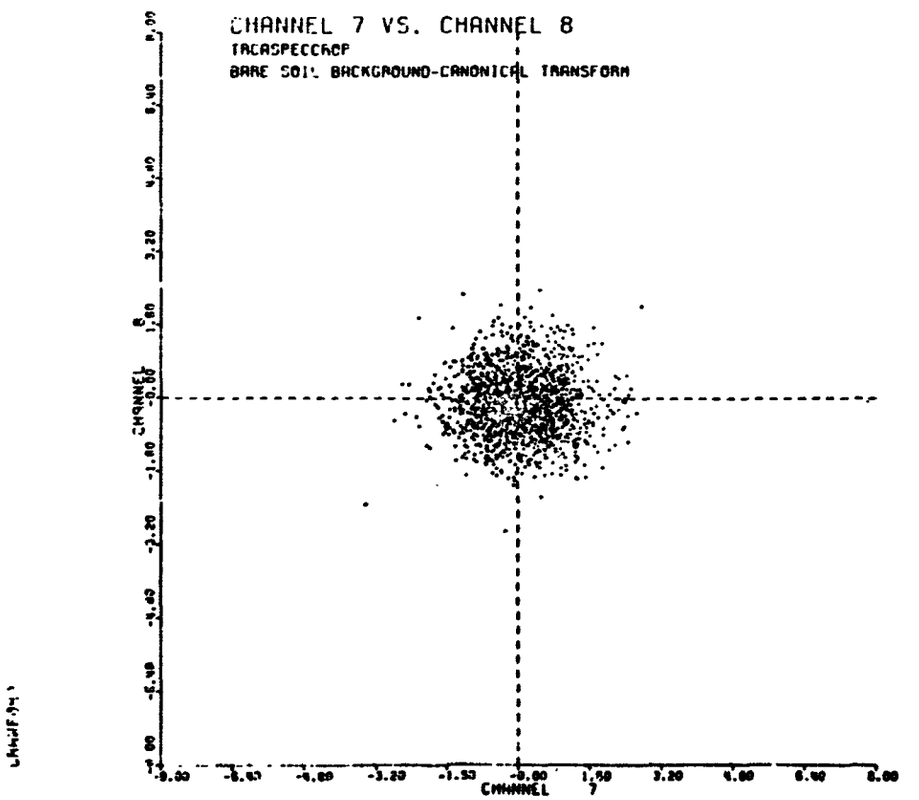


Figure 5-42 Last Four Canonical Variables; Bare Soil

For bare soil, the eigenvalues are listed in Table 5-14

TABLE 5-14

EIGENVALUES FOUND IN CANONICAL ANALYSIS WITH BARE SOIL AS THE TARGET CATEGORY

Eigenvalues	
1st	53.7
2nd	2.2
3rd	1.6
4th	1.4
5th	0.8
6th	0.7
7th	0.6
8th	0.5

The first eigenvalue suggests that the dispersion of bare soil background observations with respect to bare soil mean value is the square root of 54, or 7.3 times as large in the direction of the first canonical variable as that of bare soil observations about the same mean value in the same direction. The canonical variables one and two scattered in Figure 5-41 are consistent with this interpretation. Similarly, the dispersion of observations measured by the second variable appears about 1.5 (the square root of 2.2) times as large for background as for bare soil. Most surprising, the interpretation is consistent even in the scatter diagrams, Figure 5-42, of variables 7 and 8. The eigenvalues are less than one. The cluster of bare soil background, which has three times the number of bare soil observations, is clearly smaller than the target cluster.

This is too much of a good thing. The definition of the good variance is apparently not as good as desired. A lack of information content in the last canonical variables should be characterized by nearly identical distributions. A decision rule which would classify correctly a number greater than 25% of the bare soil observations could be based on variables 7 and 8 alone. The problem of controlling the disposition of information more carefully has only begun. Improved definitions of good and bad variance offer subject material for further studies of Signature Data Processing.

Table 5-15 gives the corresponding eigenvalues for the other three crops. The interpretation is similar although no scatter diagrams are presented for canonical variables beyond four.

TABLE 5-15

CANONICAL ANALYSIS EIGENVALUES

Crop / Eigenvalue	Citrus	Cotton	Sorghum	Bare Soil
1st	11.2	7.9	25.8	53.7
2nd	6.3	2.8	3.5	2.2
3rd	1.8	1.5	2.0	1.6
4th	1.6	1.3	1.8	1.4
5th	1.1	1.2	1.4	0.8
6th	1.0	1.0	1.3	0.7
7th	0.7	0.7	1.0	0.6
8th	0.5	0.6	0.6	0.5

The canonical variables which are the best candidates for deletion to accomplish dimensionality reduction are those with eigenvalues close to one. In all categories, the first four eigenvalues exceed one. The next four are scattered around one, about half above and half below. The highest numbered eigenvalues, as observed for bare soil, in each case tend to fall below one.

It is too early to draw all desired conclusions concerning the comparative utility of the canonical analysis transformation with respect to other linear transformations available. The requirement for an analytical approach to distinguish between noise-like variance and signal-like variance is conceptually satisfied. While it has not been possible to complete sufficient measures of the information lost by using the method for dimensionality reduction, it already appears as though still better control should be exercised over the disposition of information into the variables retained.

SECTION 6

HARDWARE DEVELOPMENT

A breadboard version of the analog table look-up device described in Section 3.2 was produced during the study to allow evaluation of the concept and identification of troublesome areas. A complete version of the analog table look-up device requires the assembly of several elements as indicated in Figure 6-1.

Analog computation of a single conditional probability density, $f(\vec{x}(t))$ is accomplished in three steps. The video input, $\vec{x}(t)$, is a time-varying multispectral observation of unknown origin. The observation is transformed by the analog transformation network into a space of principal components or canonical variables as described in Section 5.2.2 and 5.3. The transformation requires that mean values, $\underline{\vec{x}}$, stored in the network as DC voltage levels and describing the expected value of $\vec{x}(t)$ on the condition that it is an observation of the preferred category for the transformation applied, be subtracted from $\vec{x}(t)$ before operating with matrix A.

The matrix A is the transformation applied to the adjusted video. Since the matrix transformation is accomplished by forming specified linear combinations of the components of $(\vec{x}(t) - \underline{\vec{x}})$, analog circuitry consisting of summing and inverting amplifier with selectable gains is provided to form a special-purpose analog computer.

The output variables which result, be they principal components or canonical variables, will be uncorrelated, hopefully statistically independent, and have unit variance when the transformation is applied to the preferred category. The block diagram shows eight output variables resulting from eight input variables. The number of output variables can be reduced, perhaps to two, without loss of information if a sufficiently controlled canonical analysis is used for dimensionality reduction.

In this event the next operation, which is illustrated in Figure 6-1 indicating a requirement for four flying-spot scanners, could be accomplished with a single flying-spot scanner. In the more general case illustrated however, the general density for the transformed observation $\vec{y}(t)$ is computed by an analog multiplier in the form:

$$f(\vec{y}(t)) = f_1(y_1(t), y_2(t)) f_2(y_3(t), y_4(t)) f_3(----) f_4(----). \qquad (6\text{-}1)$$

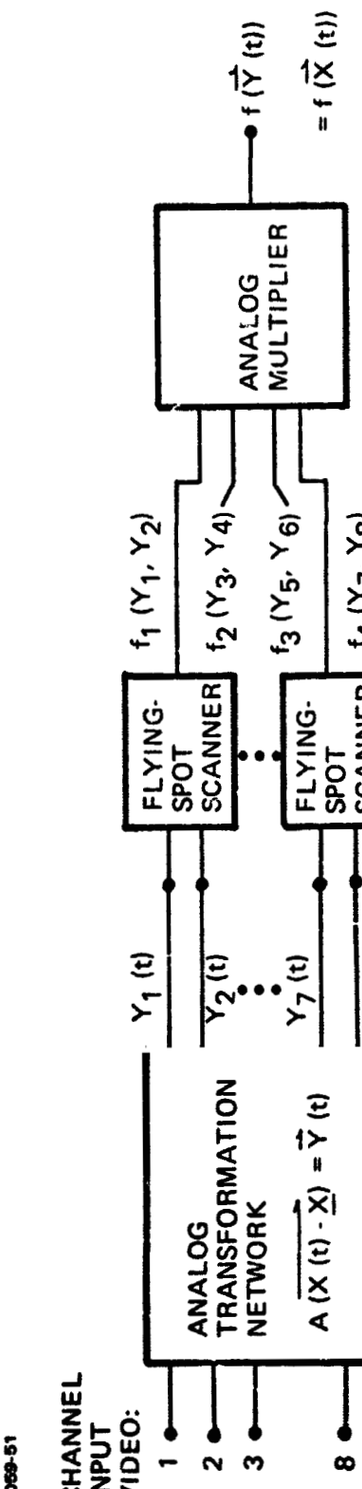


Figure 6-1 Block Diagram of Analog Table Look-Up Device

Since $\vec{y}(t) = A\vec{x}(t)$, where the matrix A specifies a non-singular transformation of $\vec{x}$, $\vec{y}(t)$ occurs with the same likelihood as $\vec{x}(t)$. Knowledge of the probability density for $\vec{y}(t)$ is equivalent to knowledge of the same for $\vec{x}(t)$.

The gain realized by performing the transformation is that of working in a coordinate system in which the expression for the density is factorable into a product of bivariate densities as in Equation 6-1.

The breadboard version of the device completed during the study includes the analog transformation network and a single, working, flying-spot scanner. Figures 6-2 and 6-3 are pictures of the completed breadboard units installed in the multispectral signature data processing laboratory. A laboratory tape playback machine provides multispectral video recorded during flight as input to the analog transformation network. Two stationary card readers are used for input coefficients to the transformation network. Each coefficient controls switches in an 8-bit ladder network of precision resistors; the total resistance of the network is controlled in this way to select the weighting for each channel of input for each linear combination.

A single pair of the transformed variables is then selected to deflect the beam of the flying-spot scanner. A functional schematic of the flying-spot scanner is illustrated in Figure 6-4. The output obtained is a bivariate conditional probability density function. It is conditional on the hypothesis that the input video $\vec{x}(t)$ has the same origin as the scatter diagram transparency scanned.

The analog output, when used to produce imagery, provides only a mapping of the bivariate conditional probability density for the target category defined by the transparency. It does not resemble imagery that would be produced by applying the maximum likelihood decision rule to several such conditional probability densities.

Examples of probability density imagery obtained in this way are presented following the discussion of the design, calibration, and performance of the breadboard unit.

6.1 HARDWARE DESCRIPTION

The transformation network (Figure 6-5) electronically forms eight linear combinations of the eight scanner channels. The output signals from the transformation network are then routed to an analog table look-up device which consists of cathode ray tubes (CRTs), computer-generated transparencies placed on the face of the CRTs, and photomultiplier tubes. The electron beams are deflected to various positions on the CRTs depending upon the pairs of signals selected to drive the horizontal and vertical axes. The photomultiplier tubes are then used to determine whether the electron beam is positioned in an opaque, transparent, or borderline region of the transparency.

REPRODUCIBILITY OF THE ORIGINAL PAGE IS POOR.

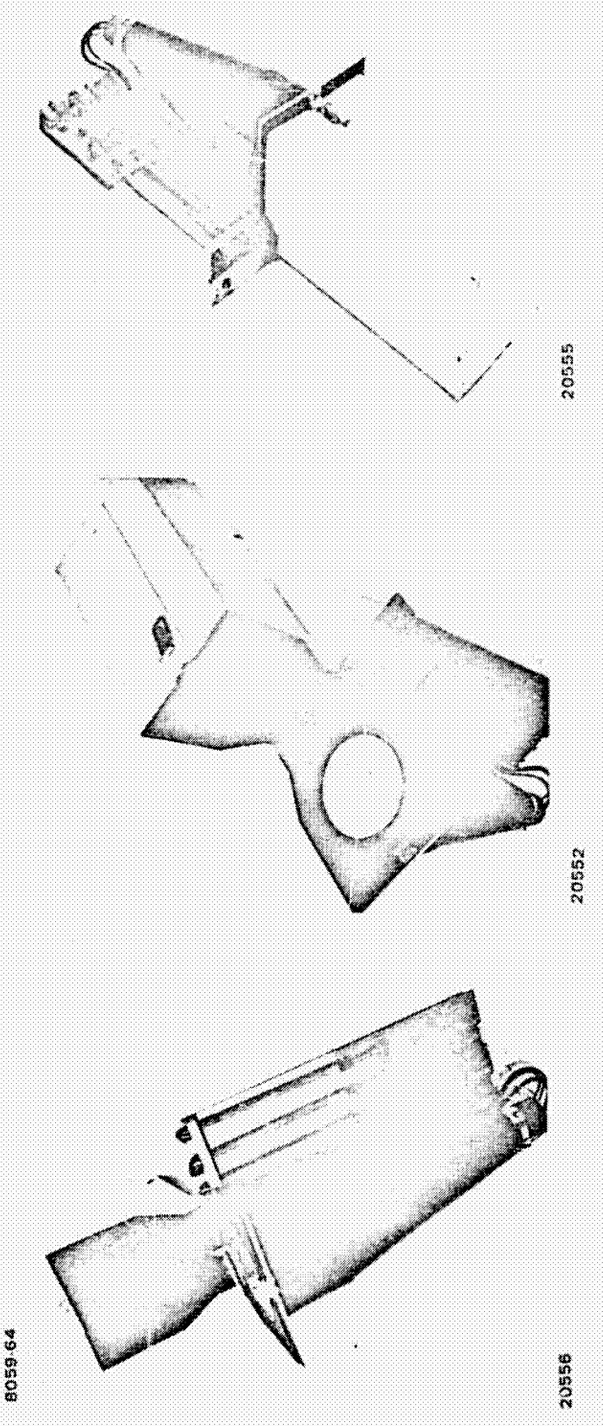


Figure 6-2 Three Views of Flying-Spot Scanner



20553

Figure 6-3 Front View of Transformation Network Installed and Flying-Spot Scanner

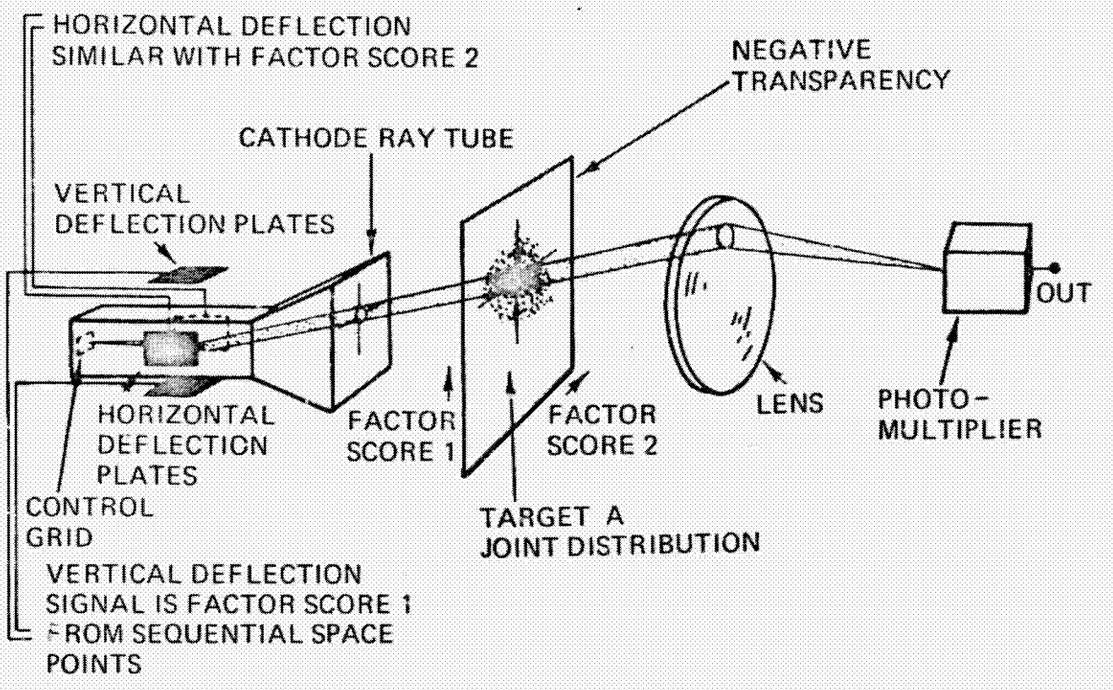


Figure 6-4 Functional Schematic of Analog Table Look-Up Device

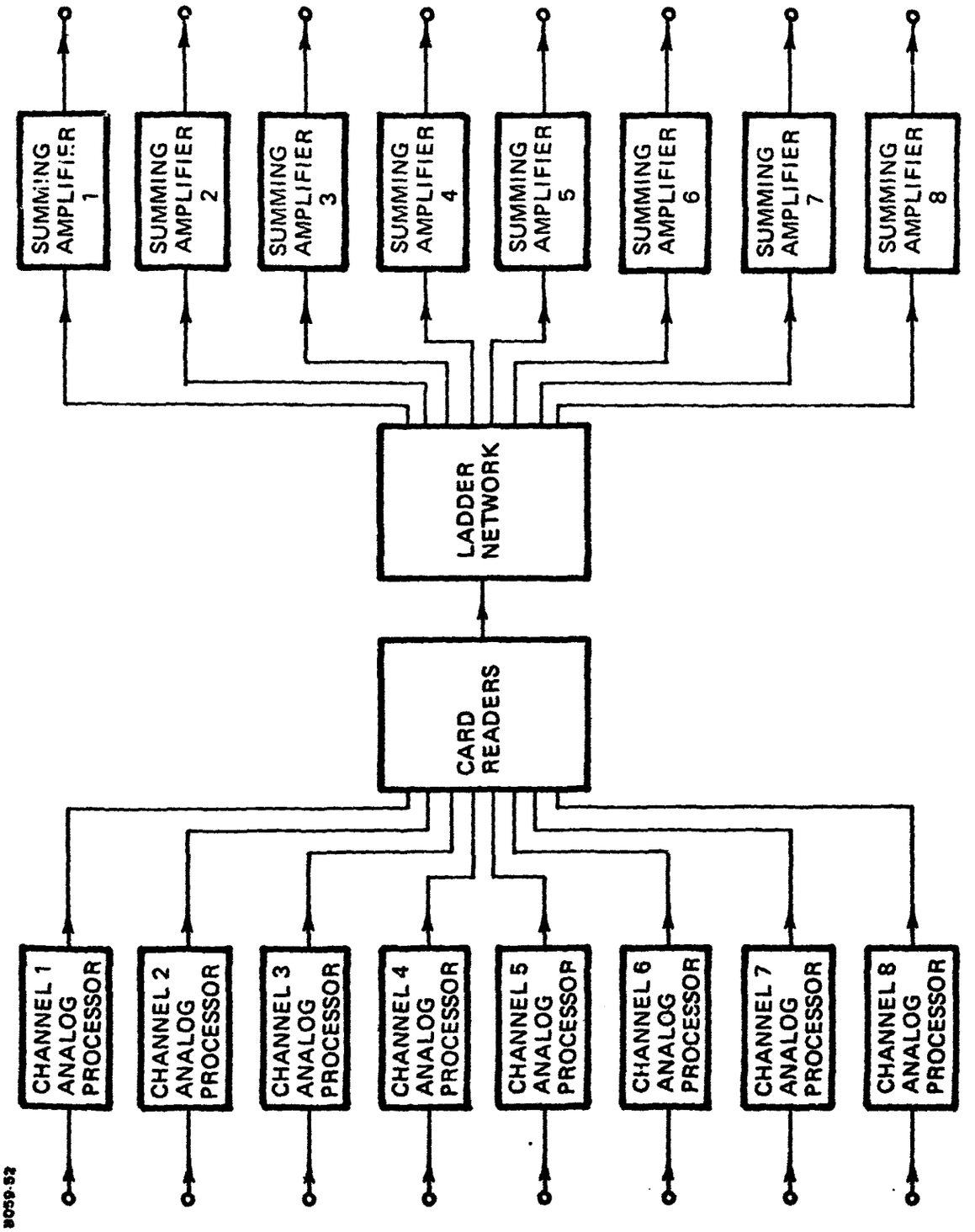


Figure 6-5 Transformation Network

For the following discussion see Figures 6-6 and 6-7. The tape recorder scanner channels are fed to the inputs of LM301 operational amplifiers (A1) through 10K input resistors (R_1). The gains of the analog processors are controlled by switching in various size feedback resistors (R_3) in the first stages. Emitter followers are placed inside the feedback loops to provide for sufficient drive capability. The non-inverting input of the operational amplifiers is supplied with DC voltages under the control of the card reader. The second stage of the analog processor consists of a unity gain inverting amplifier. It is identical to the first stage, with the exception of the fixed gain and zero DC offset. The output signals from the first and second stage are routed to the card reader. A card is then programmed to select the proper signs of the coefficients and also switches in eight-bit ladder networks which attenuate the input signals in multiples of 1/256.

The signals from the outputs of the ladder network are then fed to LM 301 operational amplifiers (A3) which are wired up in their inverting configuration and provide negative unity gains for each of their inputs. Thus, the output from a single amplifier is the negative sum of its eight inputs. The outputs from the summing amplifiers are then fed to limiters which inhibit the signal from exceeding a -2 v to +2 v dynamic range.

The outputs from the limiters are then paired up and fed to deflection amplifiers. These consist of single-ended input, balanced output differential amplifiers. The input is fed to the base of T_1 while a variable DC level is supplied to the base of T_2 to provide for a beam centering control. The emitter resistors (R_6 and R_7) are fed from a constant current source (T_3, R_9, R_{10}, and R_{11}). The variable resistor R_5 adjusts the voltage gain of the amplifier. The output from the collectors of T_1 and T_2 is then routed to deflection plates on a CRT.

The CRT electronics are straightforward, consisting of a resistor divider chain to properly bias the CRT grids. The CRT utilizes P16 phosphor which has a decay time (100% to 10% persistence) of 0.12 μsec.

The photomultiplier tube electronics are also straightforward, consisting of a string of 100-v zener diodes to properly bias the PM tube dynodes. The output signal is developed across an 18K load resistor (R_{19}).

A sine-cosine generator is formed by operational amplifiers A7 and A9. The sine and cosine outputs are followed by a buffer stage which provides for gain control by varying feedback resistors R_{22} and R_{28}. The outputs of these buffers are sent to a pair of deflection amplifiers which allow the deflection electronics to be calibrated.

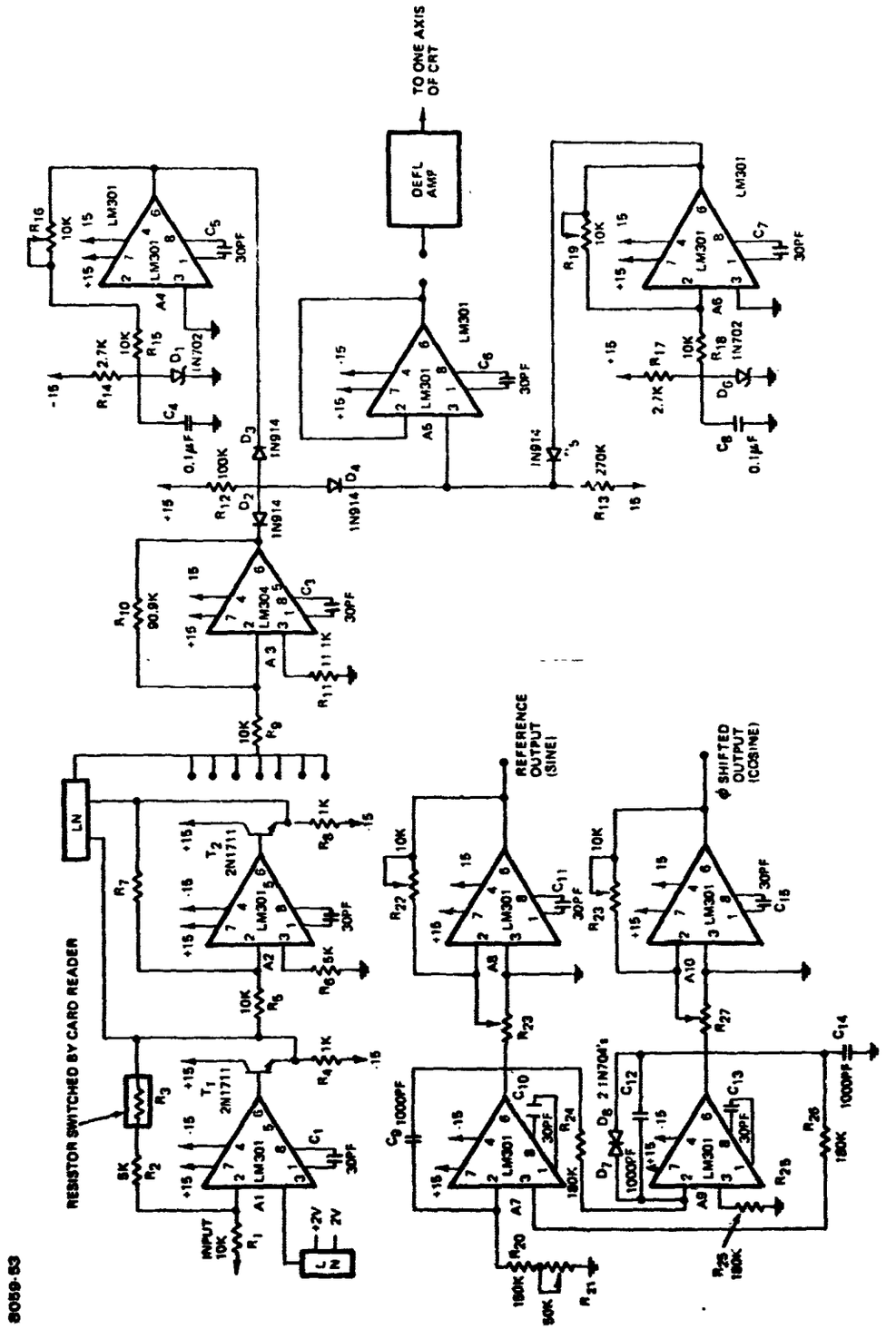


Figure 6-6 Analog Processor and Calibration Electronics

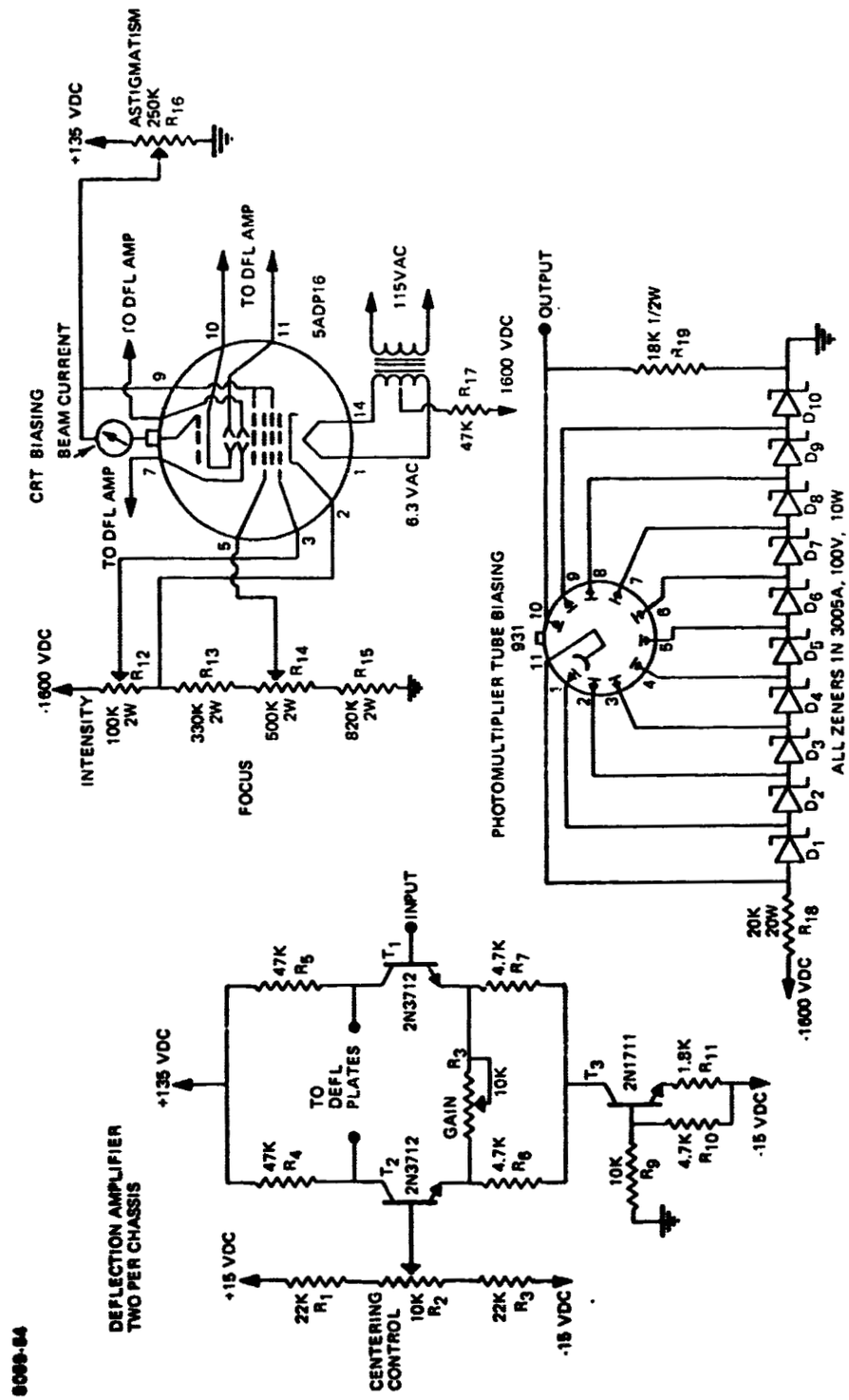


Figure 6-7 Flying-Spot Scanner Electronics

6.2 SYSTEM PERFORMANCE

To calibrate a single flying-spot scanner, several known voltages must be deflected to predetermined positions on the face of the CRT. To accomplish this, a test transparency consisting of concentric circles is placed over the face of the CRT to define the desired deflection positions. Initially, the horizontal and vertical inputs are grounded. The horizontal and vertical offsets are varied until the electron beam is positioned at the center of the concentric rings on the test transparency. The horizontal and vertical inputs are now connected to a sine-cosine generator. The peak signal from this generator is switchable to either 1, 2, or 3 v. For the 1-v position, the vertical and horizontal gains are adjusted such that the test trace and smallest concentric ring on the test transparency are coincident. The 2-v sine and cosine waveforms are then switched into the deflection amplifiers. The next largest concentric ring on the test transparency and the test trace should now be coincident. If not, the gains are readjusted slightly until coincidence is achieved. The 1-v peak waveforms are then switched back into the deflection amplifiers to ensure that the test trace and smallest circle are still coincident. While only one circle is actually required for calibration, a series of concentric rings also serves as a check on system linearity. The 3-v peak signals are used to set the dynamic range of the limiting circuitry. This guarantees that the electron beam deflection will not be larger than the physical dimensions of the transparencies and hence eliminates the possibility of erroneously detecting a transparent region of the transparency.

The system bandwidth was measured at 125 kHz. The temperature drift for the transformation network, deflection amplifiers, and CRT caused less than 1% change in both the horizontal and vertical displacements of the electron beam.

The variation in light output as a focussed, constant current electron beam is scanned across the entire face of the CRT and is less than 5%. To realize less than 5% variation, it was necessary to burn-in the CRT for 100 hr. Burn-in was accomplished using a raster-type beam deflection waveform, such that the beam was deflected to all phosphor particles. Beam current was kept at 25 microamperes during the entire burn-in period. If the electron beam is defocussed while keeping the beam current constant, an increase in brightness is observed.

Under normal operation the electron beam is in focus only in a small region of the CRT. Thus, additional variation in light output is introduced because the electron beam is deflected to regions where defocussing results. A third effect which produces an apparent brightness variation is the integrating effect of the photomultiplier tube. Light is also being emitted from the past locations of the electron beam; thus, the signal current from the photomultiplier is dependent upon the past history of the electron beam. The cumulative effect of the three mechanisms can produce an apparent light variation of as much as 40%.

6.2.1 Imagery

Probability density imagery has been produced as a first step towards fullblown use of the analog table look-up device to produce genuinely enhanced imagery. The example illustrated in Figure 6-8 is a map in which the darkest areas are those with a large value of citrus probability density. The area shown is a portion of flight line 11 as flown in June of 1969. Section 5.1 describes the data source in greater detail. Three strips of imagery are presented. The top strip is obtained by adding together all eight channels of video and film recording the result. The center strip is similarly produced, but is annotated to indicate the known identity of the crops growing in selected fields.

The last strip is probability density imagery; it is a bivariate smoothed density function produced by deflecting the flying-spot scanner beam with the first two principal components of citrus observations through a scatter diagram. The citrus scatter diagram used for this purpose is among those shown in Figure 6-9. It was prepared as described in Section 5.2.2 where it is also shown with similarly prepared diagrams from cotton, sorghum, and bare soil. A negative transparency of the 2000 citrus observations, scattered in the plane of principal components one and two, is the information storage medium for the general bivariate distribution.

The probability density imagery is a demonstration of the ability of the flying-spot scanner to recover and display or make available for further computation the information contained in a bivariate conditional probability density. The histograms presented in Section 5.2.2 are useful for interpretation of probability density imagery. They show the distribution of conditional probability density imagery computed from principal components. Figure 6-10 shows this distribution for the first four principal components of citrus. The rightmost values of $f(\vec{x} \mid \text{citrus})$ are those corresponding to a dark gray tone on the imagery. The histogram shows that values towards the right of the scale occur more frequently for citrus than for any of the other three categories shown; but the frequency with which non-citrus categories occur is usually larger than the citrus frequency. Thus, in Figure 6-8, it should be no surprise that non-citrus categories occasionally show up.

Histograms of the type introduced in Section 5 and shown for illustrative purposes in Figure 6-10 show that the dynamic range of probability density extends over several orders of magnitude. The result is not surprising, but it does not place a similar dynamic range requirement on the devices used in the laboratory.



Figure 6-8 Probability Density Imagery Produced From Analog Table Look-Up Device

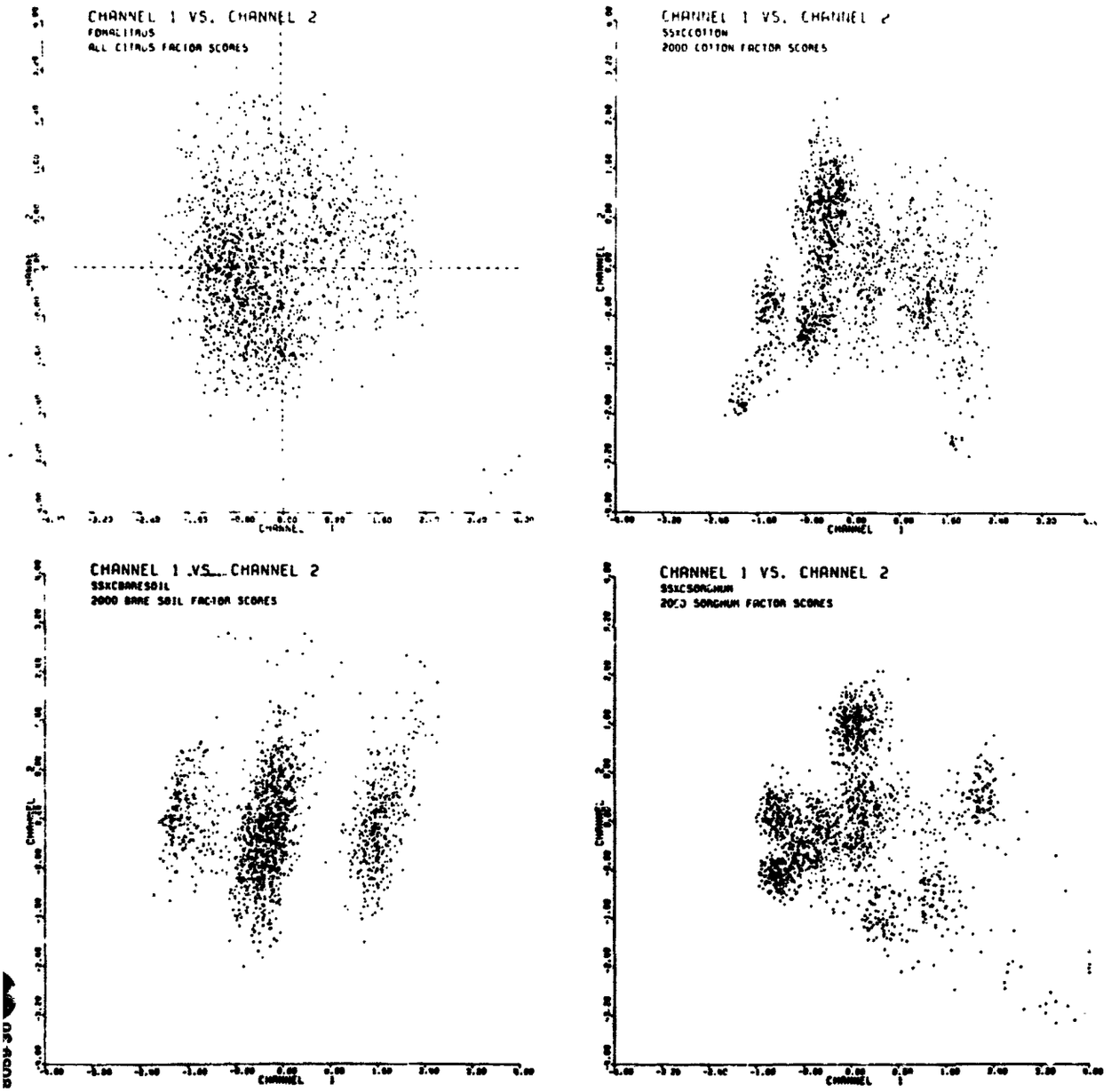


Figure 6-9 Scatter Diagrams 2000 Factor Scores of Four Crops 1 vs. 2

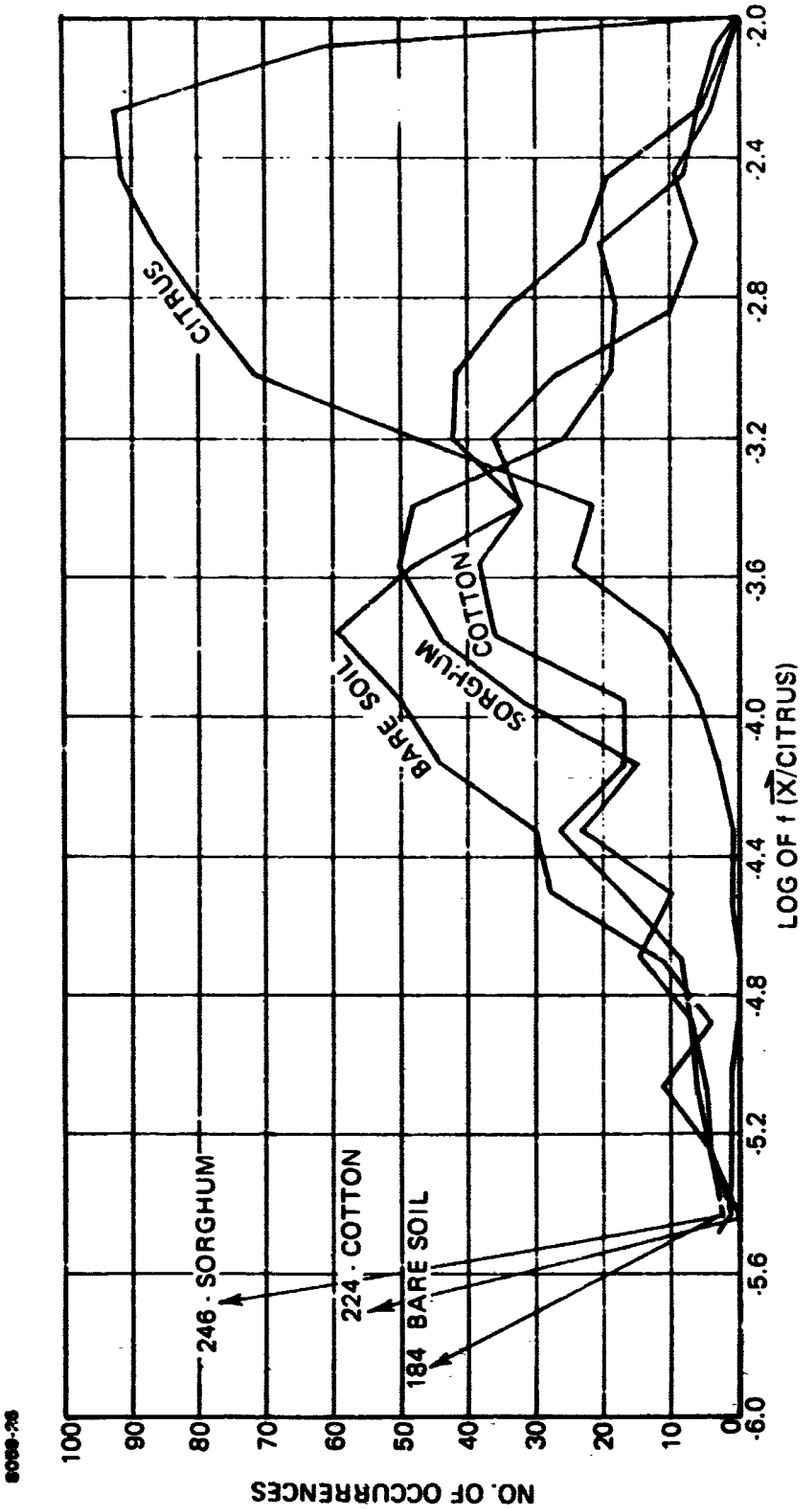


Figure 6-10 Distribution of f(x/citrus) for Samples of N = 600

The uncertainties in some of the parameters used in classification procedures may well exceed in magnitude the noise introduced by the hardware. The dynamic range available in photomultipliers and in film density is safely in excess of that available in the CRT required for the ATLUD. In the final analysis, the upper limit to the dynamic range required for usch a system is that which an interpreter can make use of when looking at imagery. While the CRT has less range than probability density, its output range is reasonably well matched to the range used by human observers.

Notice also that the top strip of imagery shows a number of cloud shadows on the ground. Near the left end of the strip a road cuts diagonally across the imagery. It is interesting to compare the cloud patterns in the top strip, the field identification in the center strip (where the same clouds show up incidentally), and the conditional probability density mapping of citrus in the bottom strip.

The density mapped in the imagery is bivariate in the first two principal components instead of tetravariate in the first four as analyzed in the histograms presented earlier. A histogram prepared from the bivariate density for citrus is presented in Figure 6-11.

6.2.2 Smoothing

The smoothing effect of the flying-spot scanner has been directly evaluated by using it with a conventional raster scan to produce video, and film recording the video. By raster scanning a scatter diagram with the CRT beam in a focussed and unfocussed condition, the amount of defocuss required to produce a smoothly variable probability density which uses the full range of gray scale dynamic range available can be determined. The scatter diagram containing 2000 sorghum observations was selected for the evaluation because of the multiple clusters to be smoothed, yet resolved by the flying-spot scanner. The scatter diagram from which the negative transparency was produced is also shown in Figure 6-10. Figure 6-12 shows prints made from the imagery exposed by raster scanning the transparency. The upper picture is badly undersmoothed. It was produced by setting the CRT focus adjustment at the position of best focus.

The lower picture looks like a satisfactory, smooth representation of a population probability density which could have yielded the sample used to prepare the scatter diagram. The focus position used to produce this result was also used to produce the probability density imagery of Figure 6-8.

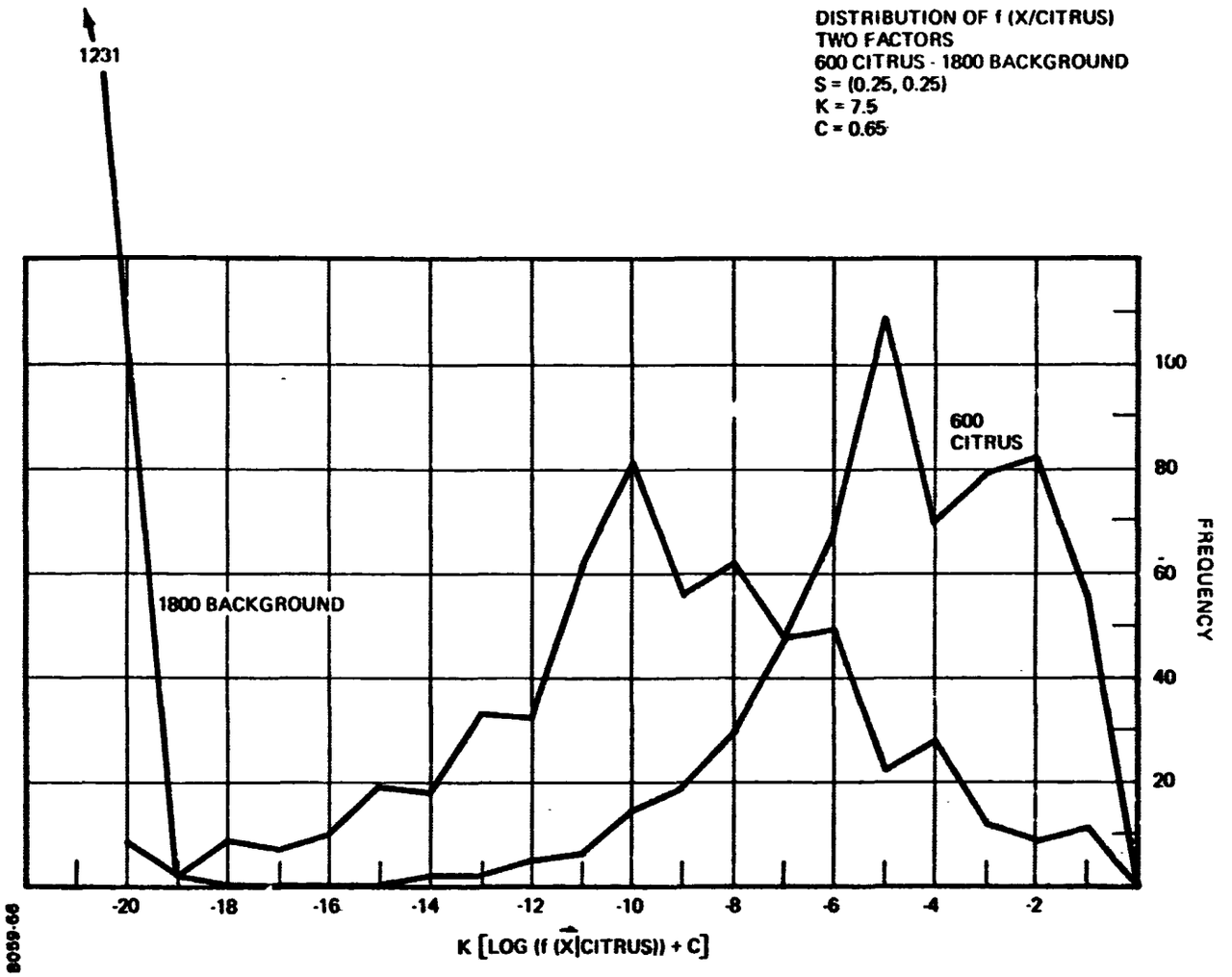
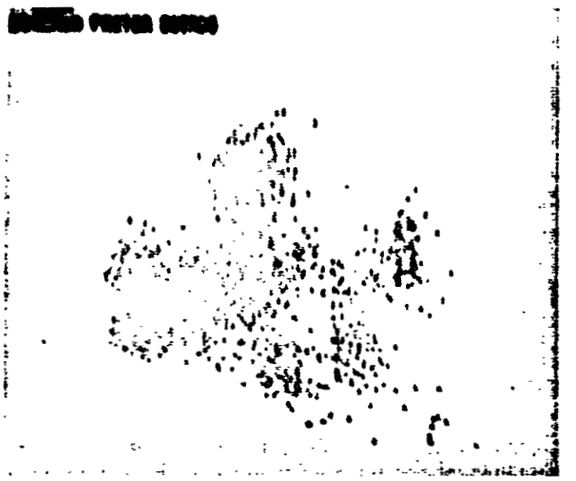


Figure 6-11 Histogram Prepared From Bivariate Density for Citrus



(a) Best Focus



(b) Best Smoothing Focus

Figure 6-12 Smoothing Effect of Flying-Spot Scanner

Smoothing parameters employed to make imagery with this device are most conveniently obtained when the desired smoothing parameters are equal for the two variables. The role of the computer analysis of smoothing parameter selection is to determine the best selection for the smoothing parameters without such a restriction. Compensation to accommodate the hardware can be performed by applying the appropriate metric scaling to the variables as necessary in the transformation network.

SECTION 7

CONCLUSIONS

Three types of conclusions are drawn from the study results and inferences based on experience during the study. Type one conclusions are those regarding the theoretical and analytical methods which influence the ability to accomplish Signature Data Processing; these are discussed under the heading Analytical Problems in Section 7.1. Type two conclusions concern the description of a research tool to accelerate the rate at which solutions to analytical problems can be identified and evaluated. This tool, described as a Signature Data Processing System (SDPS) in Section 7.2, is a general-purpose multivariate processor functionally specified to store large amounts of information and apply analytical techniques at high data throughput rates. Simultaneous user judgments and human factor considerations are possible by making use of real data rate display devices already provided in other equipment that will be available. Functionally oriented engineering specifications for this system are presented in Volume III.

Type three conclusions are drawn from experience gained during the study with the breadboard version of a portion of the recommended SDPS. Critical design properties of the flying spot scanners which comprise the Analog Table Look-Up Device are identified. Some of the options that should be preserved for comparative evaluation and to allow for modular expansion of system capability are identified.

7.1 ANALYTICAL PROBLEMS

A non-parametric representation of a general multivariate probability density is recommended on the postulate that no representation is superior to that obtained by preparing a multivariate histogram from a sufficiently large sample drawn from the population of interest. For reasons discussed in Section 2 (limited information storage capacity of computers), a sample is stored instead of the density function. The computational scheme of smoothing the sample histogram, discussed in Section 4.3, is employed to evaluate the density function at any desired point.

The desirability of a procedure for obtaining reasonable and objective representations of population probability densities by direct methods requiring neither operator intervention nor iterative trial-and-error methods is self evident.

It is concluded that the development of such a procedure requires a direct method of choosing smoothing parameters. During this study, trial-and-error methods have been used to select smoothing parameters that seem to give plausible

results. While the computation of probability densities using the smoothing methods — even the relatively fast hyperpyramid version — is itself a time-consuming procedure, the time is well spent, provided that reasonable results are obtained. But without an objective method for selecting comparable smoothing parameters for all distributions to be compared, several attempts are required to obtain reasonable results. Because the criterion of reasonableness is subjective, there is no guarantee that optimal parameters are selected for a single point distribution or that distributions to be compared are provided with density functions which are comparably scaled.

The use of a smoothing method is currently feasible only for relatively small samples. This is not so much because of the long time required to compute densities for each point in the sample once, but because of the necessity for doing it repetitively. For example, with a training sample of 600 observations, four channels each, the computation of probability densities for 2400 unknown observations requires about 15 minutes when using the hyperpyramid smoothing. Running time increases as the following proportionality:

$$T_{run} \propto N_T * N_S * M \qquad (7\text{-}1)$$

where

N_T = no. of training samples

N_S = no. of unknown samples, i.e., the number of points at which the density is evaluated

M = no. of channels per observation.

For maximum utility as an evaluation standard, much larger training samples will be necessary with a corresponding increase in computational time. These times will remain prohibitive for use on data presented in an image format; sampling of image format data will always be required. The use of satisfactorily large samples (a number which must be determined) will require direct methods for determining optimal smoothing parameters.

The techniques which produce optimal, i.e., maximum likelihood, solutions to multivariate classification problems are well known and straightforward when a proper method of evaluating conditional probability densities is available. The first requirement to make such a method available is the determination of how to choose, directly, smoothing parameters. This is a key step.

Given the capability to choose, directly, optimal smoothing parameters, computational time is further reduced to a minimum by dimensionality reduction. Dimensionality reduction is conducted to place the useful information in the variables retained. Linear transformations of the canonical type discussed in Section 5.3 show promise of providing the control necessary to distinguish noise-like variance from information-containing variance. With properly chosen definitions of "good" and "bad" variance (see Section 5.3, Canonical Analysis), linear transformations can be found which concentrate information in a subset of the transformed variables while assigning the noise-like variance to variables which can be deleted from the problem.

By applying the method of information divergence, the utility of transformations and definitions of "good" and "bad" variance can be measured. Proper application of the divergence method again requires the ability of direct selection of optimal sets of smoothing parameters.

The process of dimensionality reduction permits two benefits, both of which reduce the computational time:

1. Computational time is reduced by reducing the number of variables.

2. Computational time is reduced by permitting the use of smaller training samples.

Both benefits are indicated in the proportionality stated by Expression 7-1.

The second benefit accompanies the first. The number of samples required to represent the distribution of a population increases exponentially with the number of variables retained. The precise judgment of the sample size providing an adequate representation calls for the definition of terms and criteria. Such precision is unnecessary to recognize the impact of the proportionality on the sample size requirement. In a space of two channels, with 100 resolution elements on each channel, 10,000 cells are available for each observation. An adequate sample size does not necessarily require that all cells be occupied; but among the cells which are occupied, many cells should be occupied by more than one observation. If a third variable is added to a sample of adequate size when observed with two variables per observation, the number of occupied cells increases more slowly than the number of cells available. One million cells instead of 10,000 become available. It is reasonable to assume that roughly 100 times as many samples will be necessary to achieve a similar percentage of cell occupancy in the space of higher dimensionality.

It is desirable, therefore, to reduce the dimensionality to the smallest possible number. Dimensionality reduction is easily obtained by using only a portion of the data channels; information is eliminated at the same time. A more satisfactory approach to dimensionality reduction is obtained by performing a linear transformation of the canonical analysis type.

A twofold reduction in computational time is obtained by applying a linear transformation to obtain dimensionality reduction without a corresponding reduction of information content. Computational time is reduced by reducing: (1) the number of variables and (2) the size of a sample required to represent the population adequately.

It is concluded, therefore, that the priorities in the development of analytical techniques for multivariate signature data processing are:

1. Methods for direct selection of smoothing parameters

2. Study of the methods of canonical analysis for optimum dimensionality reduction.

Section 7.2 describes an SDPS. The system is configured to aid in the development and evaluation of the priority analytical techniques. It is furthermore a high-speed signature recognition device capable of computing conditional probability density functions at real-time rates.

The density function computations are performed in an analog manner, and the computation is the analog of the smoothing technique which requires the selection of smoothing parameters. With analog computation to assist the investigator, many of the time-consuming and expensive aspects of the analytical problems become less serious.

7.2 SIGNATURE DATA PROCESSING SYSTEM

It is concluded from the work conducted during the study that an SDPS which satisfies the study objectives defined in Section 2 will serve not only as a system to apply decision theory, but also as an important tool to assist in the solution of the analytical problems identified in Section 7.1. This section is a discussion of the utility of the recommended SDPS. Detailed functional specifications are presented in Volume III.

A block diagram of the subject system is shown in Figure 7-1. The system is modular and, as shown, interfaces with the Ground Data Station (GDS) of the Multispectral Data System (MSDS) under construction for NASA/MSC (Contract NAS9-9089). In this way, the initial configuration of the SDPS will provide only for performing the data processing operations. The tape input equipment, a control computer, and the display and output devices already available in the GDS are used also by the SDPS. System elements to be provided initially by the GDS are so annotated in the block diagram. The modularity of the system permits the selective addition of the shared system elements to the SDPS so that it may be built into a self-contained system.

Only those elements of the SDPS which operate on the data to extract information are discussed here. Included are the Transformation Network, the Multiple Threshold Gates (MTGs), the Analog Table Look-Up Device (ATLUD), and the Decision Network.

The Transformation Network applies a linear transformation of the type specified by principal component analysis or canonical analysis to the 24 channels of input data, or to a subset thereof, provided by the MSDS-GDS. Two purposes are served by properly selected transformations of this type:

1. The variables are transformed to a coordinate system in which they are more nearly statistically independent. This allows the probability density function to be factored into a product of density functions.

2. Information from all original variables is concentrated in a subset of the transformed variables to permit reduction of dimensionality with minimal loss of information.

Both of these benefits are applied to advantage in the processing operations performed after the transformation is completed. A decision boundary test is applied by the MTGs and a maximum likelihood test is applied by the ATLUD and the Decision Network in series.

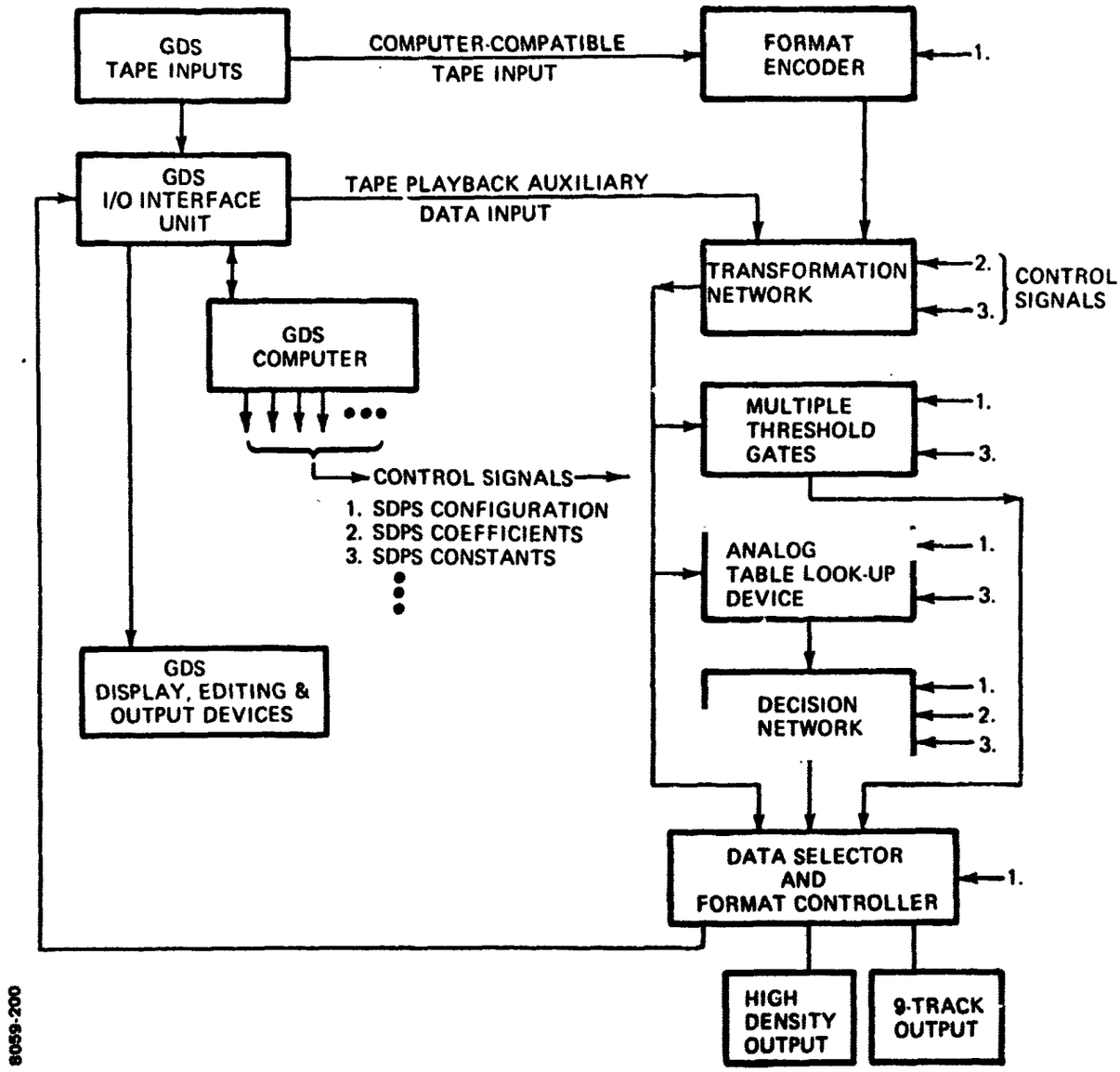


Figure 7-1 Functional Block Diagram of SDPS

A Decision Boundary Mode is one of the three operating modes specified in Volume III for the SDPS. The test involved is simple and fast but quite effective nonetheless. Each of the variables retained after reduction of dimensionality is tested by the MTG to determine whether it falls within a predetermined range which characterizes the variable in question when it is obtained from an observation of a specific category. Such a range for each of the variables retained defines a hypersurface in the form of a hyperparallepiped. This hypersurface is a decision boundary, which defines approximately the regions R_i and C_i as discussed in Section 2. An observation falling within such a hypersurface is assigned with a maximum likelihood to the category corresponding to the decision boundary.

The merits of such a processing scheme are that the decision boundaries are simply and quickly defined by inspection of scatter diagrams such as those presented in Section 5.3. A minimal information storage problem is presented; only the upper and lower limits of each variable for each category must be retained. The operation can be performed rapidly with either digital or analog methods. Digitally time-consuming and electronically touchy analog arithmetic operations are unnecessary; logical operations alone suffice. The output products of such a decision-making operation are logical variables easily encoded for computer-compatible tape recording at a density of eight categories per character. For simultaneous display of all categories identified, a color display and film recorder offers the necessary degrees of freedom. These are provided by the MSDS-GDS.

On the negative side, it should be noted that precisely correct decision boundaries, e.g., the R_i and C_i of Section 2, are not hyperparallelepipeds. For simple problems, little harm is done by approximating the C_i, which are not exhaustive or necessarily mutually exclusive, by hyperparallelepipeds. But decision boundaries defined in this way coincide only approximately with the boundaries of the regions specified in the theorems of Section 2.

Furthermore, classification achieved in this way does not present maximum information in the display prepared. The information that a particular decision is the maximum likelihood does not speak of the relative likelihood of other decisions. The spatial information contained within a spatial region assigned to the same category is totally suppressed. Not all maximum likelihood decisions have equal likelihood. The conditional probability of correctness can vary considerably. Mapping such decisions preserves spatial information. Spatial information is difficult to process by machine, but relatively easy for a user to process by inspection and experience. To suppress this information denies the user access to information readily available before spectral processing. For some problems, a disservice is performed in this way.

Another operating mode—the Maximum Likelihood Mode - is available for use in problems where the decision boundary approach is inadequate or undesirable. The additional information presented in the display which results, and the additional precision with which a maximum likelihood region is defined, are purchased at the expense of additional complexity in the hardware which accomplishes the processing. This mode uses the ATLUD and the Decision Network serially to precess the variables presented by the Transformation Network. The video signal from the Transformation Network serves as input for several sets of flying spot scanners, conceptually operating as described in Section 6, for the computation of several conditional probability densities. Each of the probability densities is the likelihood that the observation which produced it belongs to one of the categories being considered. The probability densities, which in the most general case can be handled by the ATLUD, are computed as the product of bivariate probability densities. Thus, a tetravariate probability density is computed as the product of two bivariate densities. One flying spot scanner is used for each bivariate density.

The output of the ATLUD is a complete set of conditional probability densities. The operation performed by the ATLUD may be regarded as a nonlinear singular transformation of a set of input variables into a set of conditional probability densities.

Maximum likelihood decisions, and variants thereupon as described in Theorems 1, 2, and 3 of Section 2.2, are based on these densities. Decisions which can be expressed as logical variables just as in the Decision Boundary Mode are achieved by assigning each observation to the category with the largest conditional probability density. This is, clearly enough, the maximum likelihood decision as defined by Theorem 1. It is superior to that obtained by the Decision Boundary Method because the shape of the hypersurfaces which define the decision boundaries is no longer restricted to hyperparallelepipeds.

One of the negative aspects of the Decision Boundary Mode of operation is countered in this way. Regions which are nonexhaustive, to limit the probability of erroneous classification in the manner of Theorem 2 (Section 2.2), can also be achieved. To do so requires the use of the Decision Network.

The other negative aspect of the Decision Boundary Mode of operation was the restriction of output variables to logical ones which supress much of the spatial information. Among the output options when operating in the Maximum Likelihood Mode is providing the probability densities as final output. Maximum likelihood decisions can in this way be displayed with minimal loss of spatial information. To accomplish this using the color display devices of the MSDS-GDS, it is only necessary to associate each category for which a likelihood is computed with a separate color. The color displayed is that of the category with the greatest

likelihood; at constant hue, the saturation is then modulated with the value of the likelihood (i. e., conditional probability density) of the maximum likelihood category. The Decision Network plays no essential role in this procedure.

The use of ATLUDs for computing probability densities allows the operations digitally performed by the smoothing procedures to be performed in an analog manner. The selection of smoothing parameters for analog computations can be accomplished even by trial-and-error procedures without incurring the expense of doing so with digital computation. The criterion of reasonableness is more easily evaluated if complete bivariate densities with gray scale as the third dimension are prepared. The analog signal representing probability density is played into a film recorder to accomplish this. Examples are shown in Figure 6-12. The problem of selecting smoothing parameters is simplified in three ways with the analog computation. First, the sample size can be made very large without increasing the time and expense required to perform the computation. Second, with large samples the selection of smoothing parameters becomes less difficult. Third, the smoothing parameter is expressed in the analog computation as a CRT spot size. Defocussing of the beam increases the smoothing parameter while smoothing can be virtually eliminated by focussing well. A variable smoothing method might be explored whereby the smoothing selected would depend upon the local spot density of the bivariate distribution. The beam could be defocussed until some minimum threshold of light output is obtained; conversely, in regions of high spot density the beam could be reduced in size and intensity until some upper limit is obtained. The probability density would then be a combined function of the light output, as discussed earlier, and the feedback voltage required to obtain the desired focus voltage level.

The direct computation of numerical values for smoothing parameters, as the requirement is described in Section 7. 1, might be accomplished by brute force with a simulation of such a procedure. The requirement, however, is for a more elegant approach which will result in a net saving of computational time. The ability to perform acceptable analog smoothing with an ATLUD does not eliminate the requirement of Section 7. 1. It does allow progress to continue in the development of signature processing hardware while, in the analytical problem area, attention is directed to the very difficult problem of determining how to select numerical values for smoothing parameters.

Operations performed by the Decision Network could be performed instead by converting the ATLUD output to digital form and passing the information to a small computer such as that found in the MSDS-GDS. This is also true of the operations performed by the MTG and the Transformation Network. In each case, it has been elected to recommend special hardware for the operations in order to increase the data throughput rate. By providing parallel computation channels and

special-purpose high-speed computational hardware, the rate of data processing can equal the rate of data collection. This is about 10^5 observations per second of multiple channel data. The serial computation and general-purpose computational hardware of a computer cannot approach the necessary speed.

The first of two operations available in the Decision Network is formation of linear combinations of the probability densities from the ATLUD. These are of the form

$$F_i = H + \sum_{j=1}^{M} G_{ij} f_j \quad , \quad i = 1, M$$

where

f_j = probability densities from ATLUD

M = number of assignment categories considered.

This operation is suppressed by letting the $G_{ij} = 0$ and $H = 1.0$. This suppression allows the $\{f_j\}$ to be the final output of the system. Otherwise, $H = 0$. When the operation is performed, two purposes are served. First, the weighting coefficients, G_{ij}, may be assigned values analogous to the coefficients "a" and "b" of Theorem 2 (Section 2.2). In this case, the MTG may be employed to determine logical variables for final output according to the truth of the relationships

$$f_i \geq F_i \qquad i = 1, M$$

The boundaries of the regions, C_i, within which these relationships are true are the boundaries of regions within which categories i are the maximum likelihood decisions and the probability of incorrect assignment is limited to the desired value. By applying Theorem 3 (Section 2.2) to generalize the utility of the coefficients, G_{ij}, the boundaries thus defined may also be caused to satisfy other criteria, e.g., minimizing cost, etc. These boundaries have the properties of those defined in the Decision Boundary Mode, but they are general in shape and are not similarly restricted to hyperparallelepipeds. Maximum likelihood boundaries which define mutually exclusive and exhaustive regions are specified by letting the $G_{ij} = 1/M$.

Second, the weighting coefficients may be assigned values in accordance with a priori probabilities and relative cost of decisions to allow computation of a conditional probability for each category. This is the second of the two operations performed by the Decision Network. It requires computation of ratios of the form:

$$P_i = \frac{Q_i f_i}{F_i} \quad i = 1, M$$

The coefficient, Q_i, is specified as the product of the a priori probability of a category i observation occurring and the relative cost judgment (or other weighting information employed by an algorithm of the user's choice) by which the decision in favor of category i is to be influenced.

This provides another continuous variable suitable for use in a color-coded decision display. As in the case of the display of the maximum likelihood, display of the maximum probability preserves the spatial information which belongs to image data. Display of the color corresponding to the category with the largest conditional probability is also a display of the maximum likelihood decision. By displaying probability instead of likelihood, the relative cost and a priori information is easily included. Because probabilities lie in the 0 to 1 range, scaling may be simplified; the full range, however, will not often be required for display purposes. A display variable with a dynamic range well suited to the display used can be derived from conditional probability to suit specific user requirements.

Other computations which present some difficulty when performed by computer may also be foreseen as applications for the Decision Network. The functional specification of the SDPS has been deliberately written in a modular and flexible form. The design and construction of the system will reflect this modularity and flexibility so that performance may be modified and capability extended as experience suggests. The explicit functional specifications are directed at achieving operational processing capability. The system capability can easily be extended to allow a direct supporting role in the evaluation of linear combinations for use in dimensionality reduction.

The probability densities computed by the ATLUD are computed from variables in the space defined by the Transformation Network. This is a space of reduced dimensionality. The computation of information divergence is used as one tool to evaluate the effectiveness with which the total amount of available information is concentrated in the variables produced by the Transformation Network. The computation of information divergence involves repetitive computations of probability densities. This places a practical limit on the number of such computations

performed and the sizes of the samples studied. Note, however, that the techniques of raster scanning the flying spot scanners in the ATLUD or using specially prepared sample tapes for the SDPS input (e.g., prepared by using the MSDS-GDS editing capability) will permit the computation of information divergence to be performed in an analog manner. The Decision Network capability can be extended to do this by adding the operations of integration and logarithm computation to those already specified.

7.3 HARDWARE PROBLEMS

Design studies, including a state-of-the-art survey, should be performed on some of the hardware elements forming the SDPS. None of the elements described require nominal component performance in excess of that which is currently available. Some choices lie in the selection of analog or digital methods and the location of conversion operations where necessary. Because the technology used in the components is now improving rapidly, selections of this kind should be deferred as long as possible to ensure consideration of all options.

One element of the SDPS, the Transformation Network, been constructed in breadboard fashion for the processing of eight-channel data. A portion of the ATLUD has also been constructed in breadboard form. A single flying spot scanner provides a bivariate probability density function. Some conclusions regarding the development of the corresponding elements for use in the design of an SDPS are drawn from experience with this equipment.

The Transformation Network used for dimensionality reduction is a device which performs an extremely large number of multiplications and additions. The input data for the SDPS are presented in digital form, and the Transformation Network is the first element to operate on the data. The only element of the SDPS which necessarily operates on data in analog format is the ATLUD. It is possible, therefore, to perform all necessary conversions as part of the ATLUD operations, allowing all other elements to operate on data in digital format. The trade-off required to make the selection between analog and digital electronics will involve the possibility of time sharing and serial processing with digital methods against the parallel processing and higher throughput rates with analog methods. High-speed digital electronics becoming available may offer throughput rates favorably comparable to those obtainable with analog methods. Miniaturized analog electronics, however, are also being offered with vastly improved performance characteristics. The selection will therefore depend upon the current state-of-the-art when the design is specified in detail.

The large number of coefficients and constraints required for operation of the Transformation Network make it desirable to automate the job of setting up the transformation. Using a computer for this job places no extra burden on the computer's time during operation, unless a time-sharing approach requiring frequent changes to accommodate serial transformations is selected. An annotation function is also served by the computer. The coefficients employed in the transform are to be encoded and appended to the housekeeping data track. For this reason, an equally acceptable method of specifying transformation coefficients and constants would be via manually set thumbwheels interrogated and documented by the control computer.

Similar considerations apply to the other elements for which the digital/analog option is available.

Problems identified for resolution during the design of the ATLUD include:

1. Excessive and changing phosphor noise on the CRTs of the flying spot scanners must be corrected.

2. Motion of the beam between positions of desired deflection leads to undesired spurious boundary enhancement.

The phosphor noise problem produces a nonuniform light output for constant beam energy as the beam is deflected. A long burn-in period is suggested by manufacturers to limit the amount of nonuniformity. This burn-in period does limit the variation to some extent. After a period of normal operation, without a raster scanning beam, the most frequently used portion of the screen has a decidedly lower light output than provided at the completion of the burn-in period. Flying spot scanners normally correct for phosphor noise variation by controlling the beam energy via a feedback loop. A beam splitter is placed between the CRT face and the transparency to be examined. The portion of the beam unattenuated by the transparency is directed to a photomultiplier (PM). This PM signal is used to modulate the beam intensity as required to keep the light output constant. Such a control loop was not provided in the breadboard version of the ATLUD, but will be a requirement for the final design.

The problem of spurious boundary enhancement produced when the beam moves from one position to the next is solvable when the initial input data format is digital. When consecutive observations, presented to a flying spot scanner in the form of two consecutive pairs of DC voltage levels, represent spectrally different target categories, the beam moves between two final resting places on the CRT face. To suppress the spurious output obtained during the motion of the beam, the beam may be blanked during this period or the PM output may be ignored during this period and for a sufficient time thereafter to allow the transients to decay. The clock pulse which accompanies the input data may be used as an indicator to determine the periods during which the analog output data should be ignored. This procedure is recommended for inclusion in the design specification.

A third problem, which cannot be investigated in detail without phosphor noise rejection, is that of calibrating the ATLUDs to ensure that each of the probability densities computed is scaled in a comparable manner. To allow maximum likelihood decisions to be made properly, each of the densities must integrate to the same value. A raster scan of the flying spot scanner during the unused portion of the duty cycle is suggested as an approach for performing this scaling. The PM

output, integrated over several raster periods, could be used as a feedback control of longer time constant in the manner described for phosphor noise rejection.

Deflection gain drift and drift in the undetected spot position have not been sufficiently stable for operation of a single flying spot scanner. For operation of a system of multiple flying spot scanners in a complete ATLUD, a calibration scheme should be devised to correct for such drifts or, if they are found to be relatively harmless and infrequent, to alarm and shut down in the event of such failure.

None of the problems identified or foreseen during the development of an operational SDPS require technology or components not currently within reach.

APPENDIX A

CLUSTER ANALYSIS

Frequently, in the course of the analysis of multivariate data, it is useful to represent a population probability density distribution in some parsimonious but reasonably accurate way. Usually, some simple parametric form is assumed and the parameters are estimated from a sample.

Even when it would be otherwise difficult to justify, the multivariate normal distribution is used because it is easy to determine the parameters and also easy to evaluate the density expression.

When a sample distribution exhibits excessive skew, kurtosis, multimodality, or other departure from normality, alternative representations are sought. This section describes a technique for dealing with the multimodal or clustered distributions when it may be hypothesized that each cluster is multivariate normal. The procedure requires an estimate of the probability density associated with each observation—this may be accomplished by a variety of methods, among which a particular one described by Specht* will serve as an example.

If the observations are resequenced to be in the order of descending probability density, the first observation may clearly be taken to be the mode of the first cluster. Each succeeding observation is then assessed with regard to the probability that it belongs to one of the existing clusters or, alternatively, that a new cluster must be formed.

Of fundamental importance in the assessment is the relation

$$f(\vec{x}) = f(\underline{\vec{x}}) \exp - 1/2 \, (\vec{x} - \underline{\vec{x}})^T \, C^{-1} \, (\vec{x} - \underline{\vec{x}}) \qquad \text{(A-1)}$$

which gives the probability density f for an observation vector $\vec{x}$ on the hypothesis that the distribution is multivariate normal with mean $\underline{\vec{x}}$ and covariance matrix C. The quadratic form in the right half of Equation A-1 can be replaced with the generalized Mahalanobis squared distance.

---

*D. F. Specht, "Generation of Polynominal Discriminant Functions for Pattern Recognition," IEEE Transactions on Electronic Computers, Vol EC-16, No. 3, June 1967.

$$d^2 = (\vec{x} - \underline{\vec{x}})^T C^{-1} (\vec{x} - \vec{x}) \qquad \text{(A-2)}$$

yielding

$$\frac{f(\vec{x})}{f(\underline{\vec{x}})} = \exp\left(-\frac{d^2}{2}\right) . \qquad \text{(A-3)}$$

Taking the log of both sides,

$$d^2 = -2\ell n \left(\frac{f(\vec{x})}{f(\underline{\vec{x}})}\right) . \qquad \text{(A-4)}$$

Note that Equations A-2 and A-4 can be regarded as two different ways of computing the same quantity, one using the hypothesized mean and covariance matrix and the other using the previously estimated probability densities.

In the cluster analysis algorithm to be described, the extent of the disagreement between the two computations makes an important contribution to the decision, because appreciable differences will occur only when the wrong cluster (and hence the wrong mean and covariance matrix) has been hypothesized. To allow for the difference, Equation A-4 will be rewritten using the symbol e to represent the value of d that would be expected from the densities f (x) and f ($\underline{x}$).

$$e^2 = -2\ell n \frac{f(\vec{x})}{f(\underline{\vec{x}})} . \qquad \text{(A-5)}$$

At this point, the notation will be permitted to reflect the presence of m clusters, each of which is to be tested with regard to the hypothesis that the current observation x arose from that particular cluster. Evaluating both Equations A-2 and A-5 for the ith cluster,

$$d_i^2 = (\vec{x} - \underline{\vec{x}}_i)^T C_i^{-1} (\vec{x} - \underline{\vec{x}}_i) \qquad \text{(A-6)}$$

$$e_i^2 = -2\ell n \left(\frac{f(\vec{x})}{f(\underline{\vec{x}}_i)}\right) . \qquad \text{(A-7)}$$

Since d_i is the observed generalized distance (between the observation and the center of the ith cluster) and e_i is the expected generalized distance, it is reasonable to define the variable α_i in Equation A-8 as the expected fraction of the observation relative to the ith cluster,

$$\alpha_i = \frac{e_i}{d_i} \tag{A-8}$$

and the unexpected fraction β_i then given by Equation A-9:

$$\begin{aligned} \beta_i &= 1 - \alpha_i \\ &= 1 - \frac{e_i}{d_i} \quad . \end{aligned} \tag{A-9}$$

The unexpected part of the observation relative to the ith cluster is the vector:

$$\vec{Z}_i = \beta_i (\vec{x} - \vec{\underline{x}}_i) \quad . \tag{A-10}$$

Clearly, since the covariance matrices C_i, the means $\vec{\underline{x}}_i$, and the probability densities $f(\vec{x})$ and $f(\vec{\underline{x}}_i)$ are estimated from the observations and hence are not free of errors, the unexpected fraction β_i is not identically zero even for the best choice for i.

Although a rule for making a choice has not yet been stated, it may be assumed that one exists. If the rule selects j for the correct value of i, the unexpected part of the observation for the chosen cluster is:

$$\vec{Z}_j = \beta_j (\vec{x} - \vec{\underline{x}}_j). \tag{A-11}$$

The covariance matrix of the Z_j for all previous selections can be called the unexpected covariance matrix. This matrix, denoted by W, plays a central role in the selection rule to be proposed.

The generalized Mahalanobis squared distance for the unexpected part of an observation $\vec{x}$ (relative to the ith cluster) is:

$$g_i^2 = Z_i^T W^{-1} Z_i. \tag{A-12}$$

The desired selection procedure can now be stated:

1. Evaluate g_i^2 for each cluster.

2. Let j be the value of i which minimizes g_i^2 for i = 1, m.

3. If g_j^2 is greater than a prechosen number Q, form a new cluster number m+1 centered at the given observation x.

4. If g_j^2 is not greater than Q, the observation is classified as part of the jth cluster, with appropriate modifications made to the mean $\underline{x}_j$ and covariance matrix C_j.

On the hypothesis that unexpected variance is multivariate normal, the g_i^2 will possess the chi-squared distribution and Q can be chosen to yield a specified probability of forming a new cluster by chance when a large deviation occurs.

The requirement for current inverses of the cluster and unexpected covariance matrices in spite of the frequent modifications to the matrices (by adding new members to the clusters) encourages the use of diagonalization rather than more conventional approaches to the inverse. Since any covariance matrix can be resolved into principal components of variance which are uncorrelated, the inverse required in Equation A-6 can be achieved by ordinary division of each principal component by the appropriate variance. Writing the diagonal matrix

$$D_I = U_I^T C_I U_I \tag{A-13}$$

The diagonal matrix D_I consists of the principal component variances of the Ith cluster when the matrix U_I represents a unitary transformation to principal component axes.

When an observation is to be added to the membership of the Ith cluster, the modification of the covariance matrix

$$\left[(N_I + 1)\, C_I\right]_{new} = \left[N_I\, C_I\right]_{old} + (\vec{x} - \vec{x}_I)(\vec{x} - \underline{\vec{x}}_I)^T \tag{A-14}$$

is replaced by its equivalent

$$\left[(N_I + 1)\; D_I\right]_{new} = \left[N_I\, D_I\right]_{old} + U_I^T\; (\vec{x} - \vec{x}_I)(\vec{x} - \underline{\vec{x}}_I)^T\; U_I. \tag{A-15}$$

REPRODUCIBILITY OF THE ORIGINAL PAGE IS POOR

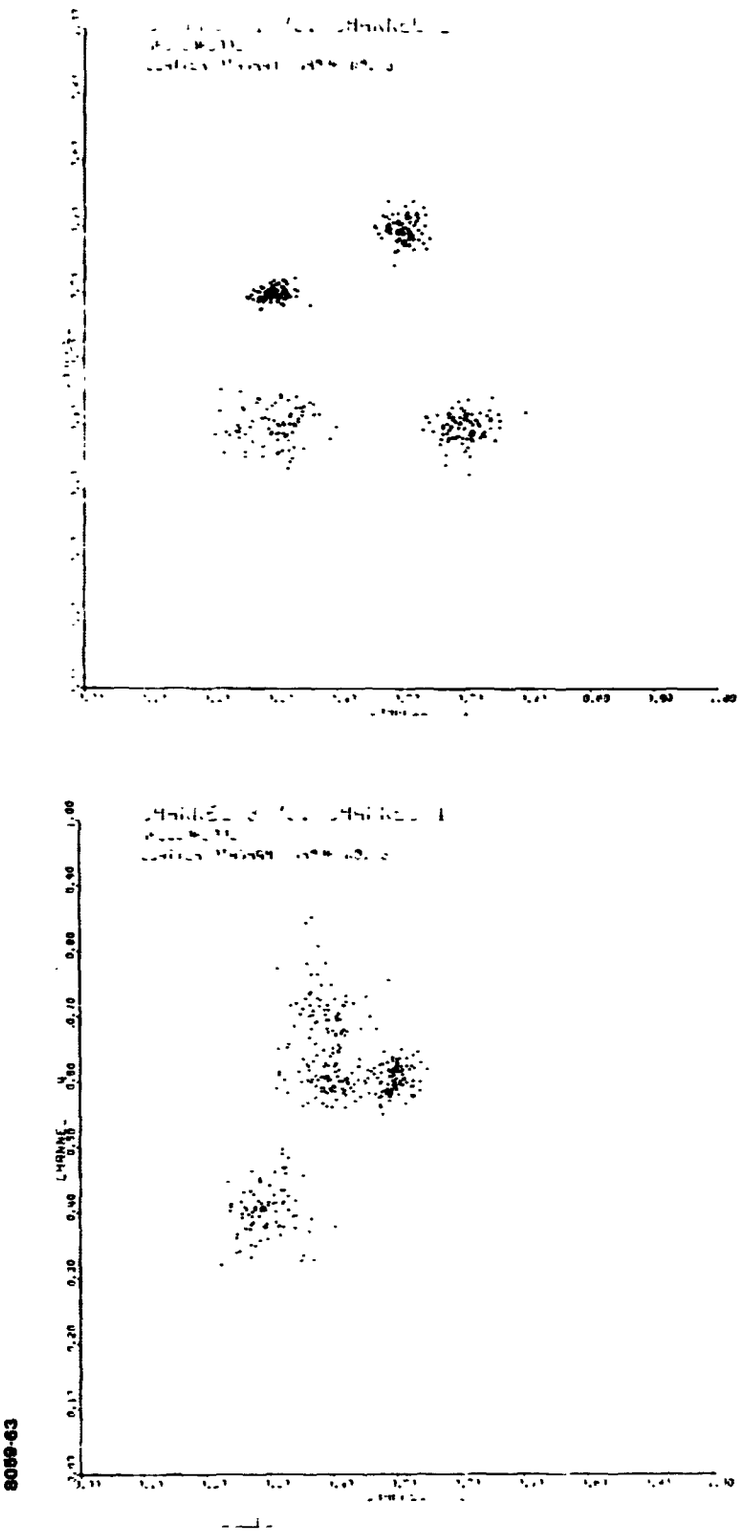


Figure A-1 Scatter Diagrams of Cluster Analyzed Data

A-5 — A-6
Missing in orig.