AD-781 146

PHOTOMETRIC ANALYSIS TECHNIQUES STUDY

R. G. Utley, et al

General Research Corporation

Prepared for:

Space and Missile Systems Organization
Advanced Research Projects Agency

July 1974

Prepared for
Department of the Air Force
Under Contract F04701-73-C-0308

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD 781146

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER CR-2-456	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER TR No. 74-100
4. TITLE (and Subtitle) Photometric Analysis Techniques Study		5. TYPE OF REPORT & PERIOD COVERED Final Report 1 July 1973-25 February 1974
		6. PERFORMING ORG. REPORT NUMBER CR-2-456
7. AUTHOR(s) R. G. Uttley, H. S. Ostrowsky		8. CONTRACT OR GRANT NUMBER(s) F04701-73-C-0308
9. PERFORMING ORGANIZATION NAME AND ADDRESS General Research Corporation P.O. Box 3587 Santa Barbara, CA 93105		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS ARPA Order 2437 SAMSO PE 63431F
11. CONTROLLING OFFICE NAME AND ADDRESS SAMSO, P.O. Box 92960 Worldway Postal Center Los Angeles, CA 90009, Attn: DYAX		12. REPORT DATE June 1974 13. NUMBER OF PAGES 110
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS (of this report) UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Approved for public release; distribution unlimited.

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Photometry
Satellites
Discrimination
Pattern recognition

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes some studies made on the application of pattern recognition techniques to the Space Object Identification (SOI) problem using passive photometric signature data. Both simulated and measured passive photometric data (primarily, the latter) on satellite objects were compared using a variety of techniques in both the time and frequency domains. Different sample spaces with dimensionalities ranging from 14 to 50 were tested, with Principal Components Analysis being used in an effort to reduce the effective number of (Continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block 20 continued

dimensions to a small subset with minimal loss of information. The GRC Mode Determination algorith, was then applied to the sample points in the reduced subsets to compare data from the different satellite/observer pairs in order to ascertain the extent to which this type of data processing is able to separate objects which are different and group together those which are in fact similar.

Results indicate that 50-point samples in the frequency domain give best performance. Defining a statistical similarity-dissimilarity index, it was found that in most cases objects known to be similar were indicated to be so, and objects believed to be dissimilar were so specified. However, the results in some cases appear to be sensitive to the precise form of data processing and/or the dimensionality of the sample space used; this suggests that an optimal processing technique has not yet been found.

FOREWORD

This Final Technical Report covers work done on the Photometric Analysis Techniques Study (Contract No. F04701-73-C-0308) during the time period 1 July 1973 - 25 February 1974, and is written in accordance with the requirements of Item A003 of Exhibit A of the Contract.

ABSTRACT

This report describes some studies made on the application of pattern recognition techniques to the Space Object Identification (SOI) problem using passive photometric signature data. Both simulated and measured passive photometric data (primarily, the latter) on satellite objects were compared using a variety of techniques in both the time and frequency domains. Different sample spaces with dimensionalities ranging from 14 to 50 were tested, with Principal Components Analysis being used in an effort to reduce the effective number of dimensions to a small subset with minimal loss of information. The GRC Mode Determination algorithm was then applied to the sample points in the reduced subsets to compare data from the different satellite/observer pairs in order to ascertain the extent to which this type of data processing is able to separate objects which are different and group together those which are in fact similar.

Results indicate that 50-point samples in the frequency domain give best performance. Defining a statistical similarity-dissimilarity index, it was found that in most cases objects known to be similar were indicated to be so, and objects believed to be dissimilar were so specified. However, the results in some cases appear to be sensitive to the precise form of data processing and/or the dimensionality of the sample space used; this suggests that an optimal processing technique has not yet been found.

CONTENTS

CONTENTS (Cont.)

ILLUSTRATIONS

ILLUSTRATIONS (Cont.)

ILLUSTRATIONS (Cont.)

TABLES

1 INTRODUCTION

Orbiting satellites can be observed by ground-based photometers and histories of their photovisual magnitude recorded. There is a distinct possibility that such histories, either alone or in conjunction with other data and intelligence, can provide a basis for significant inferences about the external characteristics (size, shape, materials) and body motions of satellites. If so, similarities and dissimilarities between satellites could be recognized, and a basis would exist for associating satellites into groups with a common mission. It might also be possible to link external characteristics and motion to the nature of a satellite payload, thereby helping to identify what its mission is. Thus it is important to establish what can be learned from photometric histories, both in connection with concealment or disguise of our own satellites, and mission identification for Soviet satellites.

The present study is an effort to develop semi-automated methods for analysis of such photometric data. The data with which we are concerned are time histories of the radiant intensities of the targets as unresolved point sources under solar illumination, measured passively. The study is primarily oriented toward the use of portions of the existing GRC DISCRIMATON computer program to separate the photometric data from different targets into distinguishable object-class clusters in some appropriate multi-dimensional state-space. The selection of suitable state-spaces which most conveniently facilitate this separation was one of the major goals of the study.

As the first step, a small library of simulated photometric data (signature/time histories) for Agena tanks in representative orbits was generated using other existing GRC computer programs. These synthetic data were used to test the ability of DISCRIMATON to distinguish between the satellites and expended, tumbling rocket tanks. Following the success of this test, the remainder of the study was concentrated on determining similarities and differences among those satellites for which

actual measured data were available, using various sorts of state spaces. A logical future extension would be to simulate the signatures of typical satellites in representative orbits and then, using DISCRIMATON, compare the state-space points derived from the analysis of given observational data with state-space groups derived from the simulated data in an attempt to determine the (known) simulated object(s) with which the observed data can best be associated.

2 SUMMARY AND DISCUSSION OF RESULTS

This section states the principal results of the study and discusses factors which bear on their validity. One such factor is the extent of the data base, which limited the number of samples analyzed to an extent where we can not be certain about the statistical significance of the degree of dissimilarity found between data sets. If certain of the results are in fact valid, we conclude that classification methods similar to those used here can be successful if a means is developed for inferring photometric signatures in other orbits from an observed signature in a given orbit.

2.1 DATA SOURCES

The study concentrated on comparison of actual photometric data on three satellites. In addition, the photometric signatures of three tumbling Agenas were simulated based on a simplified engineering drawing and using orbits and tumble rates typical of actual spent Agenas. These were compared with the satellite signatures early in the study as a first test of the efficacy of the DISCRIMATON pattern-recognition program. The sources and nature of all the data analyzed during the study are listed in Table 2.1.

The satellites were all observed at long range (~37,000 km). The 4630 data was taken at three different sites on the same day, with considerable overlap in the observation times. It was therefore reasonable to expect a priori that these three data sets would be quite similar. We learned from SAMSO after our analysis was completed that 5851 and 4630 are sister satellites in different orbits. Apart from the preceding, we have only a very small amount of information on the orbits of the satellites, and none whatever on the relation between 5587 and the others.

2.2 PERIODICITIES

Periodicity of the observed data was evident from inspection; 4630 and 5851 were each found to have a period of 10.0 seconds, while that of

TABLE 2.1

DATA SOURCES AND ASSOCIATED CODE NUMBERS

Object/Observation Site	Actual or Simulated Data	Code Number
Agena (Orbit 1)/Sulphur Grove	Simulated	1
Agena (Orbit 2)/Sulphur Grove	Simulated	2
Agena (Orbit 3)/Sulphur Grove	Simulated	3
5851/Cloudcroft	Actual	4
4630/Cloudcroft	Actual	5
4630/AMOS	Actual	6
4630/RML	Actual	7
5587/Cloudcroft	Actual	8

5587 was 1.2 seconds. The simulated Agena data had input periodicities. However, while periodicity greatly simplified data sampling, we deliberately normalized all periods to unity before applying our classification algorithms, thereby preventing period length from playing any part as a discriminant between the data sets.

2.3 MEASURE OF SIMILARITY BETWEEN DATA SETS

Samples from different data sets were processed together through selected subroutines from an already-existing GRC computer program to assign them into groups. Both the number of groups and the assignment of samples were determined by the program.

A given data set can be (and usually is) represented in more than one class. Since a data set is for a single object (under specified viewing geometry), we are led to substitute the word "feature" for "class." We have not determined what actual physical characteristics are mainly responsible for these "features," but for reasons given below our view is that such correlations must be established if analysis based primarily on pattern-recognition is to progress further.

The overall similarity of two data sets depends not only on what "features" they have in common, but also on how frequently these features occur. A simple measure of similarity which takes into account both these factors is

$$S(i,j) = \frac{1}{\nu} \sum_{r=1}^{\nu} \frac{\min\left[\frac{N_{ir}}{N_i}, \frac{N_{jr}}{N_j}\right]}{\max\left[\frac{N_{ir}}{N_i}, \frac{N_{jr}}{N_j}\right]}$$

where i,j refer to the respective data sets, and

N_i = number of samples from data set i

N_j = number of samples from data set j

ν = number of "features" distinguished (i.e., number of groups)

N_{ir} = number of samples from the ith data set which exhibit the rth feature (r = 1, 2, ... , ν)

N_{jr} = number of samples from the jth data set which exhibit the rth feature (r = 1, 2, ... , ν)

We call S(i,j) the "Similarity Index" between the two data sets. It has a maximum value of unity when both data sets exhibit the same features the same percentage of the time, and a minimum value of zero when the sets have no feature in common.

2.4 RESULTS

2.4.1 Calibration of the Similarity Index

Measurement "noise" may appreciably reduce the similarity index from the value it would have if the noise could be completely removed. An indication of how much reduction occurs in actual data when it is

reasonable (for reasons given earlier) to suppose that the index for the corresponding noise-free measurements would be at least fairly close to unity, can be obtained by comparing code 5, 6, and 7 data. Graded by visual inspection code 5 data appears to be of good quality (low "noise" level), code 6 data to be fairly good, and code 7 data to be poor. The similarity indexes between these sets, based on the frequency spectrum of samples, are given in Table 2.2.

These results suggest to us that:

1. Data should be of at least fairly good quality, as assessed by visual inspection, to be used in classification analysis by our methods. Use of poor quality data is liable to give rise to misleading results.

2. A similarity index of about 0.5 or higher is indicative of considerable physical similarity between the observed objects.

2.4.2 Comparison of Data Sets for Different Satellites

Similarity indexes between the data sets for different satellites are given in Table 2.3. All these measurements were taken at Cloudcroft, and were visually judged to be of good quality. On the basis of our above crude calibration of the similarity index, our qualitative interpretation of these results is that:

TABLE 2.2

SIMILARITY INDEXES FOR 4630 DATA

Data Sets (i,j)	Similarity Index S(i,j)
5,6	0.560
5,7	0.076

TABLE 2.3

SIMILARITY INDEXES FOR DIFFERENT SATELLITES

Data Sets (i,j)	Satellite Identification Numbers	Similarity Index S(i,j)
4,5	5851, 4630	0.0
4,8	5851, 5587	0.033
5,8	4630, 5587	0.28

- There is little or no evidence of physical similarity between 5851 and 4630, or between 5851 and 5587.
- There is evidence of some physical similarity between 4630 and 5587.

2.5 FACTORS AFFECTING VALIDITY OF THE RESULTS

All the above results are subject to a caveat about sample size. Sample sizes were limited by the amount of data available. The number of samples in the frequency domain taken from each data set is given in Table 2.4. Each sample was an ordered set of 50 numbers, represented geometrically as a point in a 50-dimensional Cartesian space. It is

TABLE 2.4

SAMPLE SIZES

Data Set	Number of Samples
4	24
5	33
6	30
7	25
8	30

known that classification schemes in which class boundaries are estimated on the basis of the available samples alone are liable to give misleadingly low error rates when the ratio of the number of samples to the dimensionality of the samples is less than about 3, as is apparently true in all the comparisons discussed above. This implies that the similarity indexes obtained are liable to be too low. In other words, our results are biased (perhaps heavily) toward indications of dissimilarity.

However, we believe that the problem is not as severe as appears at first sight. The reason is that in all cases we found that of the 50 principal axes of the correlation ellipsoid (calculated for the pooled samples), at most 12 were more than one-fifth as long as the major axis, and at most 19 more than one-tenth as long. In other words, the spread of the data is relatively small in at least 31 dimensions. This suggests to us that the minimum number of samples needed may well be a lot closer to 50 than 150. In fact, we did compare codes 4 and 5 data in the time dimension (where more samples could be taken) using 55 code 4 samples and 117 code 5 samples. The finding was again that the two sets were dissimilar.

We must also point out, on the other hand, that our methods may be inadequate to recognize dissimilarities which in fact exist; and that our grouping algorithm is heuristic, without any claim to "optimality." Nevertheless, our methods have worked well in previous applications where results could be compared with what would be found by a classifier with complete a priori knowledge of population statistics. This is the rationale for their application in the present context.

2.6 REQUIREMENTS FOR A "SIGNATURE TRANSFER" TECHNIQUE

If the sample size caveat could be removed from the present finding that code 4 data (5851/Cloudcroft) and code 5 data (4630/Cloudcroft) sets are dissimilar, an important conclusion could be drawn about how photo-metric signatures should be compared. We were told by SAMSO (after

our results had been presented) that satellites 5851 and 4630 have the same externalities and body motion, but are in different orbits. The conclusion is that a finding of strong dissimilarity in observed data for satellites in different orbits cannot in general be regarded as implying that the satellites are dissimilar.

In our view, this shows that a requirement for successful classification of satellites is a means for inferring photometric signatures in other orbits from an observed signature in a given orbit. This might be accomplished by using our methods to correlate the signatures of various external features, calculated at the same ranges and phase angles as the observed satellites with the observed data. In effect, this would amount to the progressive build-up of a reference data base. A major obstacle is the long running time of programs that generate synthetic photometric data, coupled with the very substantial number of externalities, body orientations, and motions that would probably need to be tried. On the other hand, the degree of success inherent in the approach would be indicated by comparing the synthetic signature of a given satellite, derived from engineering drawings, with actual observation of the same satellite. This could be done at reasonable cost.

3 ORGANIZATION OF THE REPORT

Our data processing schemes are described in Sec. 4. The description encompasses the various forms of data preprocessing utilized and the pattern recognition algorithm itself. An outline of the capabilities of our computer program for generating synthetic photometric histories is also given.

The sources and extent of the satellite histories that became available during the study are listed in Sec. 5. Some excerpts from this data are shown in Sec. 6, together with some from the simulated histories of the tumbling Agenas.

Section 6 also provides specific details of how the data was sampled and preprocessed before input to the pattern recognition algorithm for both the time and frequency domains.

Finally, Sec. 7 describes and discusses the findings of our pattern recognition algorithm. It compares the results of classification in the time domain with those in the frequency domain, and shows the effect on classification of such parameters as the dimensionality of samples, various types of sample normalization, and choice of what constitutes a "sample point." Some practical questions about the validity and scope of the algorithms are answered in the appendix.

4 DESCRIPTION OF BASIC PROGRAMS AND TECHNIQUES

The algorithms and major computational capabilities required by the study are embodied in three computer programs. Two of these, PHOTARG and FACTRY, were devised by General Research Corporation and are described below. The other is a readily-available Fast Fourier Transform routine written by IBM and used to obtain the frequency spectra of selected time-intervals of recorded satellite histories.

The problem of estimating classification error probabilities is also discussed.

4.1 GENERATION OF SIMULATED HISTORIES--PROGRAM PHOTARG

Program PHOTARG computes the radiant intensity of an object illuminated by the sun and seen from an observing photometer. This is converted to units of absolute satellite magnitude (defined to be the apparent photovisual magnitude of the object evaluated at a standard reference range of 1000 km from the sensor). Because of the long running time of program PHOTARG, and because the primary emphasis of this study was intended to be on the analysis of measured photometric data from real satellites, the program was actually used only to generate representative histories of tumbling Agenas.

The program requires that the target be decomposable into a combination of certain shapes in known geometrical relationships and orientations with respect to one another and a target-centered coordinate system. The list of permissible shapes is broad enough to allow simulation of many targets of interest; the shapes are:

1. Spheres and segments of spheres
2. Cones, conic frustra, and segments of either
3. Cylinders and cylindrical segments
4. Rectangular plates
5. Discs and segments
6. Ogives and segments
7. Prolate spheroids and segments

In addition, the program requires a table giving the (measured) bidirectional reflectance of the material of which each shape is composed.[1]

4.2 THE PATTERN RECOGNITION ALGORITHM--PROGRAM FACTRY

Program FACTRY consists of two main subprograms: FACTOR, which performs a process called "Principal Component Analysis" on the input data samples, and LERNMOD, which embodies the essential pattern recognition procedures. Both are subroutines from a more comprehensive program, DISCRIMATON, developed by GRC in 1968 for use in reentry vehicle/decoy discrimination work for SAMSO/Aerospace.

The LERNMOD grouping algorithm is heuristic, and cannot claim "optimality" in any sense, though in previous applications it has often performed better than the human eye. It was developed by experimentation with samples from normal distributions, primarily in up to ten dimensions. Consequently, before reaching LERNMOD, the input samples, whose dimensionality here is usually 50, are preprocessed through FACTOR to assess the minimum dimensionality of the space in which they can reasonably be considered to lie.

Deciding what dimensions can be discarded after FACTOR has been applied involves the risk that important classification information is being rejected. A simple illustration of this is given in Sec. 4.2.1 below.

4.2.1 Principal Component Analysis--Subprogram FACTOR

The steps in Principal Component Analysis (PCA) are:

1. Calculate the correlation matrix of all the samples to be processed.
2. Find the eigenvalues of the matrix (i.e., "diagonalize" the matrix); discard those eigenvalues which are small by comparison with the largest.

3. Find the eigenvectors corresponding to the retained eigenvalues and project the sample points into the subspace defined by these eigenvectors.

The rationale for this procedure is as follows. Each member of the sample set is initially represented as an N-dimensional vector. Let us now suppose that each of these vectors is actually a linear combination of M specified vectors, where $M < N$; in other words, all the samples lie in an M-dimensional subspace of the original N-dimensional space. Then it can be shown that the sample covariance matrix has rank M , and the eigenvectors corresponding to the non-zero eigenvalues are a set of basis vectors defining the M-dimensional subspace containing all the sample vectors. Thus, if we have a situation where some eigenvalues, though not zero, are small by comparison with the largest, we can reasonably (but not always correctly) disregard the dimensions defined by the corresponding eigenvectors, and thereafter use the projections of the samples into the space defined by the retained eigenvectors instead of the samples themselves.

In Principal Component Analysis, the initial samples are first subjected to a change of scale in each dimension so that the standard deviation in each dimension is unity. This removes any dependence on units, and (what is most important in the context of the present study) has the effect of assigning equal importance to the same percentage fluctuation in each dimension rather than to the same magnitude fluctuation. This change of scale transforms the initial sample covariance matrix into the corresponding correlation matrix.*

---

*The process known as Factor Analysis omits the step described in this paragraph, and diagonalizes the covariance matrix. It therefore attaches equal weight to the same magnitude fluctuation in each dimension. In the cases treated here there is no fundamental objection to using Factor Analysis rather than Principal Component Analysis, and we show some results for both in Sec. 7.

In discarding dimensions in the way described above, we ran some risk of discarding important information for discrimination. Essentially, application of PCA inherently assumes that separation between different data classes is likely to be greatest in those directions where the spread of the projected sample points is greatest. Though this is often the case, it need not be, as Fig. 4.1 illustrates. Here, the maximum spread is in the direction x' (the eigenvector corresponding to the larger eigenvalue), while the spread in the y' direction (the eigenvector corresponding to the smaller eigenvalue) is comparatively small. But discarding the y' dimension would clearly result in elimination of all useful discrimination information.

We cannot be sure that (less easily perceived) errors of this kind have not been made in the present study. However, this could only have occurred in one case, where near-simultaneous observations from two different ground stations on the same distant satellite were compared. In the other cases, where data on different satellites was being compared,

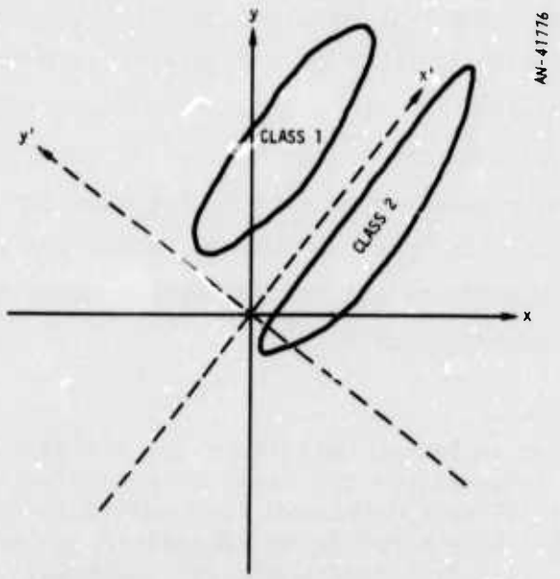


Figure 4.1. Illustration of Potential Rejection of Important Classification Information Following Principal Components Analysis

we found perfect or near-perfect separation of the data sets; the discarded information could therefore have at most a small effect on these results.

4.2.2 The Sample Grouping Algorithm--Subprogram LERNMOD

The technique used here is based on an algorithm used previously by Specht and Sebestyen to generate a smooth estimate for a probability density function when a number of samples is given. The structure of the grouping algorithm is based on the knowledge that clusters of points are likely to correspond to local maxima in the estimated density function. However, processing the data requires several additional steps.

These steps, which are described below, determine by an iterative procedure the locations of concentration centers in the data space, and then assign each point to one of those modes on the basis of maximum probability. The iteration continues until all points are accounted for. The algorithm can decide for itself how many groups are present; this need not be an input to the program by the user.

The algorithm begins with evaluation of a few basic parameters. The first of these is the standard deviation τ of the smoothing normal distribution used to evaluate the approximate overall density function. When the standard deviation is used for a distribution known to represent a single type, a value for τ derived from the second-order statistics is reasonable. In the present circumstance, however, the samples may conceivably be derived from several such widely separated distributions that an equivalent formula would yield a value of τ that is too large. This could, in turn, shift or even eliminate the peaks corresponding to local maxima. A parameter that is less sensitive to the placement of individual unimodal distributions is the nearest neighbor distance averaged over the population and denoted $\bar{d}_{min}$. However, to avoid completely losing touch with the second-order statistics, the geometric mean of the eigenvalues of the correlation matrix (g) is also evaluated; the chosen parameter, after a number of trials, was

$$\tau = \min(\bar{d}_{min}, 2\sqrt{g}/\sqrt{N}) \tag{4.1}$$

where N is the number of samples to be clustered. Two additional parameters, whose use is explained below, are given in terms of the same quantities plus the dimension n of the space by

$$\rho_o = \min\left\{\sqrt{g}, \max\left[0.15n(N)^{1/n}\ \bar{d}_m, \sqrt{g}/4\right]\right\} \tag{4.2}$$

$$P_T = \frac{0.3989}{N} + \left(4 \times 10^{-3}\right)\frac{(N)^{1/2n}}{n} \tag{4.3}$$

The iterative process can now begin. Its progress at various stages will be illustrated by what it does to the samples shown in Fig. 4.2. The main steps in the iteration are as follows.

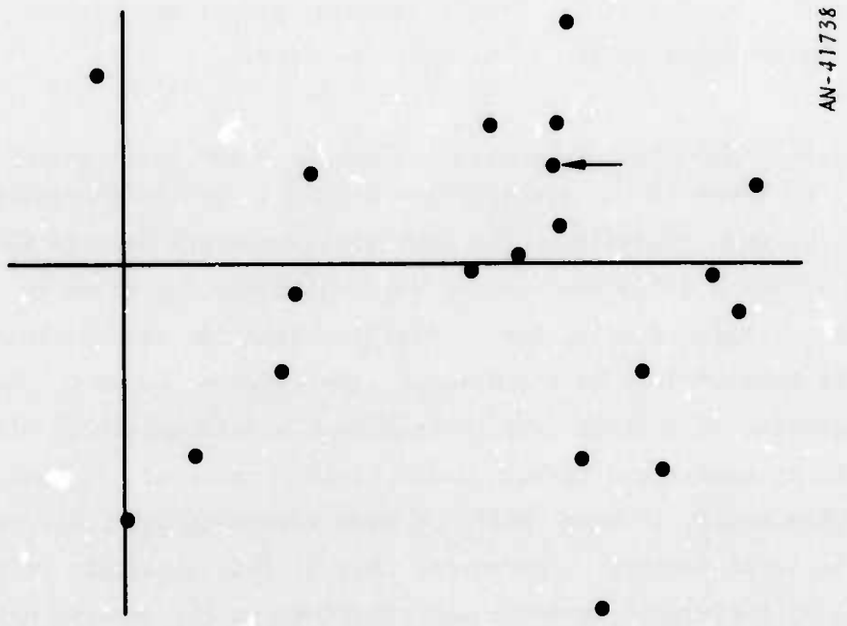


Figure 4.2. A Typical Set of Twenty Random Samples (arrow indicates point at which $p(\vec{x})$ is maximum on the first iteration) in a Two-Dimensional Sample Space

Step 1. The "probability" $p(\vec{x}_i)$ is evaluated at each of the unassigned samples points, $\vec{x}_i (i = 1, 2, \ldots, N')$, using the formula

$$p(\vec{x}) = \frac{1}{N'} \sum_{i=1}^{N'} e^{-|\vec{x}-\vec{x}_i|^2/2\tau^2} \tag{4.4}$$

N' is equal to N on the first iteration.

Step 2. The unassigned samples are ordered according to decreasing probability, as given by Eq. 4.4.

Step 3. If the maximum probability is less than P_T, the iteration is terminated. This criterion eliminates considering isolated points for the following steps.

Step 4. The unassigned sample with highest probability (and not previously rejected by Step 5) is considered to be the center of a hypersphere of radius ρ_o. For five times, recompute the center as the average of the samples lying within it and form a new hypersphere with a radius increased by 10 percent of the original radius. Thus the radius of the final hypersphere is $1.5\rho_o$. The result for the samples in Fig. 4.2 is shown in Fig. 4.3.

During the five recomputations of the iteration a new sphere center is evaluated by averaging any of the original set of samples that lie within the previous sphere, whether or not they have been assigned to modes. This has a two-fold effect: first, it tends to reduce the dependence of the entire process on the size of τ, the smoothing standard deviation; second, after the first mode has been found, if the sample selected for the starting center is on the slope of the probability density function, the hypersphere iteration tends to move the center toward the peak--the circle is thus less likely to be rejected as a spurious mode by one of the following tests.

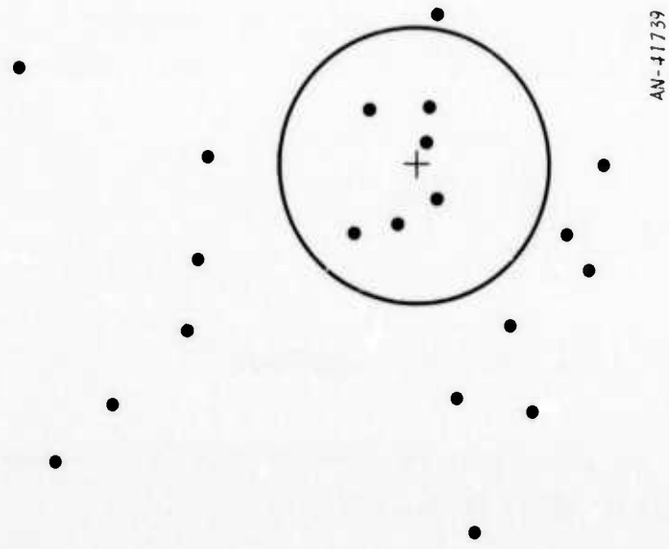


Figure 4.3. The Circle (hypersphere) Found Around the Point of Maximum Probability with Radius $1.5\rho_o$ (The center is denoted by +.)

Step 5. The final hypersphere found in the current iteration is examined for the number of samples contained. If this number is less than 5 percent of the total number of samples, the hypersphere center is rejected. In this case, the process returns to Step 4.

Step 6. The hypersphere is compared with any that may have been found previously. If the distance between its center and the center of any earlier hypersphere is less than $2.25\rho_o$, the hypersphere and its initial sample are rejected and control returns to Step 4. If the distance to every earlier center is greater than $3.0\rho_o$, control passes to Step 8, otherwise to Step 7.

Step 7. If control passes to this step, the current hypersphere intersects one or more of the previously found hyperspheres (though not to such an extent as to be rejected). Where such an intersection takes place, the point $\vec{x}(m)$ lying midway between the centers is examined.

If $\vec{x}_i$ is the sample that acted as the original center for the current hypersphere, then the current hypersphere is rejected if

$$p\left(\vec{x}(m)\right) > p(\vec{x}_i) \tag{4.5}$$

If this inequality (4.5) does not hold, then the samples lying within the previously determined hyperspheres are examined to see if they also lie within the current one. If they do, they become unassigned--but with a special flag to show that they are not to be reconsidered as potential hypersphere centers.

Step 8. The samples lying only within the current sphere are considered as initially defining a mode and assigned accordingly. The number of the current hypersphere becomes the number of the mode. The associated samples are now effectively removed from the major iteration as control passes back to Step 1. Thus, in the example, the iteration would start again, faced with the function sample points of Fig. 4.4.

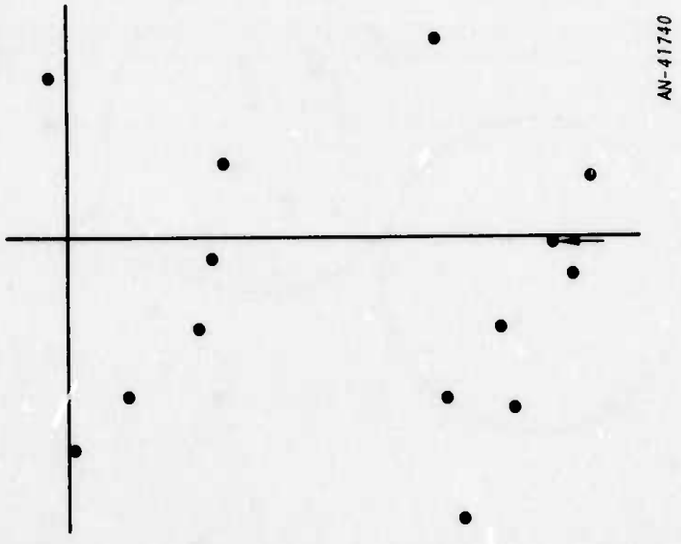


Figure 4.4. Samples Remaining After the Samples Initially Assigned to the First Mode are Removed (The arrow again indicates the point of maximum probability.)

When the iteration stops, all allowable modes have been found, being represented by the samples assigned to them. For this example, only two modes are predicted at this point, and these are represented by the points within the circles of Fig. 4.5. It remains only to assign those points as yet unidentified with a particular mode (these points may include those that were lying within two or more of the hyperspheres during the iteration). To accomplish this, the unassigned samples are first ordered on the basis of increasing distance to the closest cluster mean. Starting with the one closest to some mean, the probability $p_i(\vec{x})$ defined by the ith cluster is evaluated at the sample $\vec{x}$ with the formula

$$p_i(\vec{x}) = \frac{1}{N_i} \sum_{j=1}^{N_i} e^{-|\vec{x}-\vec{x}_j|^2/2\tau^2} \tag{4.6}$$

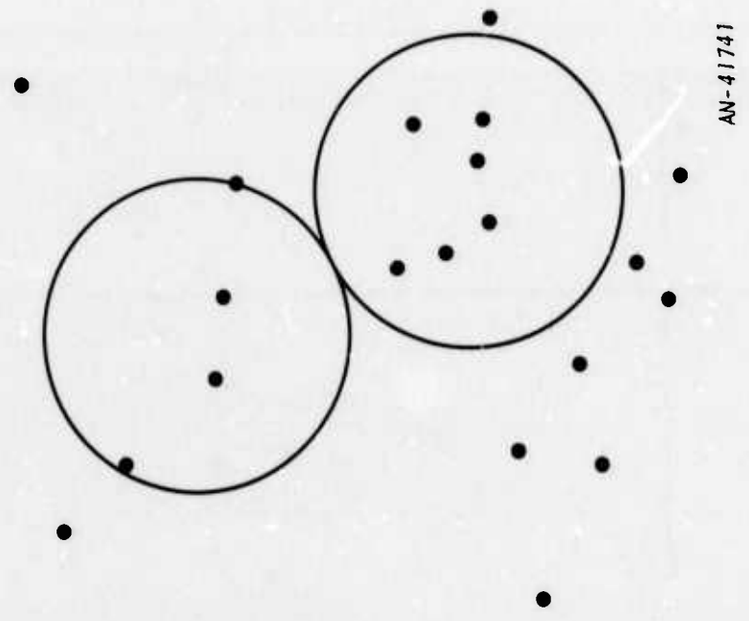


Figure 4.5. The Two Circles Found by Iteration Over the Samples

where the sum runs over the samples already assigned to the mode. The sample $\vec{x}$ is then assigned to the mode for which $p_1(\vec{x})$ is a maximum, and the process is repeated until all samples are assigned. For the sample, the samples are assigned according to the groups indicated in Fig. 4.6.

4.2.3 The Problem of Classification

Samples from the photometric history of a particular satellite, either in their original form or preprocessed in some way, may fall into one or several clusters. We can think in terms of each of these clusters representing a "feature" of the satellite's signature. Ultimately, we would like to be able to establish correlations between these "features" and the external characteristics and motion of the satellite, but this has not yet been attempted.

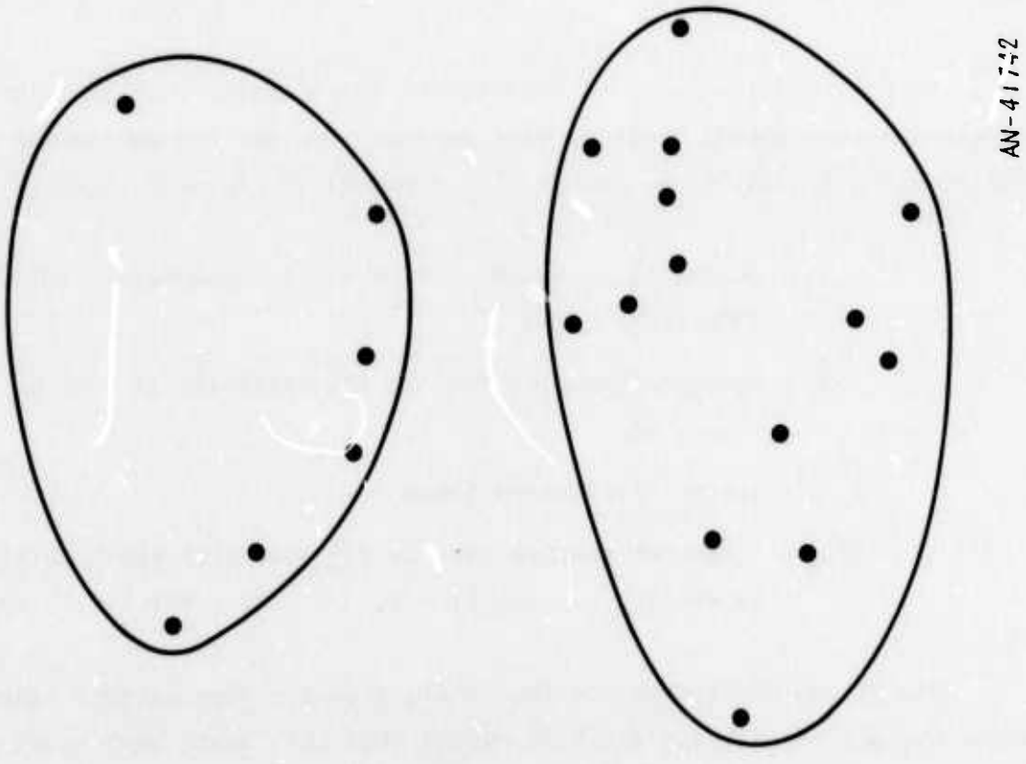


Figure 4.6. Clustering of the Samples of Fig. 4.1

Indeed, for our approach to be fully successful, such correlations would have to exist. The photometric signature of a given satellite with given body dynamics and orientation can generally be expected to depend strongly on its orbital parameters, and therefore there is a distinct possibility that two satellites of the same type in substantially different orbits will not exhibit any common features, even though we can determine (for example) that they have the same rotation period. Since such cases may well exist, we are likely to need a technique for "transferring" a satellite signature from one orbit to another. It presently seems that we would have to do this via correlation of features in one orbit with underlying externalities, then simulate the combined signature of these externalities in the other orbit. Obviously, however, this should be no different from comparing the externalities inferred from one signature with those inferred from the other. We can take one approach or the other depending on how extensive a library of signatures is available at the time.

Suppose now that we have transformed the signatures of various satellites to a common basis as best we can. We can then go through the grouping algorithm and emerge with a number of clusters. Let

k = number of satellite (or satellite/observation site) data sets

N_i = number of samples for the ith satellite ($i = 1, 2, \ldots, k$)

ν = number of clusters found

N_{ir} = number of samples for the ith satellite which fall in the rth cluster ($i = 1, \ldots, k$; $r = 1, \ldots, \nu$)

Now if two satellites are identical, then for observations taken under the same conditions we would expect that they would have equal representation in the same cluster. On the other hand, if they were

quite dissimilar, they would not be represented in the same cluster. This suggests defining a "similarity index" $S(i,j)$ between the ith and jth satellites by the equation*

$$S(i,j) = \frac{1}{\nu} \sum_{r=1}^{\nu} \frac{\min\left[\frac{N_{ir}}{N_i}, \frac{N_{jr}}{N_j}\right]}{\max\left[\frac{N_{ir}}{N_i}, \frac{N_{jr}}{N_j}\right]} \qquad (i,j = 1, 2, \ldots, k)$$

Obviously,

$$0 \leq S(i,j) \leq 1$$

$S(i,j) = 0$ implies complete dissimilarity, while $S(i,j) = 1$ implies identity.

*If N_{ir} and N_{jr} are both zero, the rth term in the sum is defined to be unity.

5 DATA BASE

Nearly all the data used in this study are recordings of actual observations of satellites by telescopes and associated equipment. Simulated data was prepared only for tumbling Agenas. Some excerpts from the data are shown in Sec. 6.

5.1 OBSERVED DATA

Observed photometric data, calibrated, digitized, and reduced to a data tape format compatible with our CDC 6400 computer, was transmitted to us via KMS Technology Center. Because GRC's software differs somewhat from most other systems, creation of a compatible data tape (called "REDDI-TAPE" by KMS) and reading it out correctly required a fair amount of effort.

The sources and quality of the data that became available to GRC in the appropriate format during the course of the study are listed in Table 5.1. If it was used in the study, data from a given source and recorded at a given site on a specified date was assigned a "code number" for identification, as shown in the table.

All the data used exhibited clear periodicity. In view of the limited time remaining after we received it, we did not use the 31 October 1973 data from Cloudcroft for 5851 and 6991; visually, they appeared similar to 4630. 5587 data was also received late, but appeared different enough to warrant inclusion (there is little enough variety in the list); both SAMSO and we were interested to see how this data would group relative to the groupings already found for 4630 and 5851.

5560 data was not used since it appeared to be aperiodic and vastly different from any of the other data available. There was nothing else of a similar type to compare it with.

TABLE 5.1

TAPED PHOTOMETRIC DATA BASE

Object ID	Observing Site	Observation Date	Quality	Comments
4630	Cloudcroft	10 November 1972	Good	Used--Code 5
4630	AMOS	10 November 1972	Fair	Used--Code 6
4630	RML	10 November 1972	Mostly Poor	Used--Code 7
4630	AMOS	15 November 1972	Poor	Tape Unreadable
5560	Cloudcroft	Several Days	Good	Not Used
5851	Cloudcroft	17 May 1972	Good	Used--Code 4
5587	Cloudcroft	9 May 1972	Good	Used--Code 8
5851	Cloudcroft	31 October 1973	Good	Not Used
6991	Cloudcroft	31 October 1973	Ranges from Good to Poor	Not Used

The apparent quality of the data, as indicated by a visual appraisal of its noisiness, varied from usually good (at Cloudcroft) to fair (at AMOS) to poor (at RML). There is always a real question as to whether data of very doubtful quality should actually be used. We did use it here because the study is exploratory in nature, and we wanted to see what our pattern recognition process would make of it.

A particular item to be noticed is that the first three data tapes on the list were taken on the same object on the same day, but from three different sites; in fact, the data intervals actually overlap to some extent. Since this satellite is at very high altitude, we would expect strong similarities to be found by the grouping algorithm.

On all data tapes received the variable recorded is Absolute Satellite Magnitude (M_S). By definition, M_S is the apparent photovisual magnitude of the target referred to a standard range of 1000 km from the sensor.

5.2 SIMULATED DATA

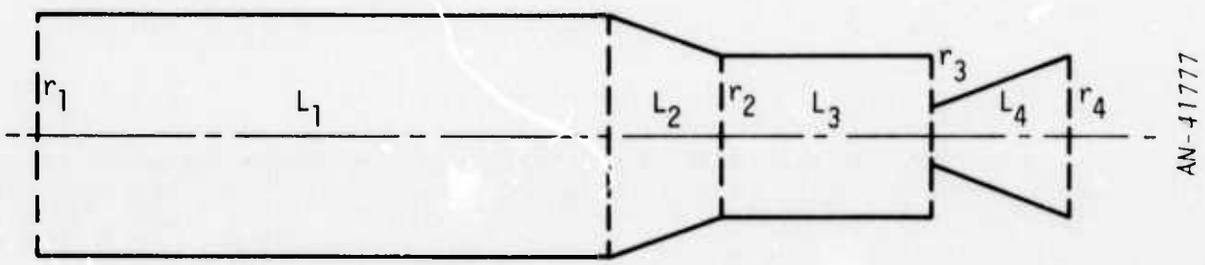
Photometric signatures were generated for tumbling Agenas in two different orbits. The first was that of 5560 (inclination ≃ 92.64 deg, mean altitude ≃ 680 km, period ≃ 100.58 min), and the second that of Agena 1963-27A (inclination ≃ 82.33 deg, mean altitude ≃ 480 km, period ≃ 93.92 min). The observing site was at the location of Sulphur Grove in all cases.

For the second orbit, two histories were generated, one with the Agena between the observer and the sun, and the other in the more usual position opposite the solar direction.

In all the cases, the Agena motion was pure tumble. The tumble axis was chosen at random in each case, but once chosen, remained fixed in inertial space. Tumble periods were selected roughly in accordance

with data on typical observed flash periods of actual orbiting Agenas (Ref. 2); the periods used were 5.0, 7.6, and 12.27 seconds.

The Agena body was approximated as shown in Fig. 5.1, with dimensions taken from an engineering drawing. The body surfaces were assumed to have the bidirectional reflectance properties of aluminum, generally in accordance with the data for "aluminum trim tape" in Ref. 1, but with modifications to incorporate some broadening of specular peaks in the reflectance curves which could occur due to small-scale undulations of the body surface. This model is evidently rather crude, but was felt to be satisfactory in the context of its intended use.



DIMENSIONS:

L_1 = 3.49 m	r_1 = 0.75 m
L_2 = 0.71 m	r_2 = 0.48 m
L_3 = 1.29 m	r_3 = 0.16 m
L_4 = 0.83 m	r_4 = 0.42 m

SURFACES:

1	CYLINDER
2	CONE FRUSTUM
3	CYLINDER
4	CONE FRUSTUM

Figure 5.1. Simulated Agena Body

6 DATA PREPROCESSING AND SAMPLING

As described in Sec. 5, the photometric data was received by us in digitized form on "REDDI-TAPES," recording Absolute Satellite Magnitude (M_S) as a function of time. These tapes were accompanied by graphical plots of the recorded data. Similar data was synthesized by us for the three tumbling Agenas described in Sec. 5.

Data samples were picked from the REDDI-TAPES for eventual input to the pattern recognition process. In most cases they were first subjected to some transformation--for example, some form of normalization, or transference from the time domain to the frequency domain by means of the Fourier transform. The sampling and preprocessing methods actually used are described below.

6.1 THE TIME DOMAIN

6.1.1 Excerpts from Data Plots

Excerpts from the plots of M_S versus time furnished by KMS Industries[3-7] for Satellites 5851, 4630 and 5587 are shown in Figs. 6.1 through 6.7. All this data exhibits very evident periodicity.

5851 and 4630 have definite quarter-periods of $\sim$2.50 seconds. The period of 5587 is $\sim$1.20 seconds.

The plots shown for 4630 are of observations taken at three different sites on the same day. The variation in the apparent quality of the data is quite striking.

Some of the simulated data for the three Agenas is shown in Figs. 6.8 through 6.10. The differences in range of the absolute satellite magnitudes in the three cases should be noted since it affected our choice of preprocessing prior to pattern recognition analysis on Agena, 5851, and 4630 data samples.

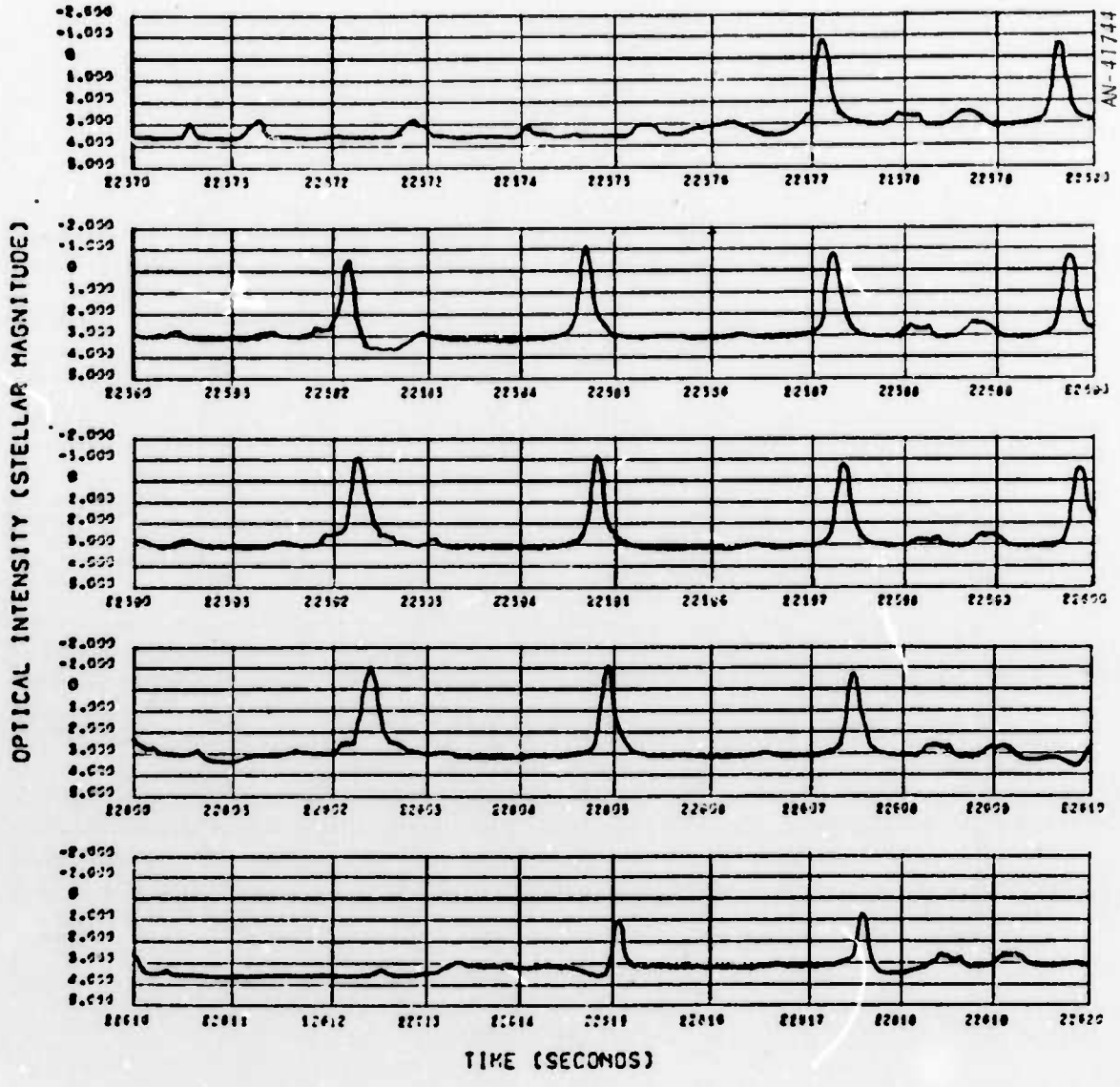


Figure 6.1. Cloudcroft Data, Satellite 5851, 17 May 1972

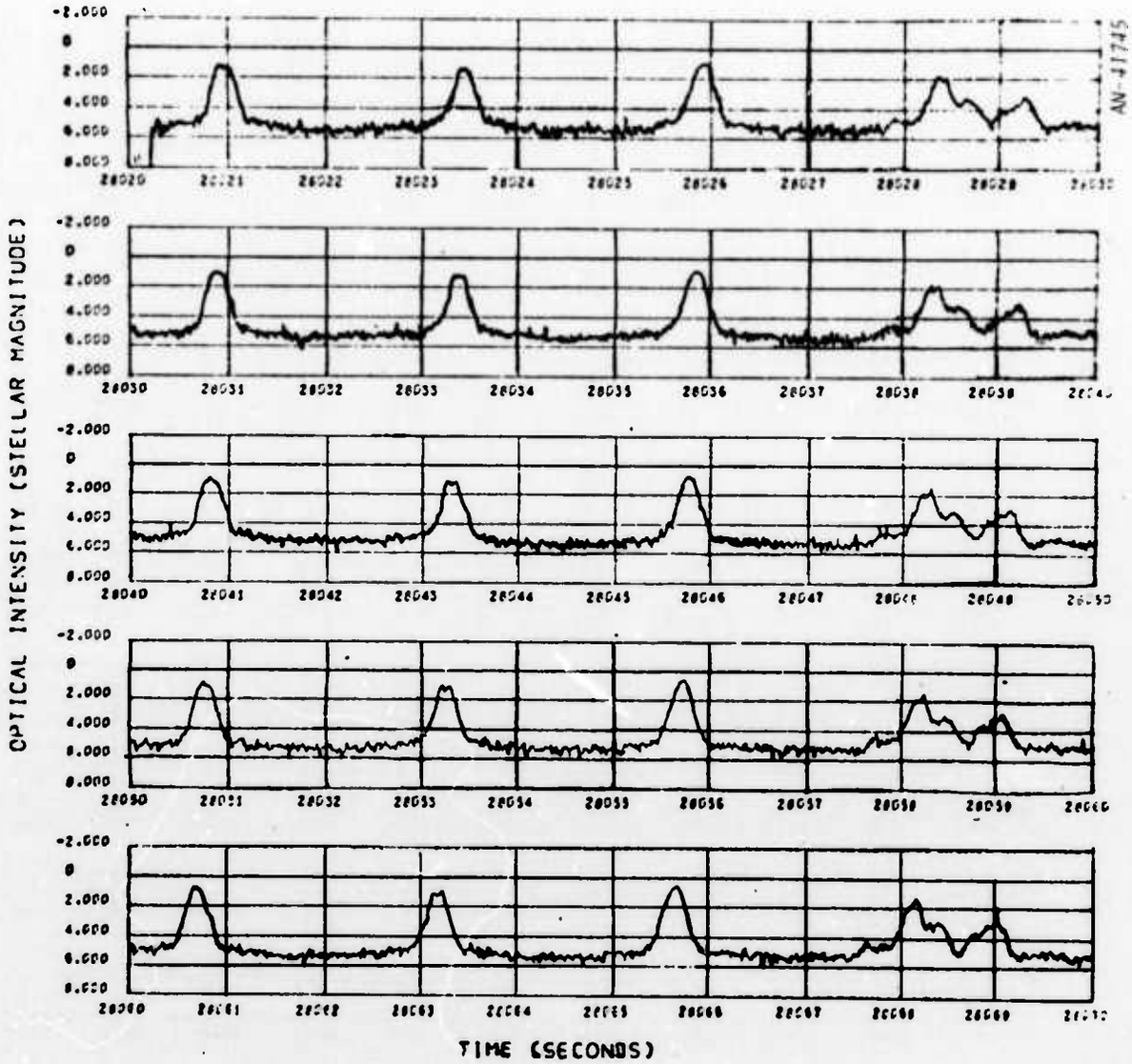


Figure 6.2. Cloudcroft Date, Satellite 4630, 10 November 1972

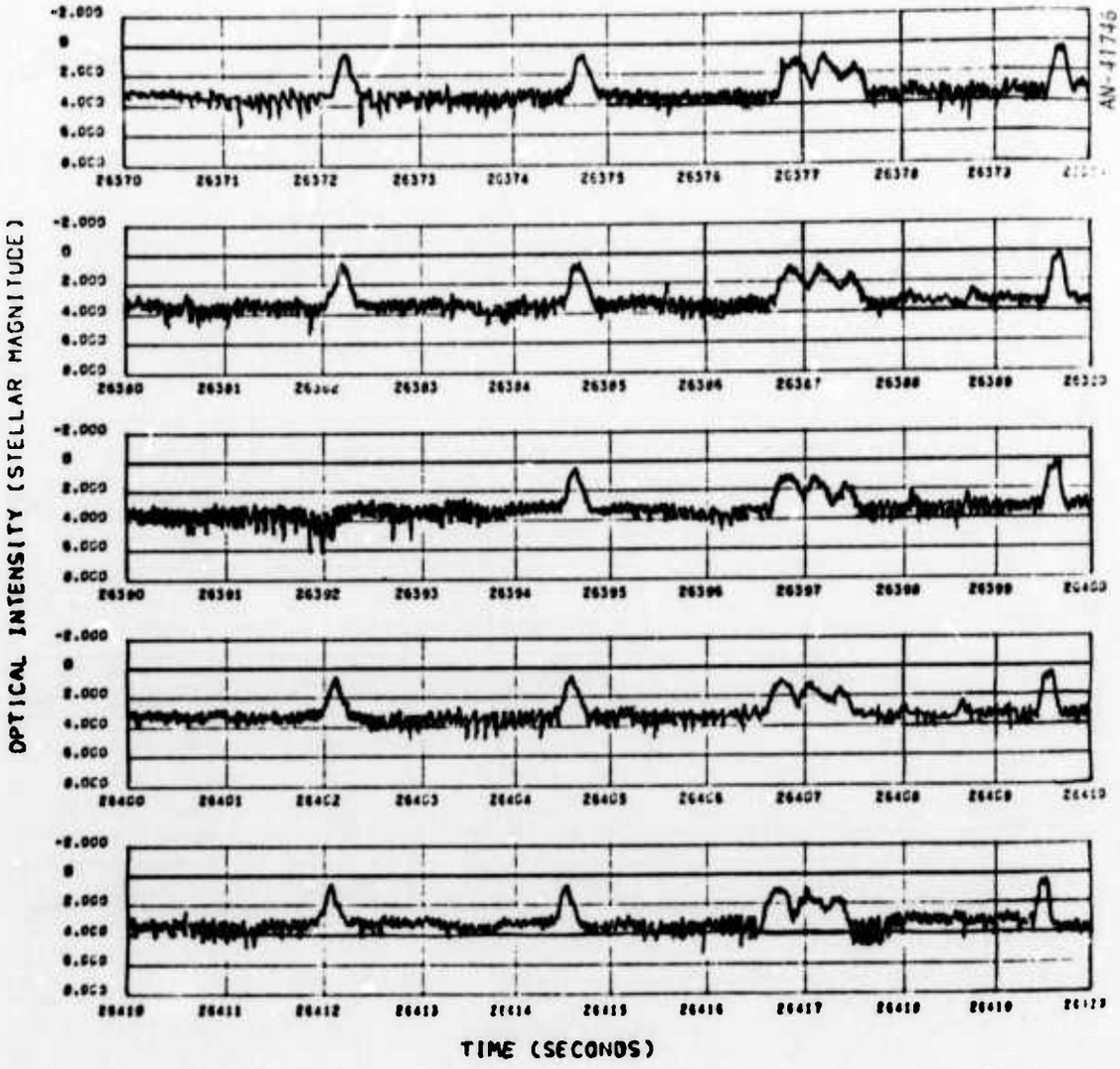


Figure 6.3. AMOS Data, Satellite 4630, 10 November 1972

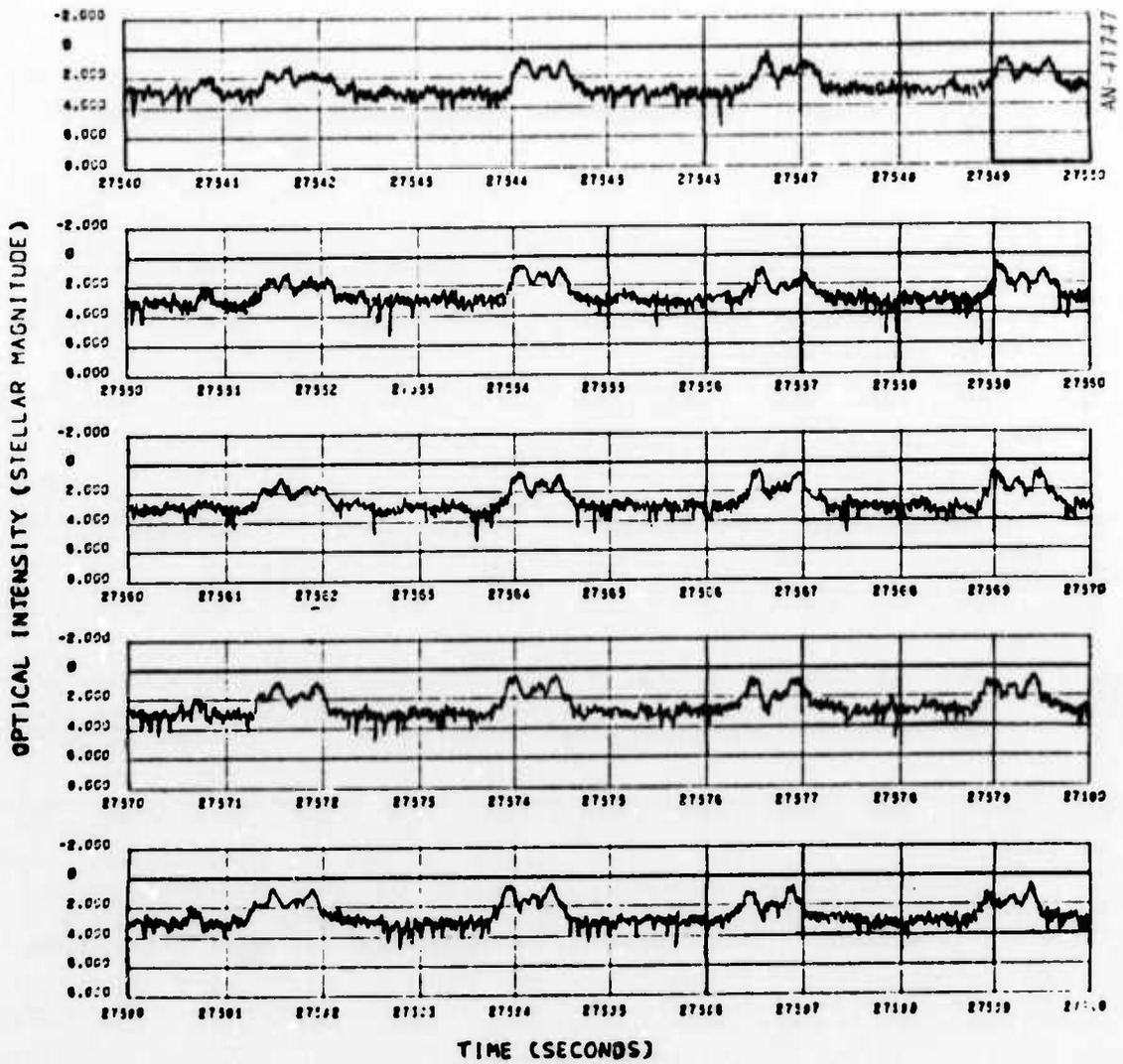


Figure 6.4. AMOS Data, Satellite 4630, 10 November 1972

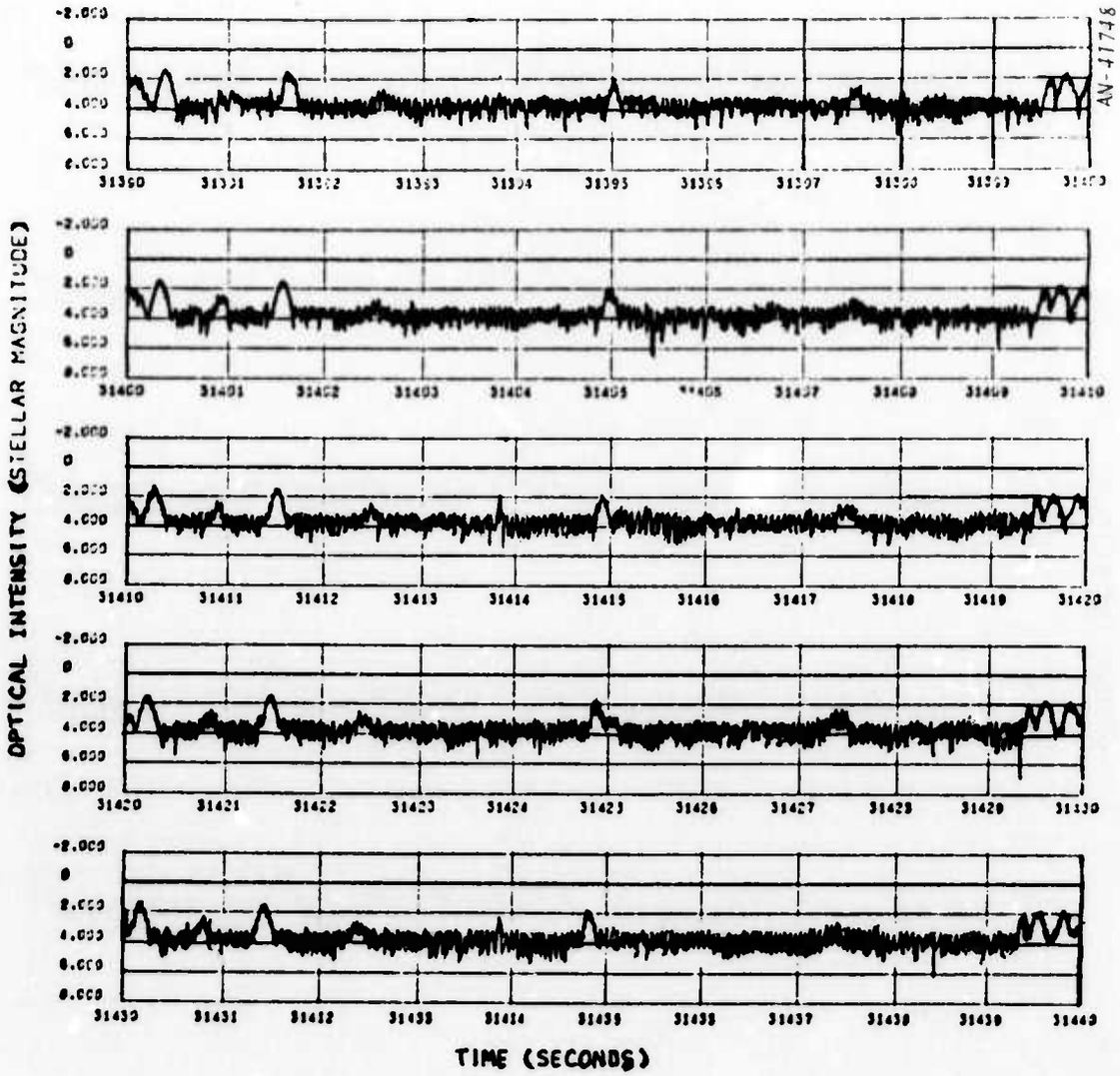


Figure 6.5. RML Data, Satellite 4630, 10 November 1972

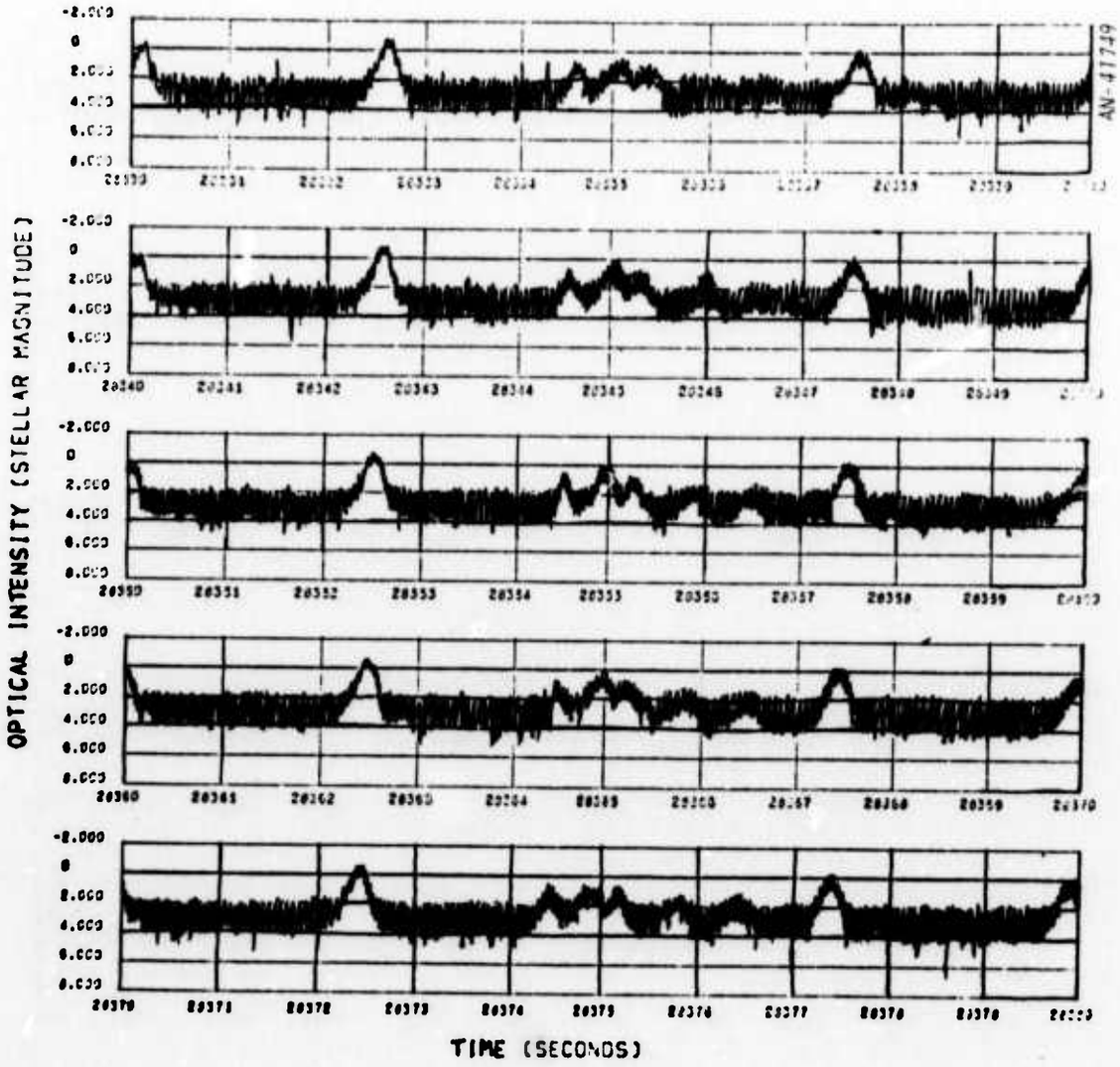


Figure 6.6. RML Data, Satellite 4630, 10 November 1972

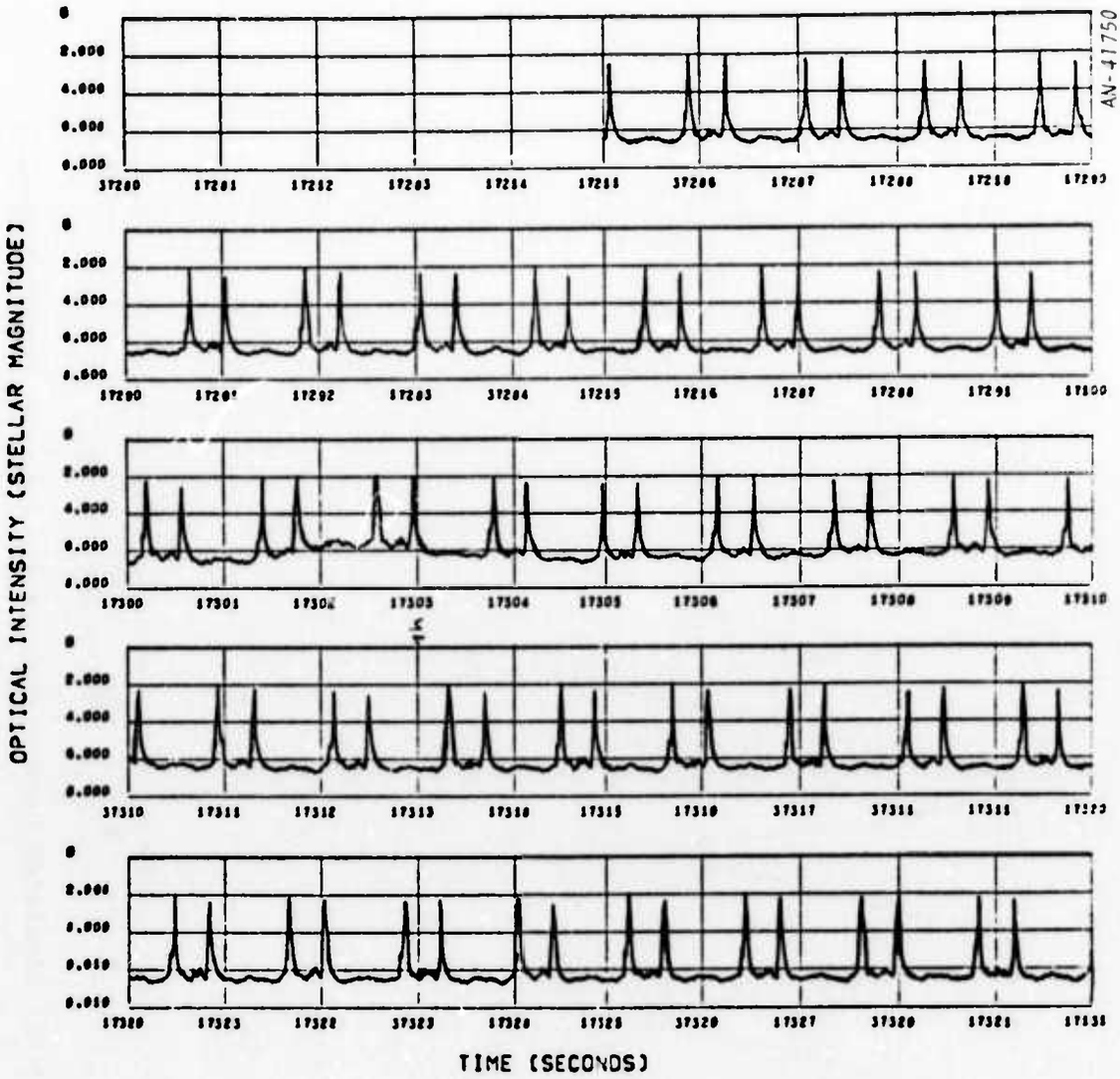


Figure 6.7. Cloudcroft Data, Satellite 5587, 9 May 1972

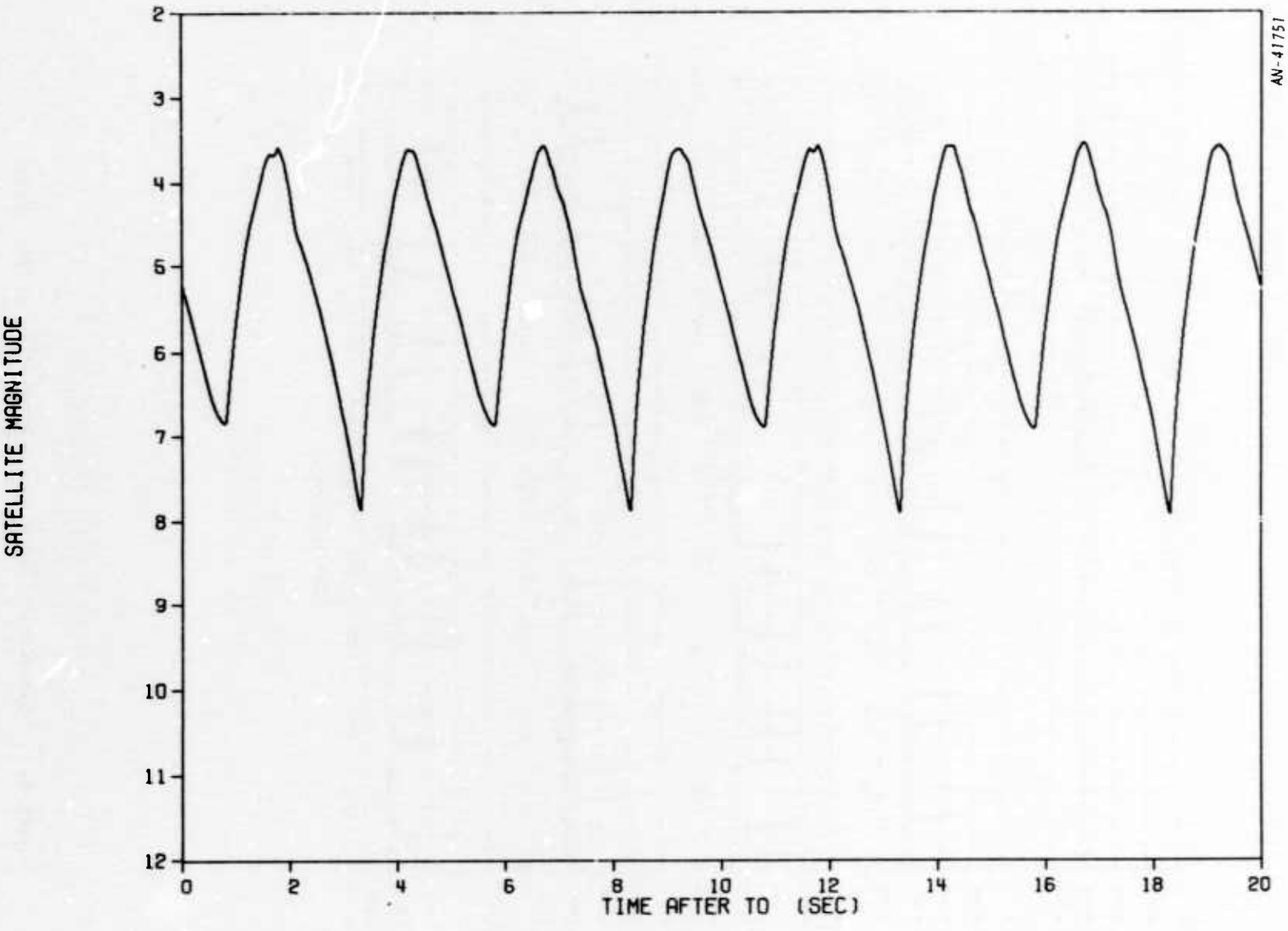


Figure 6.8. Simulated Passive Photometric Data for Tumbling Agena 1

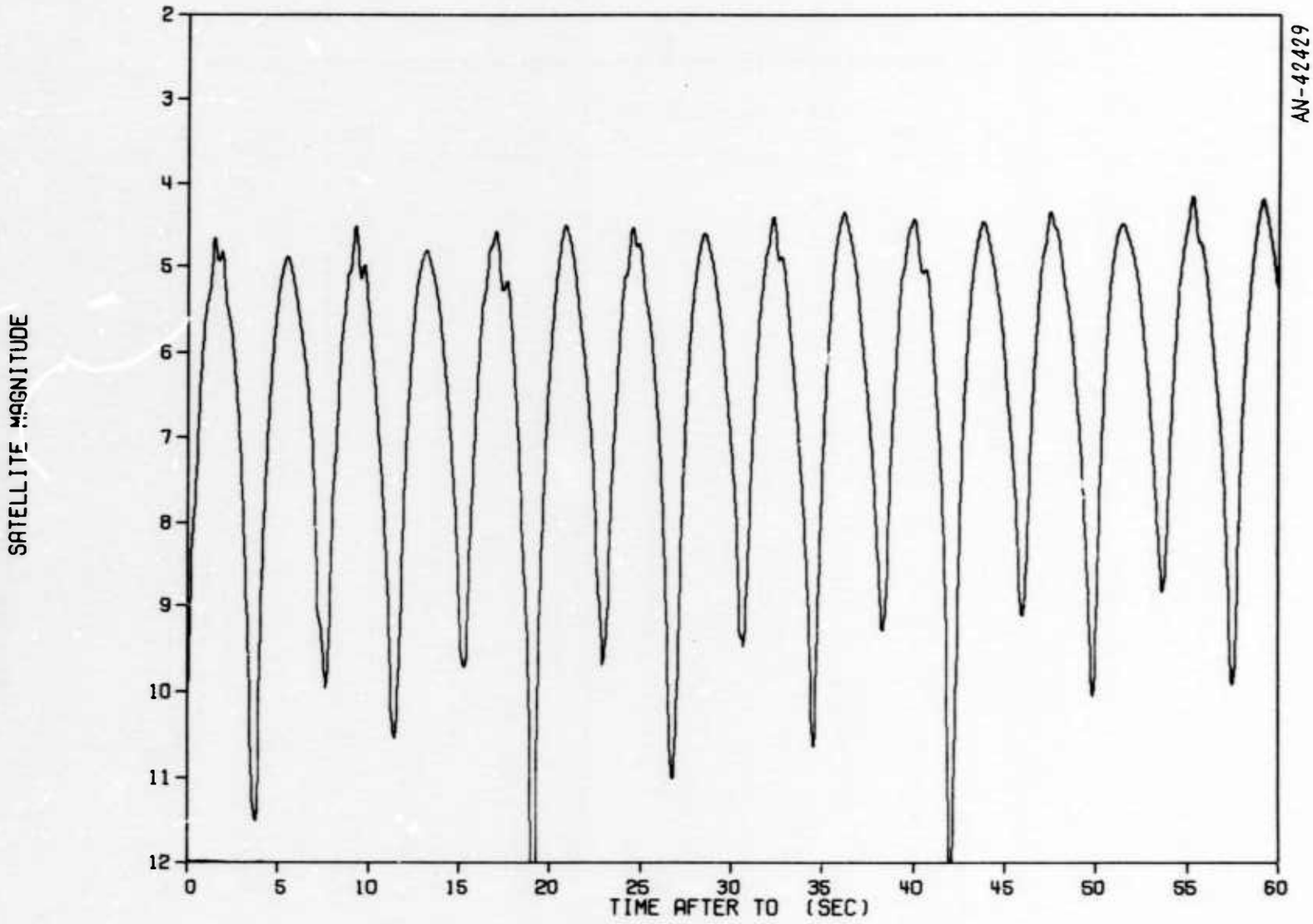


Figure 6.9. Simulated Passive Photometric Data for Tumbling Agena 2

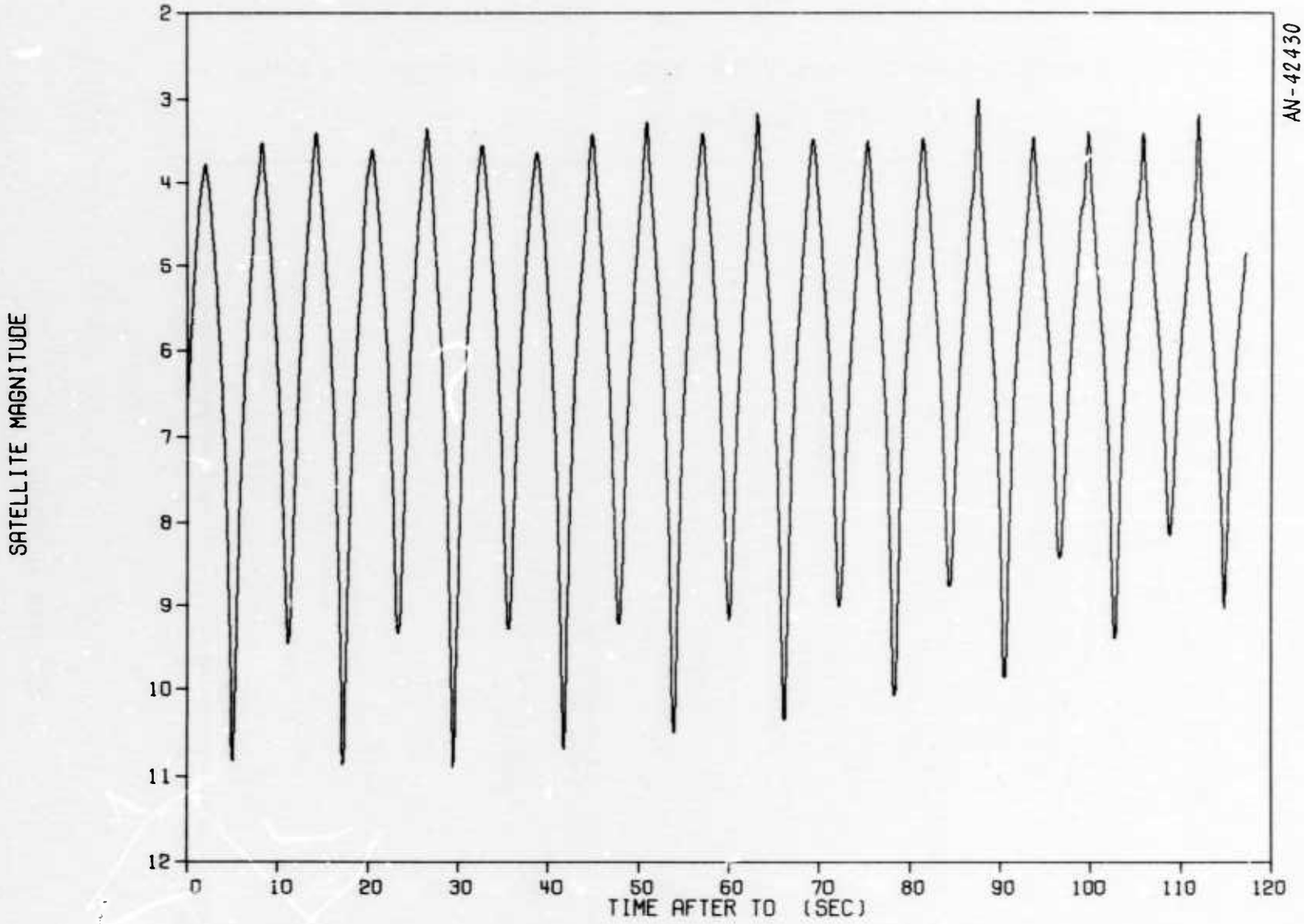


Figure 6.10. Simulated Passive Photometric Data for Tumbling Agena 3

6.1.2 Sampling and Preprocessing

Agena, 5851, and 4630 data were sampled in the time domain. 5587 data was not, since it did not become available until well after our interest had become focused on the frequency domain in which we were getting better classification results.

As stated earlier, 5851 and 4630 had the same period, while the Agena data was periodic by construction, with the periods chosen by ourselves. Thus to distinguish between Agenas and satellites on the basis of period would have been spurious. The satellite data generally exhibited one major peak per quarter period, while the Agena data generally peaked every half period. If we had merely normalized period to unity, we would still have been left with the very obvious discriminant of the number of major peaks per cycle. Taking still greater precautions to avoid possibly spurious discriminants, we chose the sample intervals to be one quarter period for the satellites and one half period for the Agenas, so that each generally contained just one peak, and normalized these sample intervals to unity.

Each data sample consisted of 50 values of M_S, represented as a point in a Cartesian space of 50 dimensions. Here and elsewhere, 50-dimensional samples were used because that is the greatest number currently allowed by the program.

There are significant differences in peak magnitudes between the three Agena cases which are probably due to aspect-angle-dependent differences in the degree to which the speculars were approached in each case; these are not, therefore, likely to be descriptive of a general observed Agena. Accordingly, when comparing Agenas and satellites, we chose also to normalize the fluctuation of M_S over the sample interval to a fixed value. This was done only in comparisons involving Agenas. It is not necessary (and indeed might obscure important classification features) where there is high confidence in the meaningfulness of the amplitude swings--as should be the case for the recorded satellite observations.

6.2 THE FREQUENCY DOMAIN

An obvious alternative to the sampling procedure just described is to take a finer "net" in the time dimension, approximate the Fourier transform on the basis of this net, and take the amplitudes at selected frequencies to be the components of a sample vector.

Two major questions arise:

1. What should be the length of a time interval on which a Fourier transform is computed?
2. What function of M_S do we wish to transform?

Our choices are defined and discussed below, together with a description of inputs to and outputs from the Fast Fourier Transform (FFT) routine.

6.2.1 Choice of Time Intervals

For each satellite, two primary considerations dictated the length of each time interval T on which an FFT was computed.

First, T had to be short enough that the available data contained more than a few intervals of length T. Second, T clearly had to be at least a single period of satellite rotation, and probably longer, to reduce the effect of any occasionally occurring disturbances in the data. We found that taking T equal to about twice the satellite period was a very good compromise. The precise value of T in each case was determined by the digitizing rate for the taped data, the choice of 2^{10} (= 1024) equispaced points in T as the basic net for computation of the FFT, and a constraint (introduced for a reason given in Sec. 6.2.4 below) that the ratio of T to the satellite period should be the same in all cases. Specifically,

- For 5581 and 4630, T = 20.48 s
- For 5587, T = 2.42 s

With these choices, we were able to use on the order of 30 intervals for each satellite/observation site pair.

6.2.2 Choice of Time-Function to be Transformed

The choice of what function of M_S to transform is, of course, somewhat arbitrary. M_S itself is a logarithmic unit, and we felt intuitively that it might be preferable to work with a more "natural" quantity such as Luminosity, I.

M_S and I are related as follows. The Apparent Photovisual Magnitude (m) of a source having the spectral distribution of the sun is related to its total Irradiance (E) just outside earth's atmosphere by the equation[8]

$$m = -2.5 \log(E) - 28.72$$

where E is measured in W/cm^2. This equation is, of course, only an approximation, especially so since no correction is made for atmospheric effects, but it is felt to be adequate for present purposes. The Radiant Intensity of an object at range r, when there is no attenuation, is given by[8]

$$I = r^2 E \times 10^{10}$$

where I is in watts per steradian and r is in km. Hence the relation between m and I is

$$m = 2.5 \log(I) + 5.0 \log(r) - 3.72$$

Since M_S is defined to be the value of m at a range of 1000 km, it follows at once that

$$M_S = -2.5 \log(I) + 11.28$$

or

$$I = 3.251 \times 10^4 \left(10^{-0.4M_S}\right)$$

In the frequency domain we could use the spectrum of I or of any function of I. After some experimentation, we finally decided to take the spectrum of $\sqrt{I}$, largely to reduce somewhat the great spread in the values in the spectrum of I itself. The use of this function should have negligible effect on the results, while improving the visual display and examination of the resulting spectra.

6.2.3 Computation of the Fast Fourier Transform and Exemplar Plots

The FFT of $\sqrt{I}$ was computed using IBM subroutine HARM. This requires that the number of sample points in the time domain over time interval T be an integral power of 2. After trying several different values we finally settled on using 2^{10} points for our Fourier Transforms; this is small enough to avoid excessive computation time yet long enough to provide good transform data. The time-domain points must be equispaced over the interval T, requiring interpolation whenever the data digitizing interval is irregular (Object 7) or incommensurate (Object 8). For Objects 4, 5, and 6, where the received data digitizing rate was 100 per second, we used only every second point, reducing the effective digitizing rate to 50 per second, in order to accommodate over two full periods ($\tau \simeq 10$ s) in each time-domain sample of 1024 points.

The spectrum of $\sqrt{I}$ consists of modulus and phase as functions of frequency f. In principle there is information concerning the target object in both the modulus and the phase functions. However, the limitation to maximum 50-point samples made it impractical to use both functions in the DISCRIMATON, and we had to make an initial selection. Various considerations suggested that there would probably be more usable information in the modulus function than in the phase function, so we decided to use the modulus of the Fourier Transform of $\sqrt{I}$ as our sample

data in the frequency domain. In doing this we were aware that some possibly significant information was being lost; it was hoped that we could include the phase function later, but ran out of time before this could be done.

The discrete spectral modulus produced by the FFT over 2^{10} points in the time domain has $2^9 + 1$ independent points in the frequency domain: these are the equispaced points at frequencies

$$f = 0, \Delta f, 2\Delta f, 3\Delta f, \ldots, 2^9\Delta f$$

where

$$\Delta f \equiv 1/T$$

The values of T used for the various objects were quoted in Sec. 6.2.1.

Of course, we could not use all of these data points: the DISCRIMATON can deal with 50-point samples at most, and furthermore the transform is likely to be distorted at the higher frequencies when interpolations have to be made. The samples were selected from among the 513 spectral moduli points in several ways as described below in Secs. 6.2.4 and 7.1.

Exemplar plots of the spectral modulus $|S(f)|$ of $\sqrt{I}$ as a function of $f(= n\Delta f)$ for various satellite/observation site pairs are shown in Figs. 6.11 through 6.16. Two of these figures are for the same satellite/site pair to illustrate the degree of difference that can occur from one time interval to another. In all cases, note that $|S(0)|$ appears at the top of the ordinate axis, and that there is a great difference between $|S(0)|$ and the average value of $|S(f)|$ over the range shown for $f > 0$. Also, the fluctuations in the value of $|S(f)|$ for $f > 0$ are sizable. These differences would be much greater if the spectral modulus of I had been used.

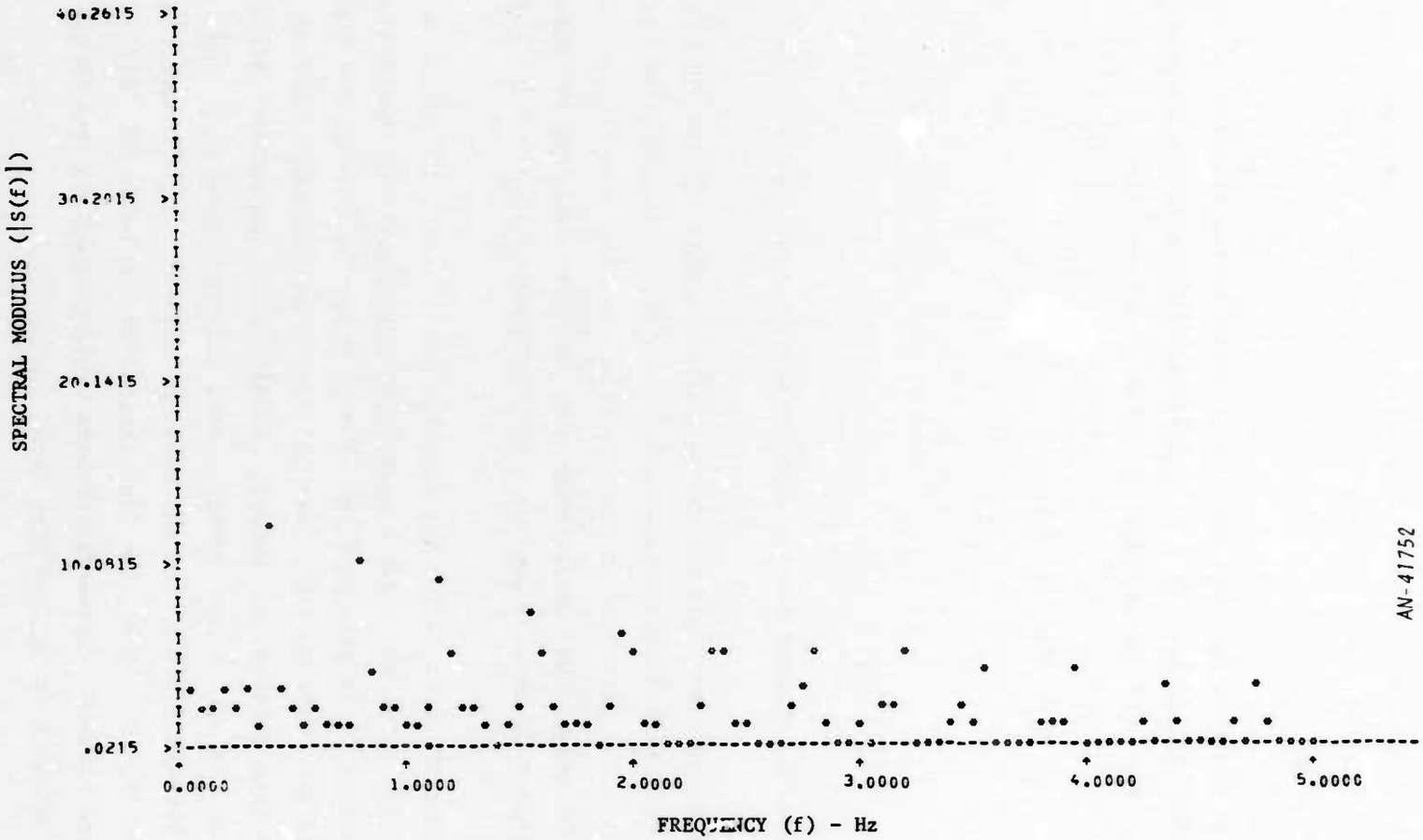


Figure 6.11. Exemplar Spectral Modulus of $\sqrt{I}$ for 5851 (Cloudcroft)

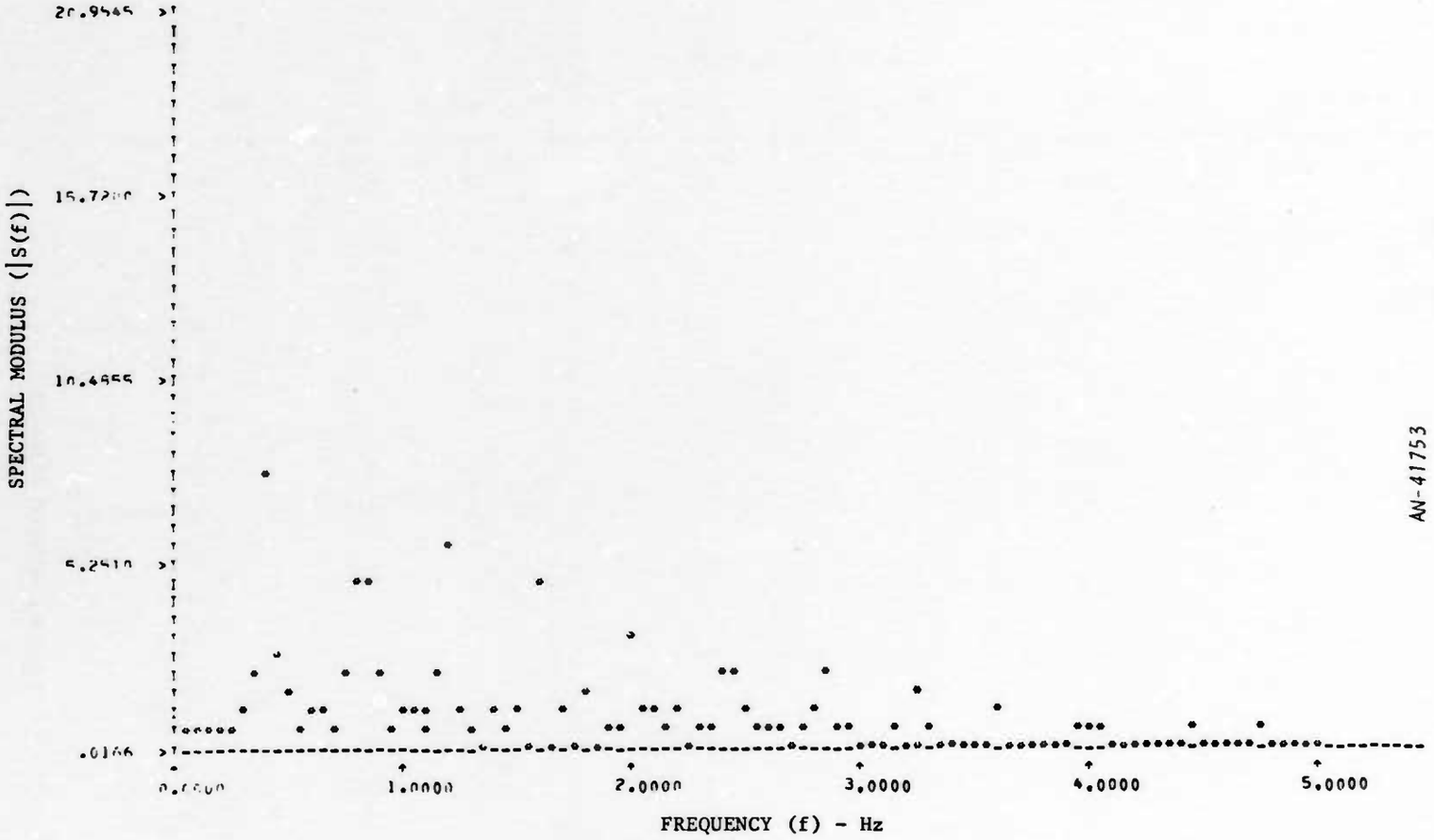


Figure 6.12. Exemplar Spectral Modulus of $\sqrt{I}$ for 4630 (Cloudcroft)

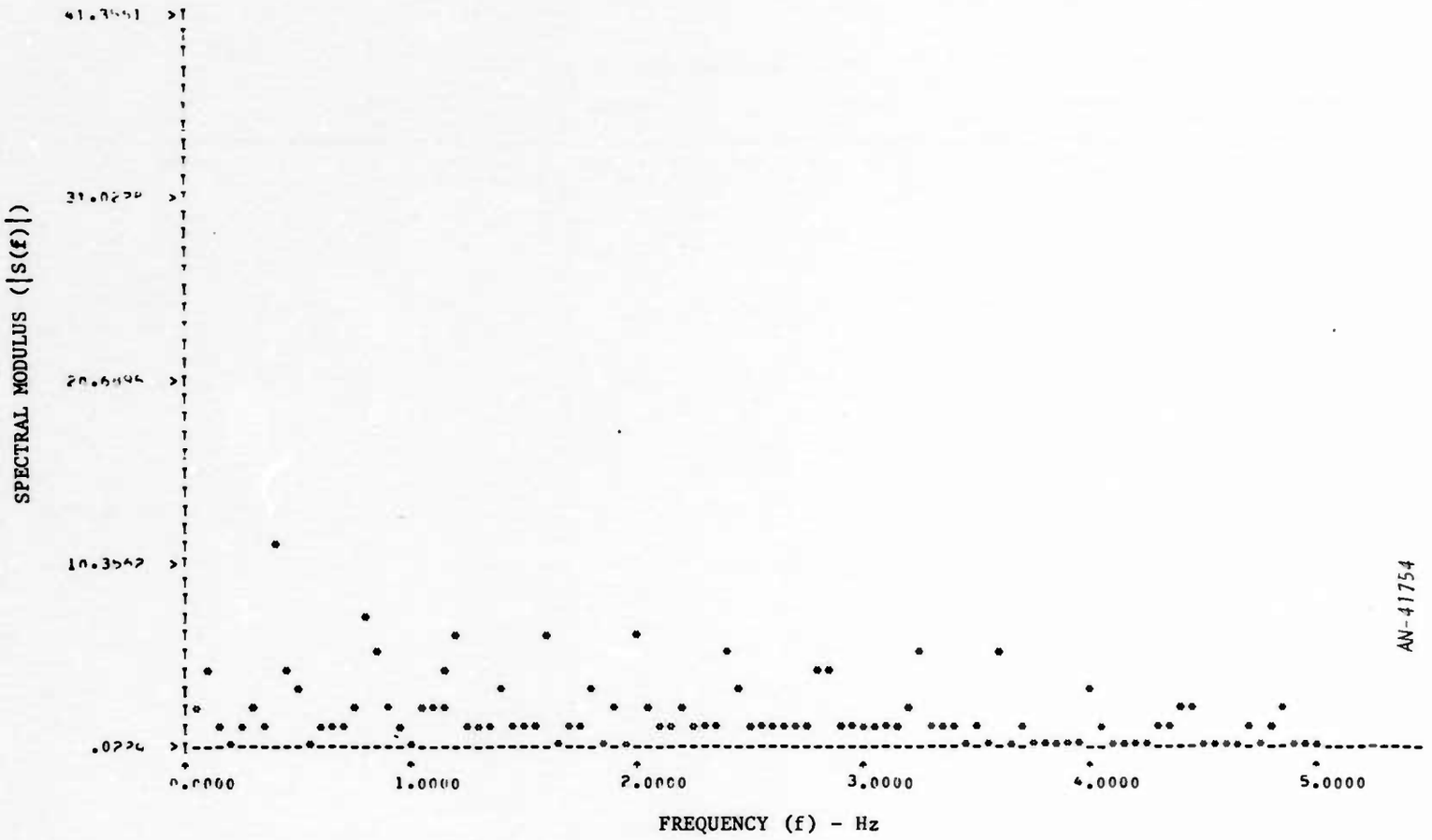


Figure 6.13. Exemplar Spectral Modulus of $\sqrt{I}$ for 4630 (AMOS)

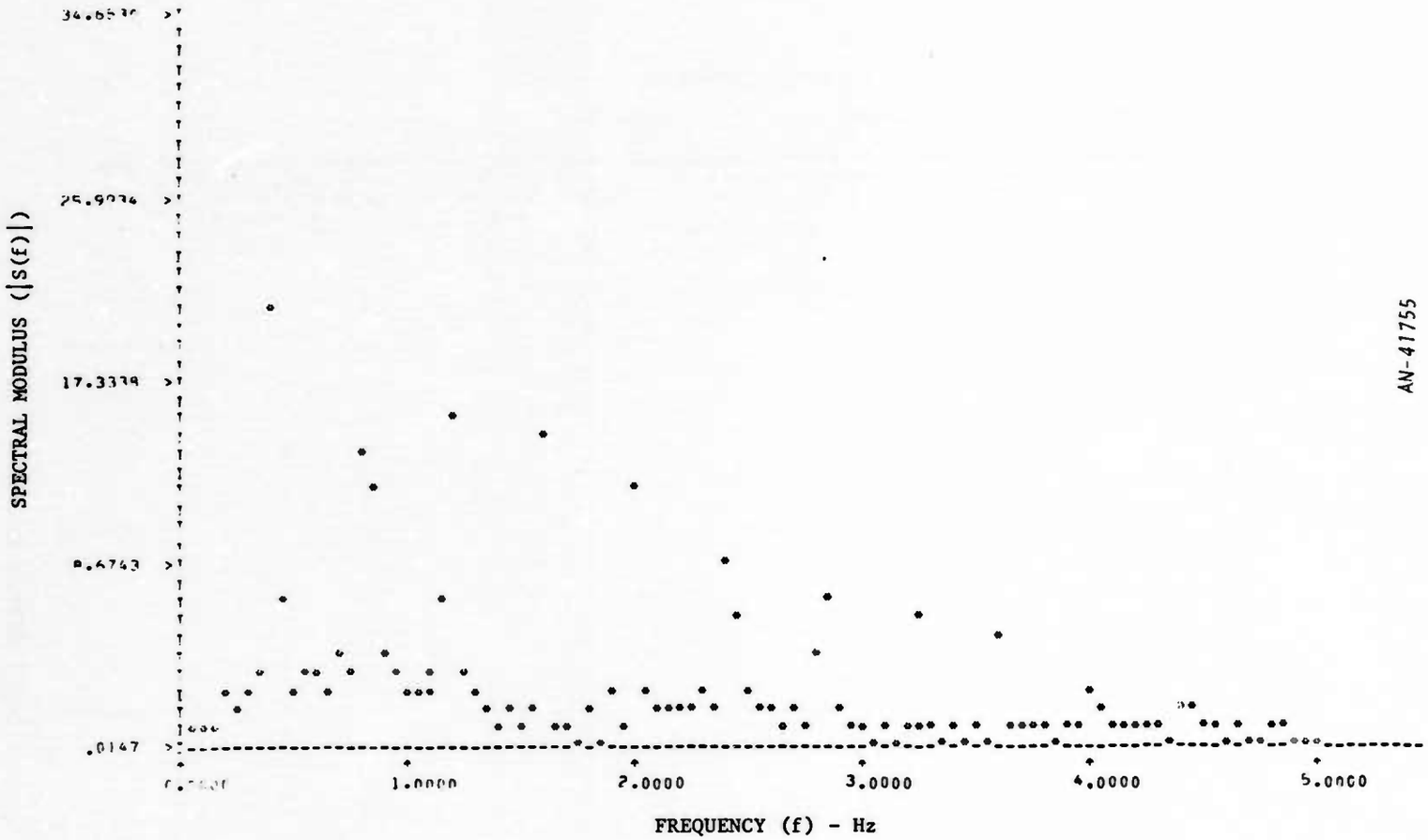


Figure 6.14. Exemplar Spectral Modulus of $\sqrt{I}$ for 4630 (AMOS)

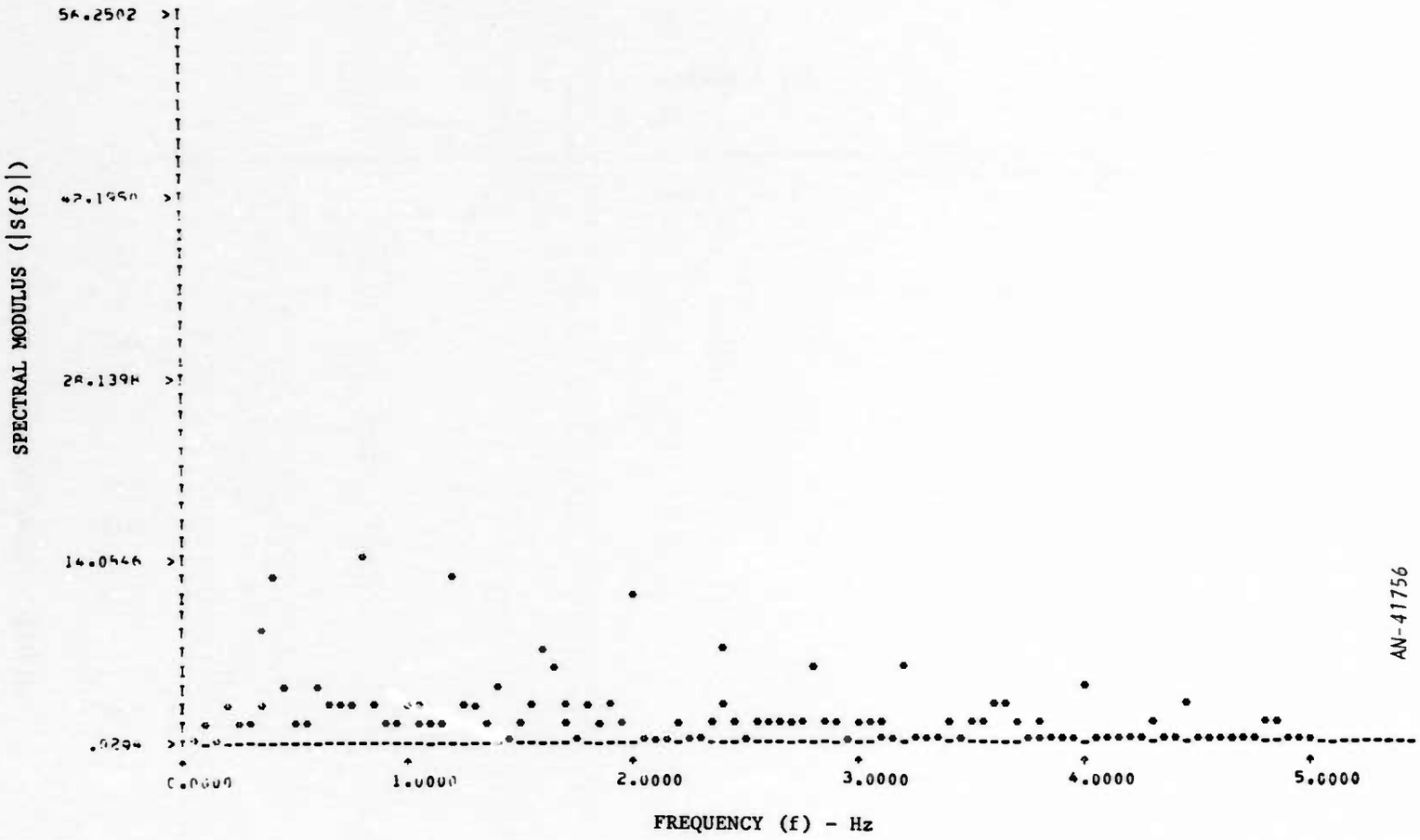


Figure 6.15. Exemplar Spectral Modulus of $\sqrt{I}$ for 4630 (RM)

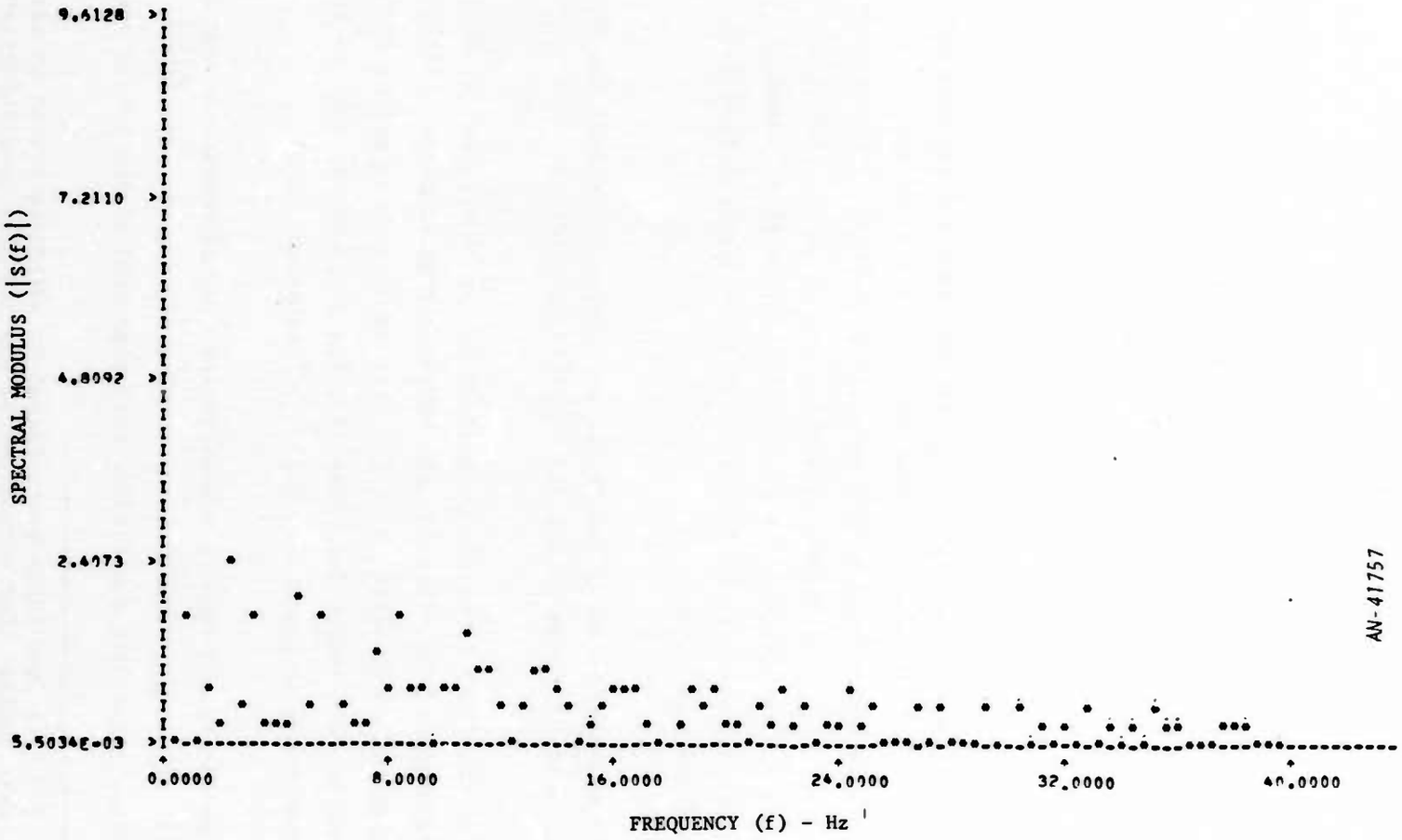


Figure 6.16. Exemplar Spectral Modulus of $\sqrt{I}$ for 5587 (Cloudcroft)

6.2.4 Sampling and "Normalization" of the Transformed Data

Each time interval of length T for which an FFT is calculated gives rise to a sample "vector" (or "point") in a Cartesian space of N dimensions, where N can, in principle, take any value from 1 to $(2^9 + 1)$ as desired. In practice we usually took $N = 50$, the largest value that the Principal Components Analysis program can currently accept. In these cases a sample vector is of the form

$$\left(|S(n_1\Delta f)|, |S(n_2\Delta f)|, \ldots, |S(n_{50}\Delta f)|\right)$$

where the values $n_1, n_2, \ldots, n_{50}$ are the same for all samples of this mode. Thus the specific value of Δf, which can vary from one object to another, does not influence classification.* The choice of $n_1, n_2, \ldots, n_{50}$ may affect classification, since no choice of 50 dimensions may contain all the significant information. However, we have no strong reason to believe that 50 dimensions was inadequate in the cases analyzed here.

In addition to using "unnormalized" sample vectors of the above form, we used two types of derived "normalized" vectors.

For the first type, the normalization was carried out by dividing each component of the vector by the corresponding value of $|S(0)|$. In particular, in cases where $n_1 = 0$, this amounts to rejecting the first component of each vector for classification purposes, as well as scaling other components to multiples of the "DC" component level.

For the second type of normalization, we calculated the mean value $(|\overline{S(0)}|)$ of $|S(0)|$ over all the samples for a given satellite/observation point pair and divided each component of the sample vectors

*Since $T = 1/\Delta f$ was taken proportional to satellite period in all cases, this implies that period per se was not used to distinguish between satellites; in effect it was "normalized out."

for this pair by $|\overline{S(0)}|$. Note that if T is such that the change in the angle subtended at the source between the sun and the observer is negligibly small, this kind of normalization should ideally be redundant, since M_S is so normalized by definition. The actual data did not conform to this condition, and we are tempted to infer that instrument calibration (at a particular site) sometimes drifted.

To remove any possibility that discrimination could occur because of instrument calibration drifts or of differences in calibration from one site to another, we also carried out the first type of normalization described above. This may be going too far; however, in the cases studied here, the type of normalization tended to improve classification. We do not know why.

Finally, some cases involved samples chosen by selecting only the first 16 peaks in the modulus function, including that at $f = 0$.

7 PATTERN RECOGNITION ANALYSIS

This section describes and discusses the results of applying the pattern recognition methodology of Sec. 5 to preprocessed data samples for Agenas and Satellites 5851, 4630 (three observation sites), and 5587. The main questions addressed are:

1. Can the Agenas be distinguished clearly from the satellites?
2. How do the various satellite samples cluster together
   a. in the time domain
   b. in the frequency domain

   when sampling retains as much information as is feasible?
3. What are the effects of various kinds of sample "normalization?"
4. What is the effect on classification of the amount and nature of the information retained in sampling?

Principal Components Analysis was carried out on the data samples in all cases. Calling the eigenvectors corresponding to the four largest eigenvalues of the correlation matrix X, Y, Z, U , respectively, plots of the projections of the sample points on the XY, XZ, XU, YZ, YU, ZU planes were printed by the computer, so that the results of the grouping algorithm could be compared with visual impressions of how the samples should be grouped. Several such plots are included here. In fact, for comparisons between Agenas and Satellites we used only the plots, since the grouping algorithm (LERNMOD) was obviously not needed. For comparisons between satellites, LERNMOD was certainly required, and was always utilized.

A summary of the principal findings, together with conclusions inferred from them, is given in Sec. 2.

The reader is reminded here of the caveat that our results could be biased in the direction of showing too great a dissimilarity between the data sets because sample sizes are too small in relation to the dimensionality of the samples (see Sec. 2).

7.1 NOMENCLATURE AND ABBREVIATIONS

For convenience we here give a reminder of the code numbers used to identify object/observation site pairs (Table 5.1), and introduce a compact notation in connection with frequency domain samples (Sec. 6.2.4).

The code numbers are given in Table 7.1.

For the frequency domain samples, recall that a sample point has the form

$$\left(|S(n_1\Delta f)|,\ |S(n_2\Delta f)|,\ \ldots\ldots,\ |S(n_k\Delta f)|\right)$$

TABLE 7.1

CODE NUMBERS FOR OBJECT/OBSERVATION SITE PAIRS

Code Number	Object/Observation Site	Actual or Simulated Data
1	Agena (Orbit 1)/Sulphur Grove	Simulated
2	Agena (Orbit 2)/Sulphur Grove	Simulated
3	Agena (Orbit 3)/Sulphur Grove	Simulated
4	S9450/Cloudcroft	Actual
5	S4630/Cloudcroft	Actual
6	S4630/AMOS	Actual
7	S4630/RML	Actual
8	S5587/Cloudcroft	Actual

where $k \leq 50$, and $0 = n_1 < n_2 < \ldots < n_k \leq 2^9$. We use the abbreviation

$$F;\ n_1,\ n_k,\ \Delta n$$

to denote a sample point

$$\left(|S(n_1\Delta f)|,\ |S\{(n_1 + \Delta n)\Delta f\}|,\ |S\{(n_1 + 2\Delta n)\Delta f\}|,\ \ldots\ldots,\ |S\{n_1 + (k-1)\Delta n\}\Delta f|\right)$$

where by definition,

$$n_k \equiv n_1 + (k-1)\Delta n$$

For example, F;0,98,2 denotes that the Fourier transform of data on all selected time intervals is sampled at the fifty frequencies $0,2\Delta f,4\Delta f,6\Delta f,\ldots,98\Delta f$, and that each time interval gives rise to a sample point with Cartesian coordinates

$$\left(|S(0)|,\ |S(2\Delta f)|,\ |S(4\Delta f)|,\ |S(6\Delta f)|,\ \ldots\ ,\ |S(98\Delta f)|\right)$$

7.2 COMPARISON OF TUMBLING AGENAS AND SATELLITES

This initial comparison involved objects with code numbers 1-5, with data sampled in the time domain and preprocessed as described in Sec. 6.1.2. The preprocessing removed any possibility that either rotation period or the maximum fluctuation of M_S over a sample interval would influence classification.

The number of samples taken from each of the data sets is given in Table 7.2.

The number of Agena samples (codes 1, 2, and 3) was small on account of usually small variation in signature from one cycle to the next. For each Agena, enough samples were chosen to be representative of the

TABLE 7.2

SAMPLE SIZES

Data Set Code Number	Number of Samples
1	7
2	15
3	18
4	51
5	119

variations seen to occur in usual plots of the entire data set. We also chose the code 4 and code 5 samples to span the (much greater) range of variations in the data evident to the eye.

The projections of the samples on the XY-plane (the plane defined by the eigenvectors corresponding to the two largest eigenvalues of the correlation matrix, and therefore the plane in which the spread of the samples is greatest), as plotted by computer, are shown in Fig. 7.1. Not all the points appear on this plot, since some may be too close to others to be distinguished separately. All the omitted Agena points are close to plotted Agena points, and omitted satellite points to plotted satellite points. The evident separation between Agenas and satellites seen in Fig. 7.1 is therefore unaffected by the points omitted from plotting.

While nominal sample size/dimensionality criteria for good classification are clearly not met in this case, we feel the figure is unlikely to be misleading because of the way the samples were selected. It is very probable that other samples from each data set would lie within the spread of those from the same set, shown in the figure. The separation between the simulated tumbling Agenas on the one hand and the

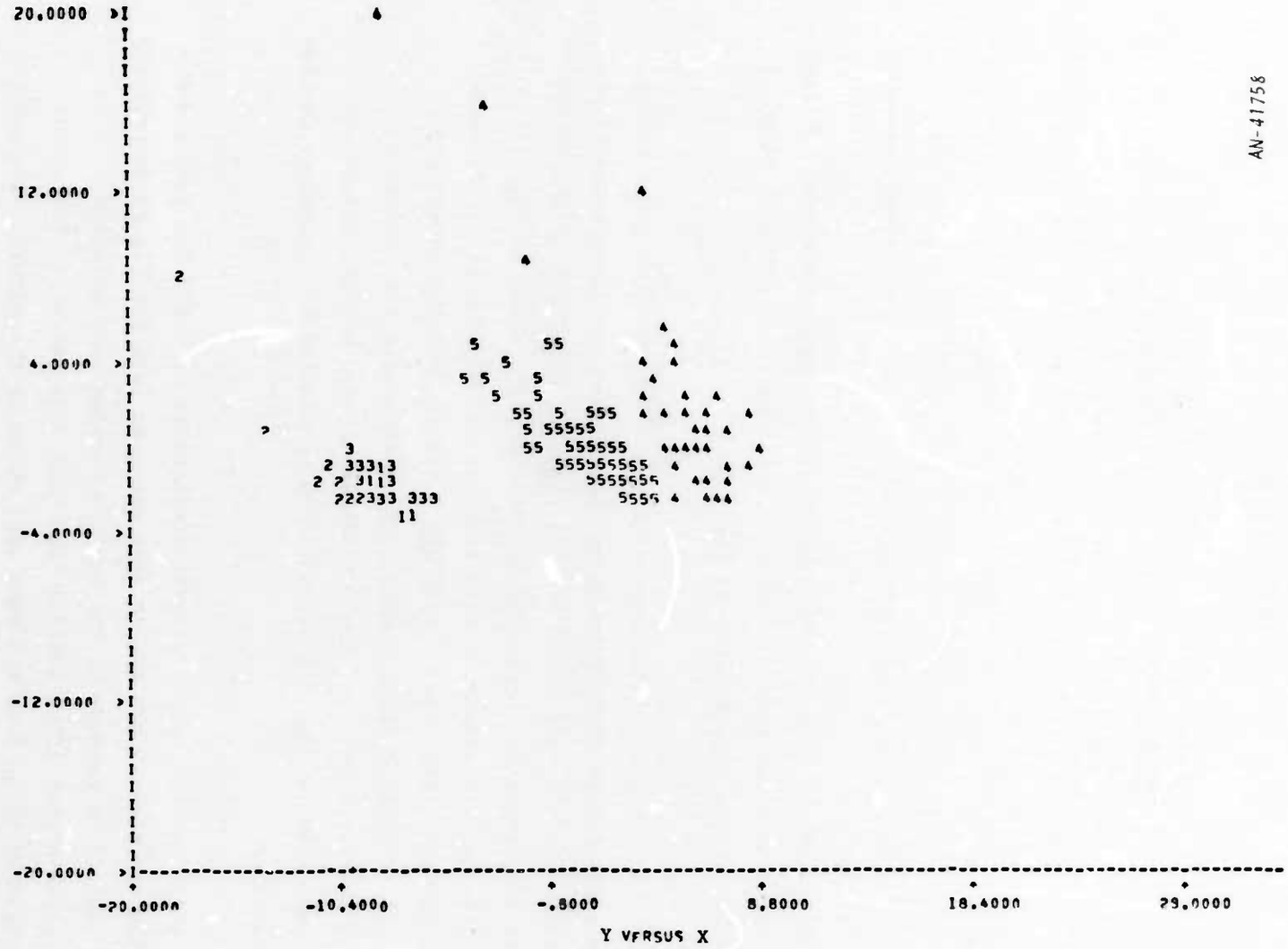


Figure 7.1. Projection of Agena, 9450, and 4630 Preprocessed Samples on the XY-Plane

observed satellites on the other hand is quite obvious. When such good separation is visually evident in the two-dimensional subspace plotted, it is unnecessary to use LERNMOD to perform the discrimination. However, in future cases when this occurs we shall also invoke LERNMOD to confirm the visual result in a higher-dimensional subspace.

7.3 COMPARISONS BETWEEN SATELLITES

Having verified that our methods could successfully distinguish satellites from tumbling rocket stages, we began making comparisons between the satellites themselves.

Comparisons were begun in the time domain. We switched to the frequency domain as soon as the FFT subroutine was successfully integrated into our computer program so that an early judgment could be made as to where further efforts should be concentrated. Results in the frequency domain proved to be rather better, and therefore almost all the results described below are for classification based on spectral analysis.

All the results in this section are based on samples having 50 dimensions (prior to Principal Components Analysis), the maximum that can be processed in our programs in their current form. The effect on classification of various forms of normalization of the samples is shown, as is sensitivity to the number of dimensions retained following Principal Components Analysis. We also comment on the effect of using the covariance matrix of the samples rather than the correlation matrix.

In the notation of Sec. 7.1, all the frequency domain results shown here are for F;0,98,2 frequency selections. Three other selections-- F;0,49,1, F;2,98,2 , and F;1,49,1 --were also tried; these gave somewhat inferior results.

The remaining question as to the effect of reducing the initial dimensionality of samples and of different ways of selecting the frequencies which define these dimensions is deferred to Sec. 7.4.

7.3.1 Comparisons in the Time Domain

Objects with code numbers 4 and 5 were compared, based on samples taken directly from the REDDI-TAPES, with no preprocessing prior to Principal Components Analysis. There are 55 code 4 samples and 117 code 5 samples, the difference in number being accounted for by the fact that there is considerably more code 5 data available and it exhibited an apparently wider variety of features than the code 4 data.

All the data was taken at Cloudcroft, and judged to be mostly of good quality.

Principal Component Analysis (PCA) found that there were 10 eigenvalues within a factor of 0.04 of the largest eigenvalue, and 13 within a factor of 0.01. Since the axes of the "error ellipsoid" are proportioned to the square root of the eigenvalues, this implies that of the 50 axes, 10 were longer than one-fifth, and 13 longer than one-tenth of the length of the major axis.

We are therefore strongly inclined to the belief that for purposes of estimating adequacy of sample size the dimensionality of the sample space need not be taken as 50, but could be reduced to about 13. If this point of view is accepted, then the sample sizes are large enough to avoid biasing the results toward too great a dissimilarity between the data sets.

Plots of the sample projections in the XY and XZ planes are shown in Figs. 7.2 and 7.3. Some points are left out in plotting because they would be over-printed on others. Since the 4's fall on other 4's and the 5's on other 5's, the loss of visual detail is minor.

LERNMOD found three groups, whose centers are indicated by X's in some of the figures, labeled M1, M2, M3 , respectively.* No

*Similar indications of group centers appear in some of the other figures.

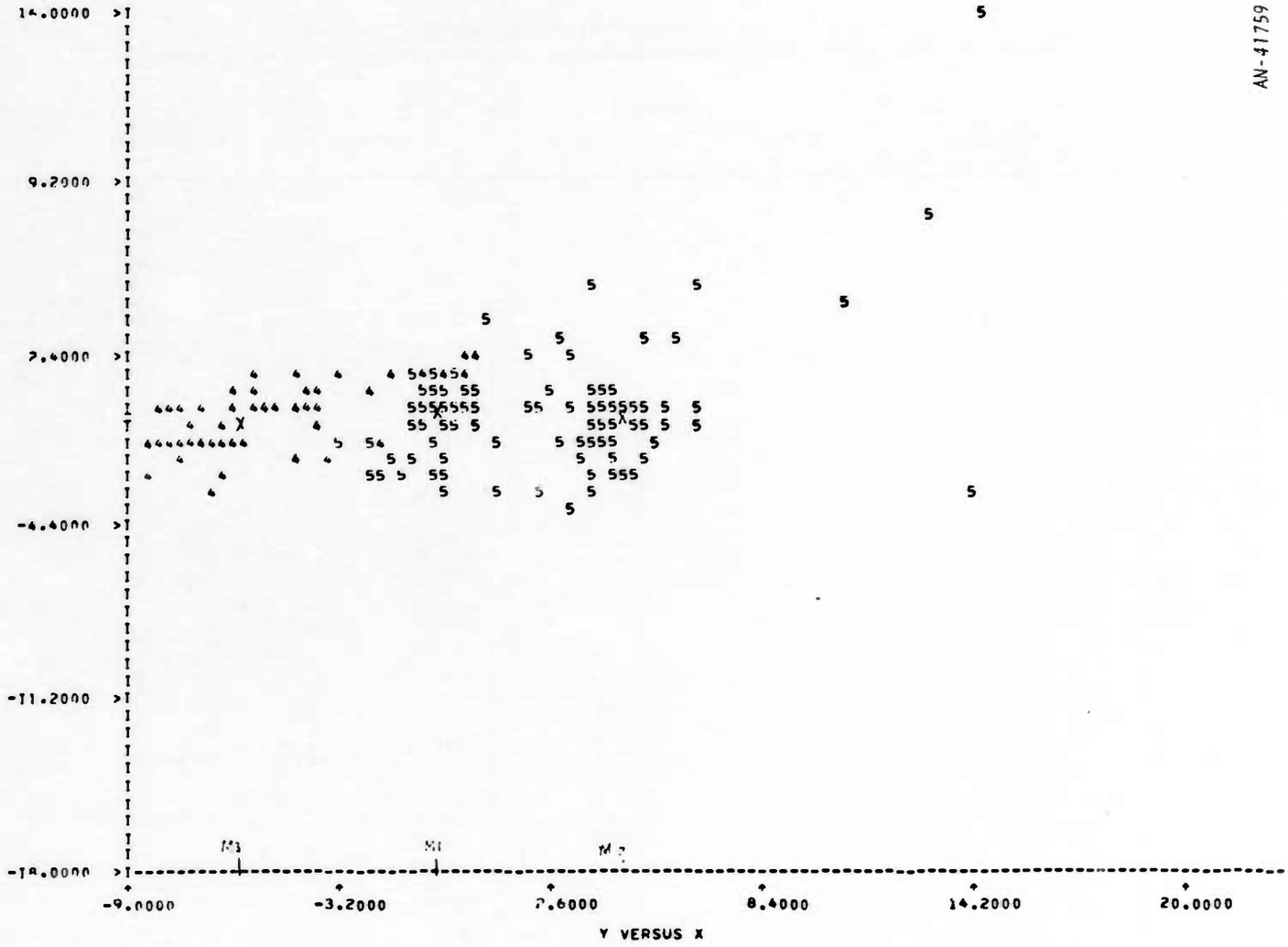


Figure 7.2. Projections of Preprocessed Samples in the XY-Plane: Code 4 and 5 Data

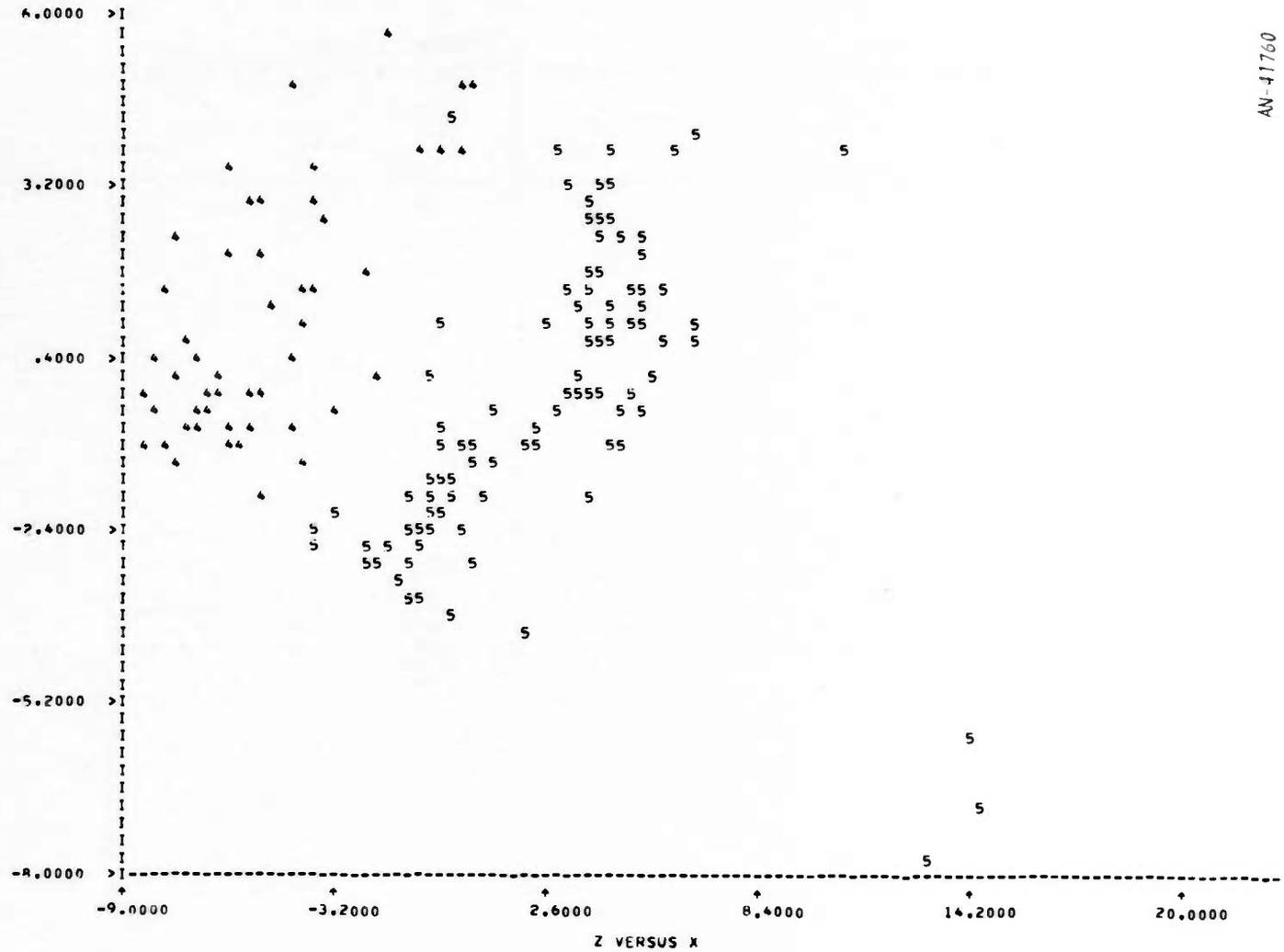


Figure 7.3. Projections of Preprocessed Samples in the XZ-Plane: Code 4 and 5 Data (Time Domain)

significance is to be attached to the group numbering. The number of samples of each type in each group is given in Table 7.3.

The value of the "similarity index" S(4,5) , as defined in Sec. 2.3, is

$$S(4,5) = \frac{1}{3}\left[\frac{1/55}{55/117} + \frac{6/55}{62/117} + 0\right] = 0.082$$

We conclude that code 4 and code 5 data are dissimilar. However, since their orbits are different, we cannot necessarily conclude that the two satellites are dissimilar.

7.3.2 Comparison of Unnormalized Samples in the Frequency Domain for Object Codes 4 and 5, and a Discussion of the Effect of Noise on Grouping

This comparison was made on the basis of 24 code 4 samples and 33 code 5 samples, all with F;0,98,2 frequency selection.

PCA found 12 eigenvalues within a factor of 0.04 of the largest eigenvalue, and 18 within a factor of 0.01, implying that 12 axes of the error ellipsoid were longer than one-fifth, and 18 longer than one-tenth of the major axis. Thus we incline to the view that sample dimensionality,

---

TABLE 7.3

NUMBER AND COMPOSITION OF SAMPLE GROUPS

	Number of Samples in Group	
Group Number	Code 4	Code 5
1	6	62
2	1	55
3	48	0

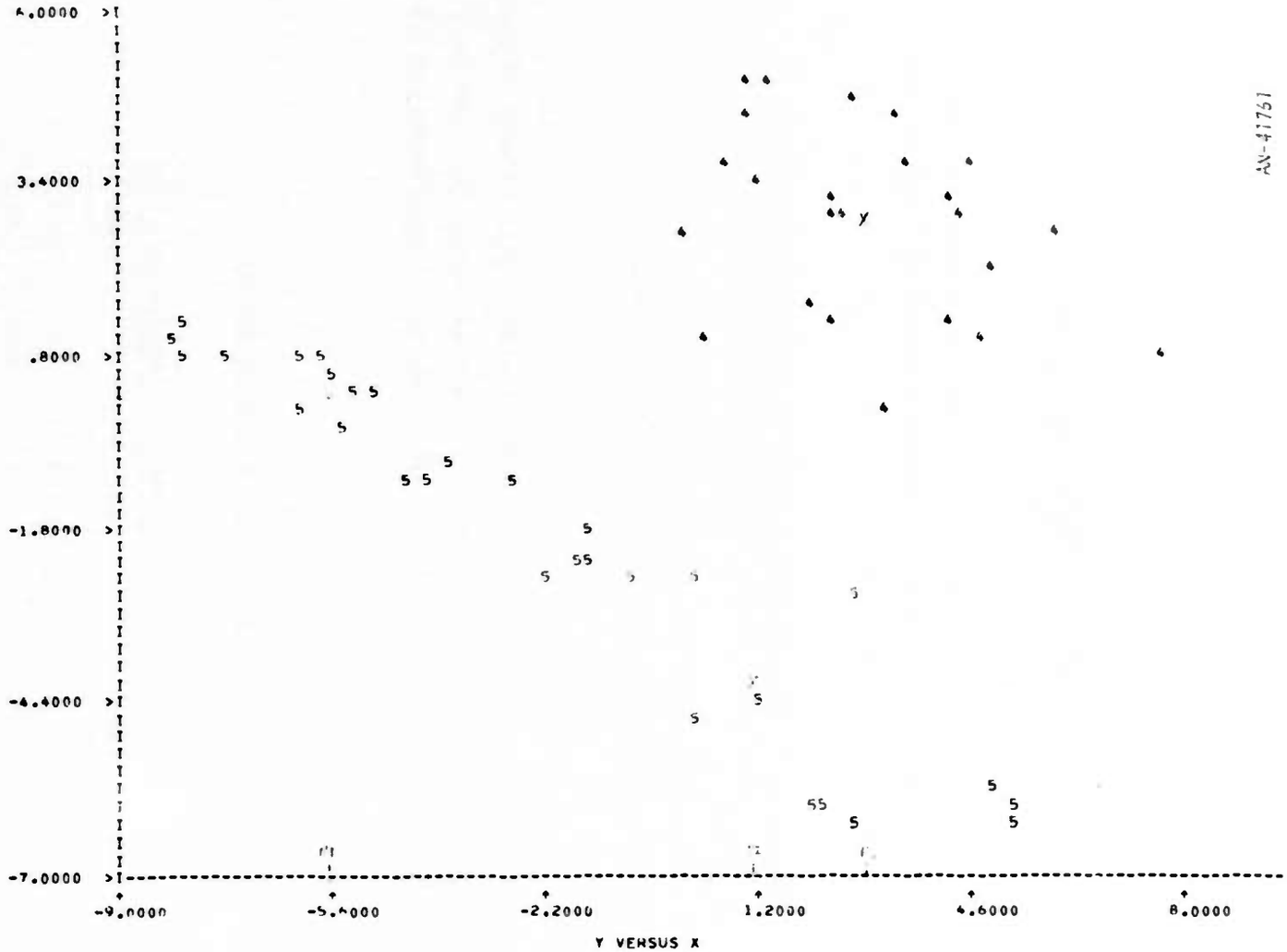


Figure 7.4. Projections of Preprocessed Samples in the XY-Plane: Code 4 and 5 Data (Frequency Domain)

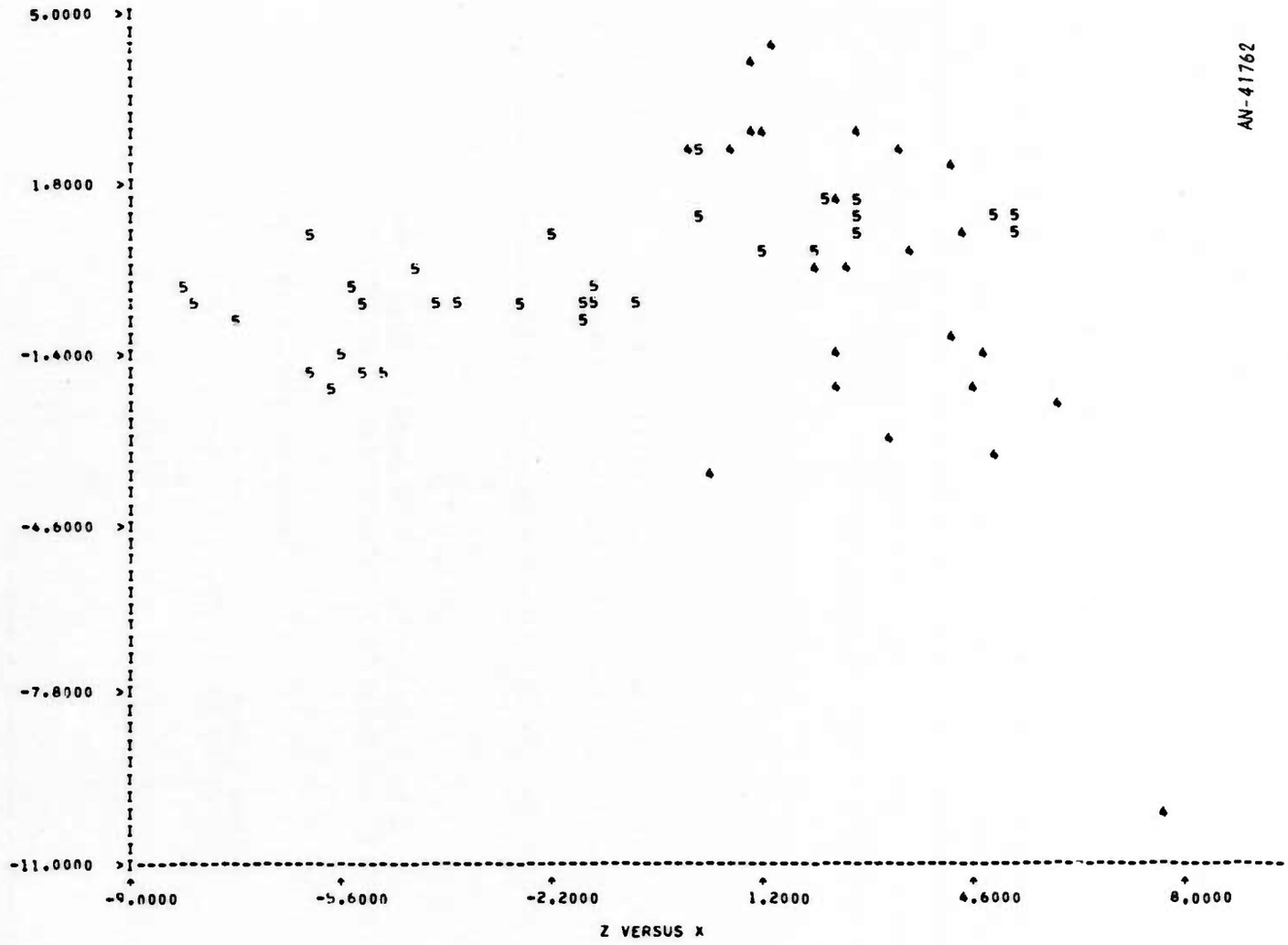


Figure 7.5. Projections of Preprocessed Samples in the XZ-Plane: Code 4 and 8 Data (Frequency Domain)

for purposes of estimating adequacy of sample size, is a lot closer to 18 than to the nominal value of 50, and could be as low as 12. Even so, our sample sizes are not large enough to avoid the possibility of results biased toward showing too great a dissimilarity between the data sets. The governing sample size of 24 was limited by the length of the code 4 data.*

Sample groupings found when two, three or four dimensions were retained following Principal Components Analysis were identical. Three groups were found, with compositions given in Table 7.4. The corresponding value of $S(4,5)$ is obviously zero, so again the conclusion is that code 4 and code 5 data are very dissimilar.

However, if we retain more dimensions when applying LERNMOD, the groupings change considerably, as revealed in Table 7.5. For five dimensions retained, the value of $S(4,5)$ is 0.14, and for six dimensions it is 0.051. Thus we would on the whole still regard the two classes of samples as quite different. The decreased ability to separate the data

TABLE 7.4

NUMBER AND COMPOSITION OF SAMPLE GROUPS; FOUR DIMENSIONS RETAINED AFTER PRINCIPAL COMPONENTS ANALYSIS

Group Number	Number of Samples in Group	
	Code 4	Code 5
1	0	16
2	0	17
3	24	0

*A sample in the frequency domain uses up approximately eight times as much of the data base as does a time domain sample.

TABLE 7.5

NUMBER AND COMPOSITION OF SAMPLE GROUPS; FIVE OR SIX DIMENSIONS RETAINED AFTER PRINCIPAL COMPONENTS ANALYSIS

Number of Dimensions Retained	Group Number	Number of Samples in Group	
		Code 4	Code 5
5	1	0	16
	2	24	14
	3	0	3
6	1	24	5
	2	0	17
	3	0	8
	4	0	3

sets indicates some deficiencies in LERNMOD, since it is evident that separations should remain the same or improve as more dimensions are kept. Consequently we always examine separations for various numbers of retained dimensions, and base our conclusions on the maximum separation found.

7.3.3 Comparison of Unnormalized Samples in the Frequency Domain for Object Codes 4 and 8

This comparison was made on the basis of the same 24 code 4 samples used in the previous comparison, and 30 code 8 samples, all with F;0,98,2 frequency selection. We remind the reader that period is not utilized in the comparison (see Sec. 6.2.4).

In this case, seven axes of the error ellipsoid were within a factor of one-fifth of the length of the major axis, and 11 within a factor of one-tenth. Projections of the samples in the XY-plane are shown in Fig. 7.6. The two sample sets are obviously very well separated (the 8's left out in printing fall within the group of 8's shown at X ≈ -5 and Y ≈ -0.1). Separations in the XZ- and XU-planes are similar. In the

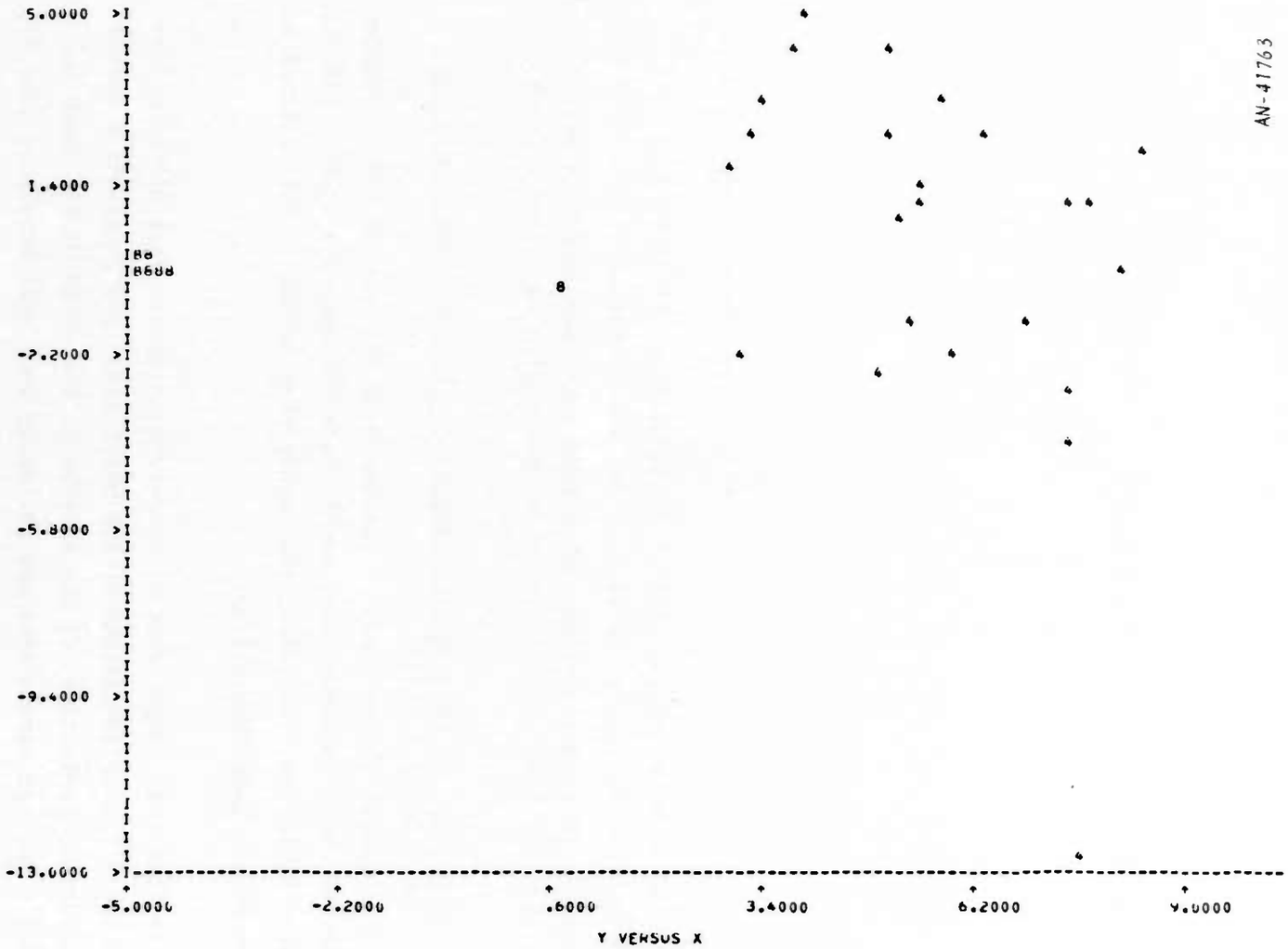


Figure 7.6. Projections of Preprocessed Samples in the XY-Plane: Code 4 and 8 Data (Frequency Domain)

YZ-plane, the 8's lie well embedded among the 4's, as shown in Fig. 7.7, and the same is true in the ZU-plane (in the figure, the 8's left out in printing lie within the central group of four at Y ≃ -0.3, Z ≃ -0.2). The tight clustering of the 8's in all subspaces, as compared with the far looser groupings for other object numbers, is interesting. This is apparently a characteristic feature of the data for this object whose cause can only be conjectured at this time; it should be possible to use cluster tightness as a discriminant, but it is not yet clear how to do this.

The sample groupings found in two, three, or four dimensions were identical. Two groups were found; group 1 containing 29 code 8 samples, and group 2 containing 24 code 4 samples together with one code 8 sample. The similarity index is

$$S(4,8) = \frac{1}{2}\left[(0/29) + (1/24)\right] = 0.021$$

We conclude that code 4 and code 8 samples are highly dissimilar. Again, this may, but does not necessarily, imply that the two satellites are dissimilar.

7.3.4 Comparison of Unnormalized Samples in the Frequency Domain for Object Codes 5 and 8

This comparison is based on the 33 code 5 samples and 30 code 8 samples previously used (F;0,98,2 frequency selection).

Two axes of the error ellipsoid were within a factor of one-fifth of the length of the major axis and 11 within a factor of one-tenth. Figure 7.8 shows the sample projections on the XY-plane, and Fig. 7.9 the projections in the XZ-plane (once again, the 8's left out in printing fall onto the tight clusters of printed 8's). Separation between the two sets in the XY-plane seems quite clear-cut, but in the XZ-plane, the distinction is not so clear.

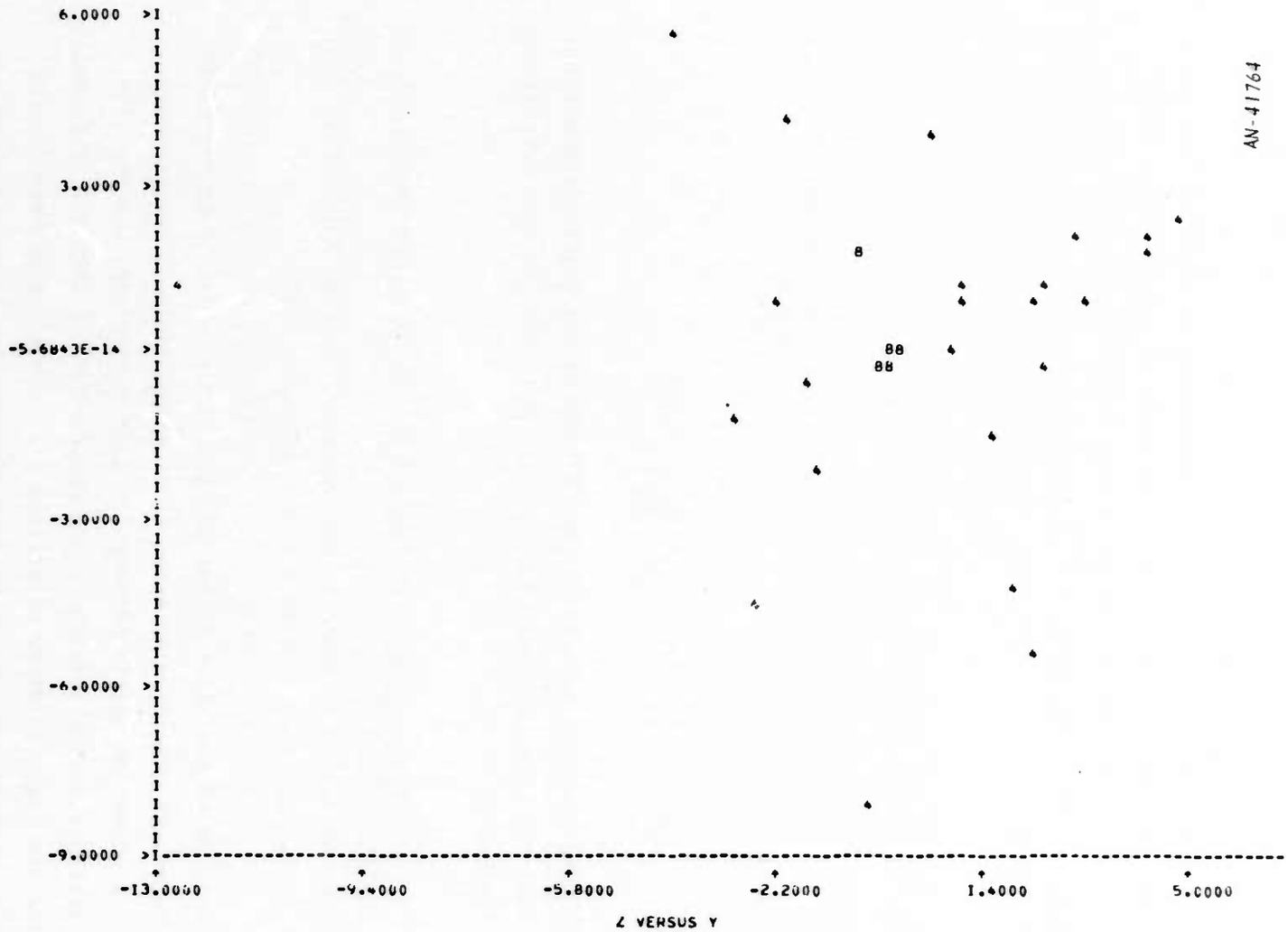


Figure 7.7. Projections of Preprocessed Samples in the YZ-Plane: Code 4 and 8 Data (Frequency Domain)

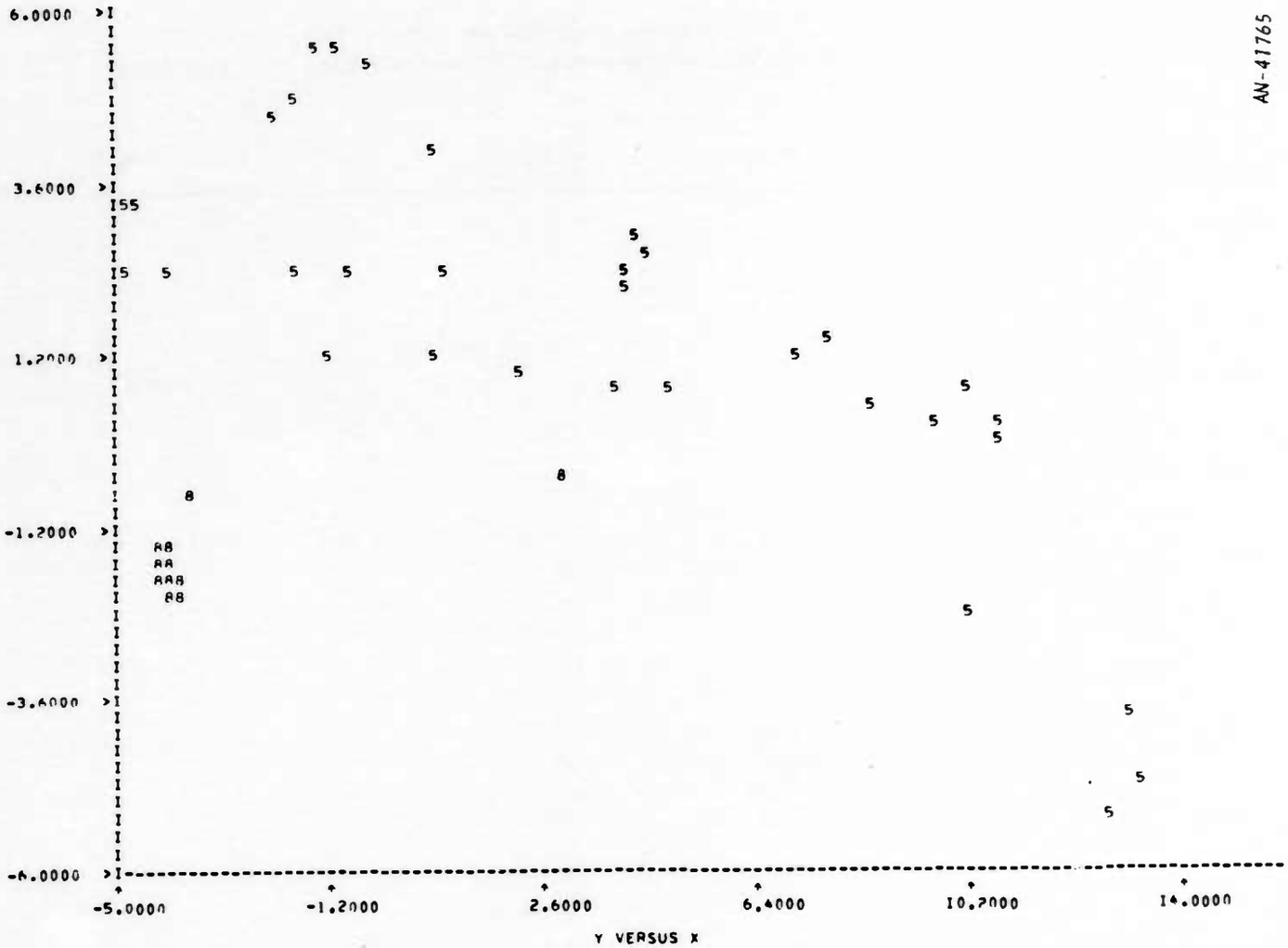


Figure 7.8. Projections of Preprocessed Samples in the XY-Plane: Code 5 and 8 Data (Frequency Domain)

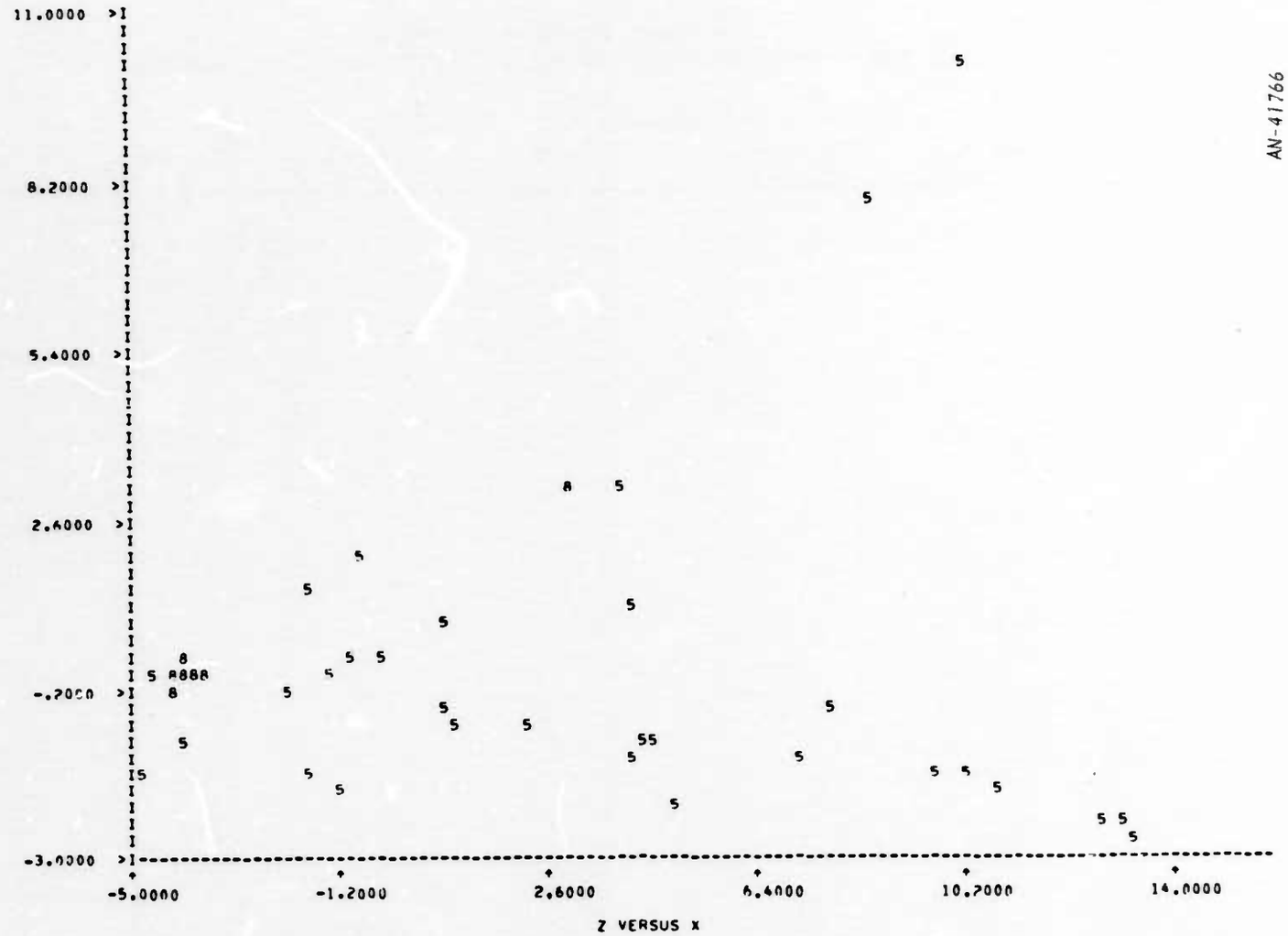


Figure 7.9. Projections of Preprocessed Samples in the XZ-Plane: Code 5 and 8 Data (Frequency Domain)

One might therefore suspect that the LERNMOD would find quite different groupings if only two dimensions rather than three were retained. In fact, its groupings were almost identical, as indicated by Table 7.6.

When three dimensions are retained, the similarity index $S(5,8)$ is given by

$$S(5,8) = \frac{1}{2}\left[\frac{16/33}{29/30} + \frac{1/30}{17/33}\right] = 0.28$$

Thus there is a moderate degree of similarity between the two sample sets.

7.3.5 Comparison of Unnormalized Samples in the Frequency Domain for Object Codes 4, 5, and 8

Having compared these objects in pairs, we felt it might be instructive to see what groupings would emerge when all three sets of samples were treated together.

Projections of sample points in the XY, XZ and YZ planes are shown in Figs. 7.10 through 7.12, respectively. Comparison of these plots with those given previously for the data sets taken in pairs reveals

TABLE 7.6

NUMBER AND COMPOSITION OF SAMPLE GROUPS; TWO OR THREE DIMENSIONS RETAINED AFTER PRINCIPAL COMPONENTS ANALYSIS

Number of Dimensions Retained	Group Number	Number of Samples in Group	
		Code 5	Code 8
2	1	15	29
	2	18	1
3	1	16	29
	2	17	1

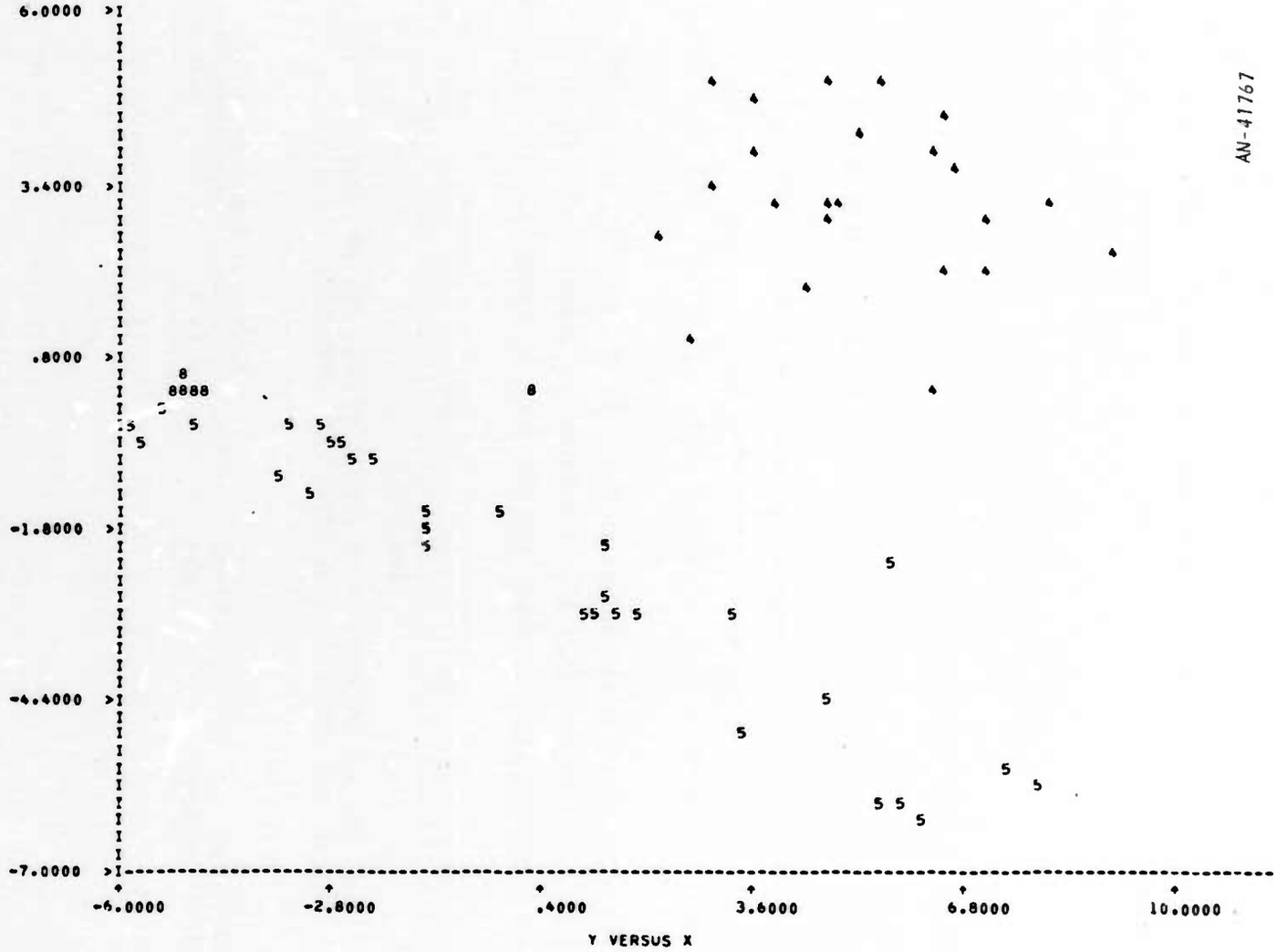


Figure 7.10. Projections of Preprocessed Samples in the XY-Plane: Code 4, 5, and 8 Data (Frequency Domain)

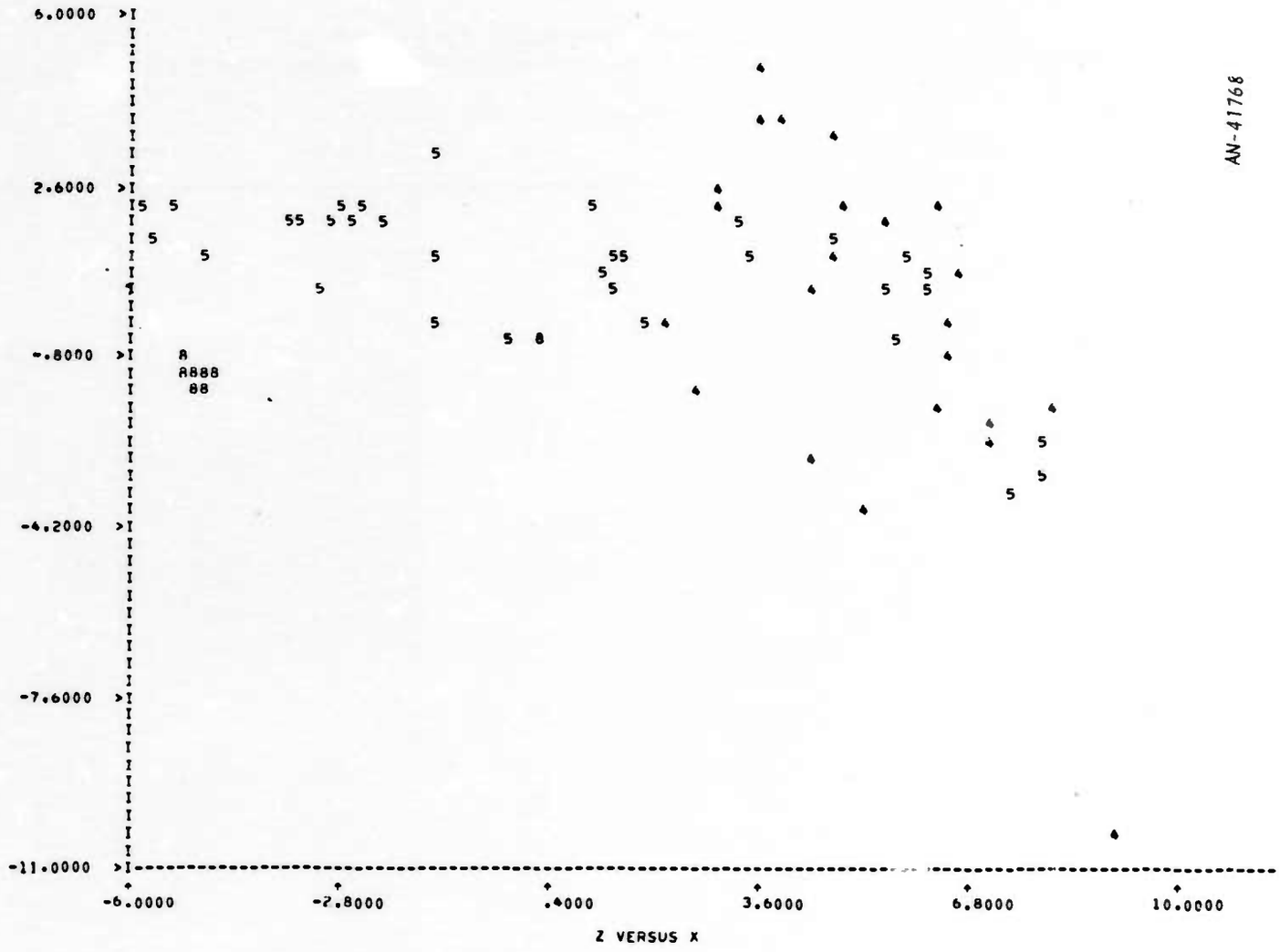


Figure 7.11. Projections of Preprocessed Samples in the XZ-Plane: Code 4, 5, and 8 Data (Frequency Domain)

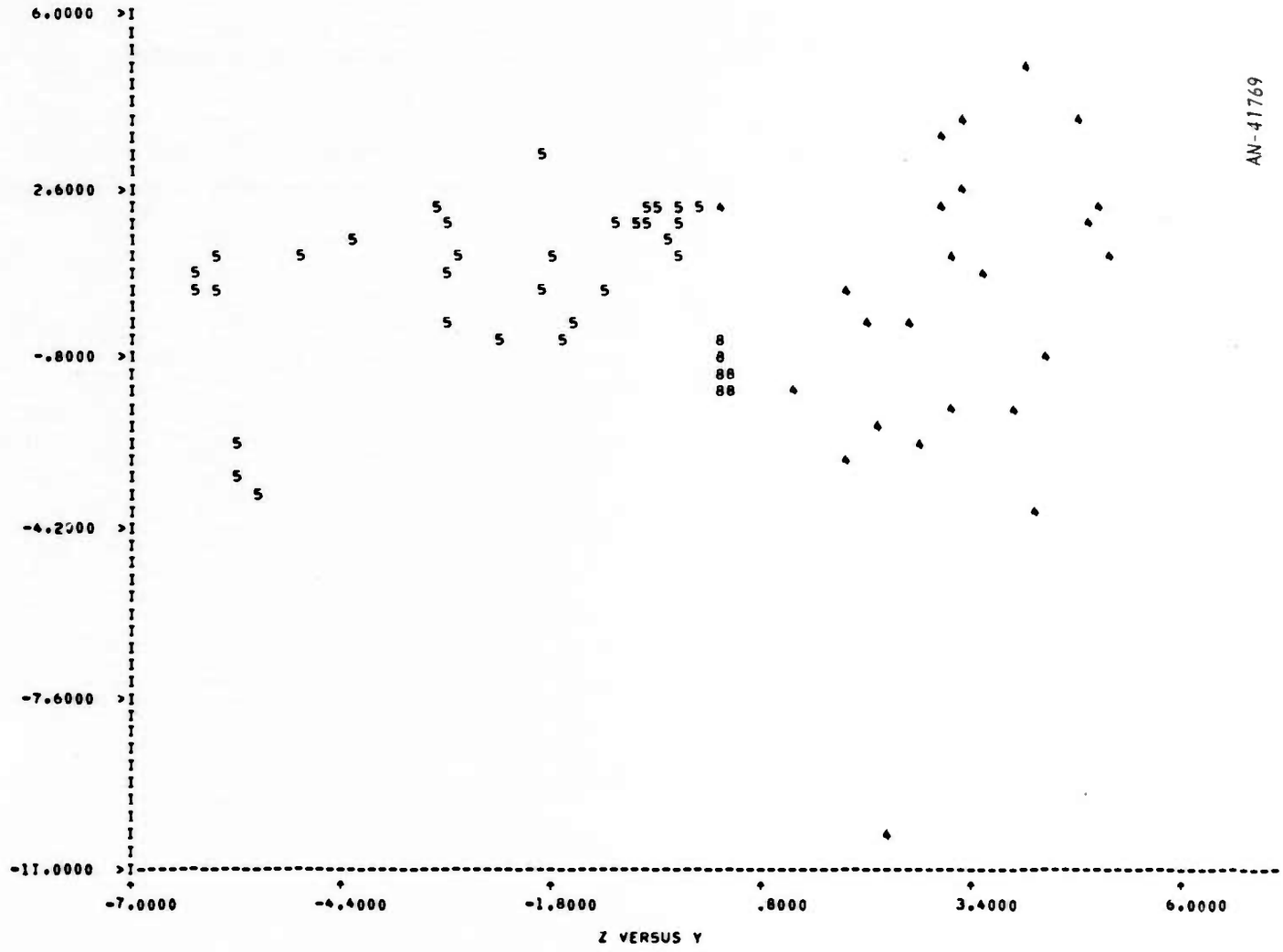


Figure 7.12. Projections of Preprocessed Samples in the YZ-Plane: Code 4, 5, and 8 Data (Frequency Domain)

that in the XY- and XZ-planes the relative location of the 4's and 8's has changed little, but the locations of the 5's relative to the 4's and 8's has changed very substantially. In the YZ-plane, the 4's have moved rather more to the right of the 8's than formerly, and again the relative location of the 5's has changed substantially. Thus the 4-8 separation dominates the directions of the longest axes of the error ellipsoid.

The groupings found in three and four dimensions are given in Table 7.7. The corresponding similarity indexes are shown in Table 7.8. The similarity index values found previously, for the data sets examined in pairs, are $S(4,5) = 0$, $S(4,8) = 0.021$, and $S(5,8) = 0.28$. Comparison of Table 7.8 with these previous results reveals that (1) if we retain three dimensions, $S(4,5)$ and $S(4,8)$ remain virtually unchanged, but $S(5,8)$ increases from 0.28 to 0.50; and (2) if we retain four dimensions, $S(4,8)$ and $S(5,8)$ remain virtually unchanged, but $S(4,5)$ changes from zero to 0.26. Thus only the value of $S(4,8)$ is insensitive to the number of dimensions retained. Moreover, it is also insensitive to whether the code 5 data is included.

TABLE 7.7

NUMBER AND COMPOSITION OF SAMPLE GROUPS; THREE OR FOUR DIMENSIONS RETAINED AFTER PRINCIPAL COMPONENTS ANALYSIS

Number of Dimensions Retained	Group Number	Number of Samples in Group		
		Code 4	Code 5	Code 8
3	1	0	16	30
	2	0	17	0
	3	24	0	0
4	1	0	16	29
	2	24	17	1

TABLE 7.8

SIMILARITY INDEXES OF DATA SETS; THREE OR FOUR DIMENSIONS RETAINED

Number of Dimensions Retained	Similarity Index		
	S(4,5)	S(4,8)	S(5,8)
3	0	0	0.50
4	0.26	0.017	0.28

Hence we draw the following conclusions concerning how our methodology should be applied:

- It is best to consider data sets in pairs rather than several at a time.
- If several sets of data are input together, two sets for which the similarity index is very small will have a very small similarity index when they are compared as a pair. This implies that under these circumstances comparison as a pair is unnecessary.

Taken together they imply that a reasonable approach to grouping of several data sets is to first input all of them together, then screen out pairs with very small similarity indexes, and finally treat all remaining pairs separately.

7.3.6 Comparison of Unnormalized Samples in the Frequency Domain for Object Codes 5 and 6

We now begin an examination which ultimately embraces code 5, 6, 7 data. These data are for the same satellite, taken from three different observation sites on the same day. Since the satellite is at very long range, it seems likely that the data should be quite similar.

The comparison discussed here was made on the basis of the 33 code 5 samples previously used, together with 30 code 6 samples, all with F;0,98,2 frequency selection.

Eight axes of the error ellipsoid were within a factor of one-fifth of the length of the major axis and 19 within one-tenth. Plots of sample projections in the XY and XZ planes are shown in Figs. 7.13 and 7.14. Visual groupings of the points would evidently tend to be controversial (the distinguishing labels on the plots are not available to aid classification).

The grouping algorithm gave the same results whether four or five dimensions were retained. It found two groups, with group 1 containing 18 code 5 samples and 24 code 6 samples, while group 2 contained 15 code 5 samples and 6 code 6 samplex. Accordingly, the similarity index S(5,6) is given by

$$S(5,6) = \frac{1}{2}\left[\frac{18/33}{24/30} + \frac{6/30}{15/33}\right] = 0.56$$

We conclude that code 5 and code 6 samples have substantially more in common than any of the other pairs previously considered. This is what we felt the result probably ought to be.

7.3.7 Comparison of Unnormalized Samples in the Frequency Domain for Object Codes 5 and 7

Twenty-seven code 7 samples with F;0,98,2 frequency selection were taken together with the 33 code 5 samples previously used. Much of the time-plot of code 7 data (from which these samples were derived) looks to be of poor quality. Therefore it was difficult to predict how the comparison would turn out. We would again expect that reasonably good quality data would show strong similarity to the code 5 data.

Ten axes of the error ellipsoid were within a factor of one-fifth as long as the major axis, and 19 within one-tenth. Projections of the samples on the XY- and XZ-planes are shown in Figs. 7.15 and 7.16. The data appears to separate in the XY-plane (at least with the benefit of the labels), but not in the XZ-plane. The separation apparent in Fig.

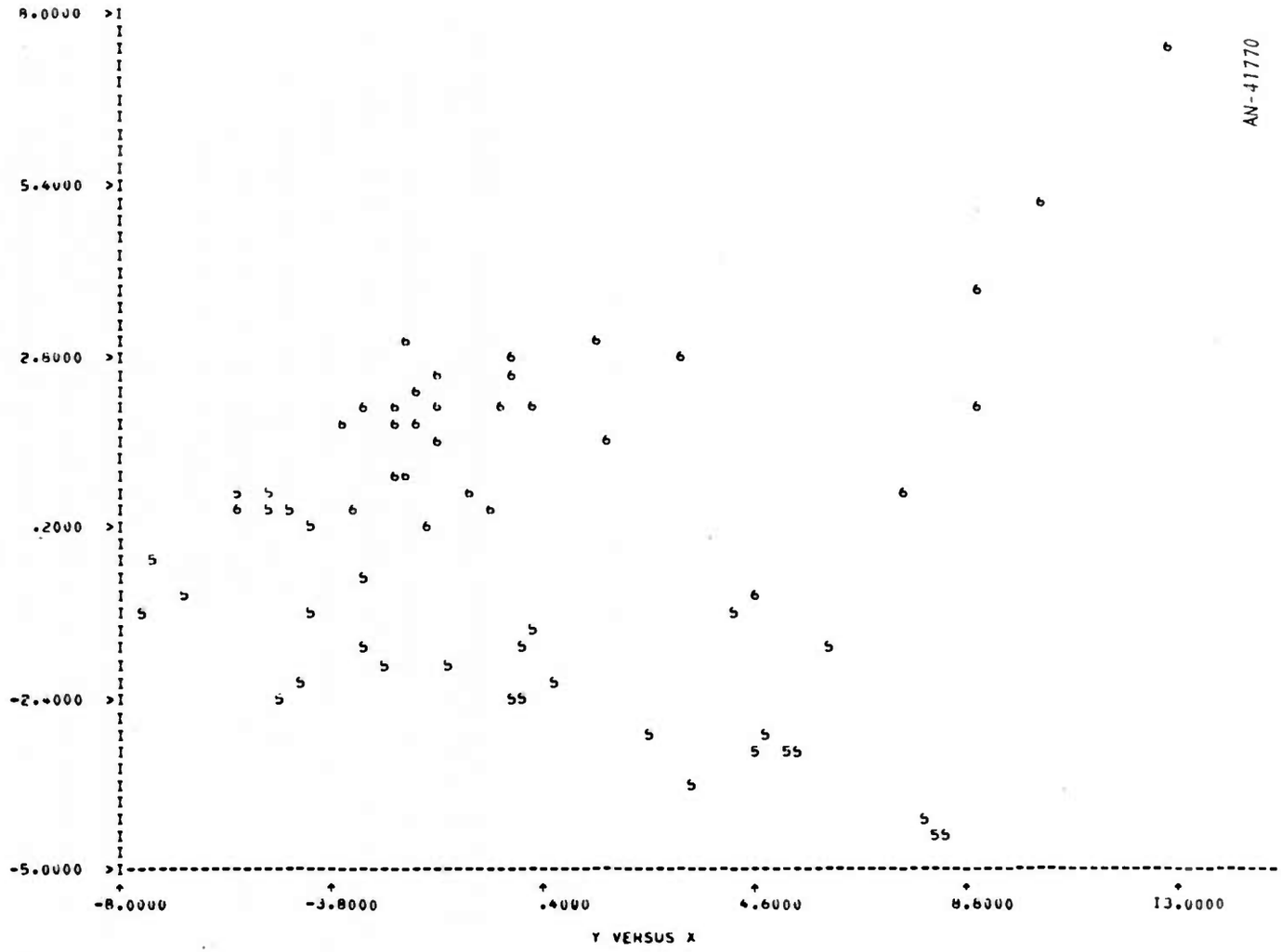


Figure 7.13. Projections of Preprocessed Samples in the XY-Plane: Code 5 and 6 Data (Frequency Domain)

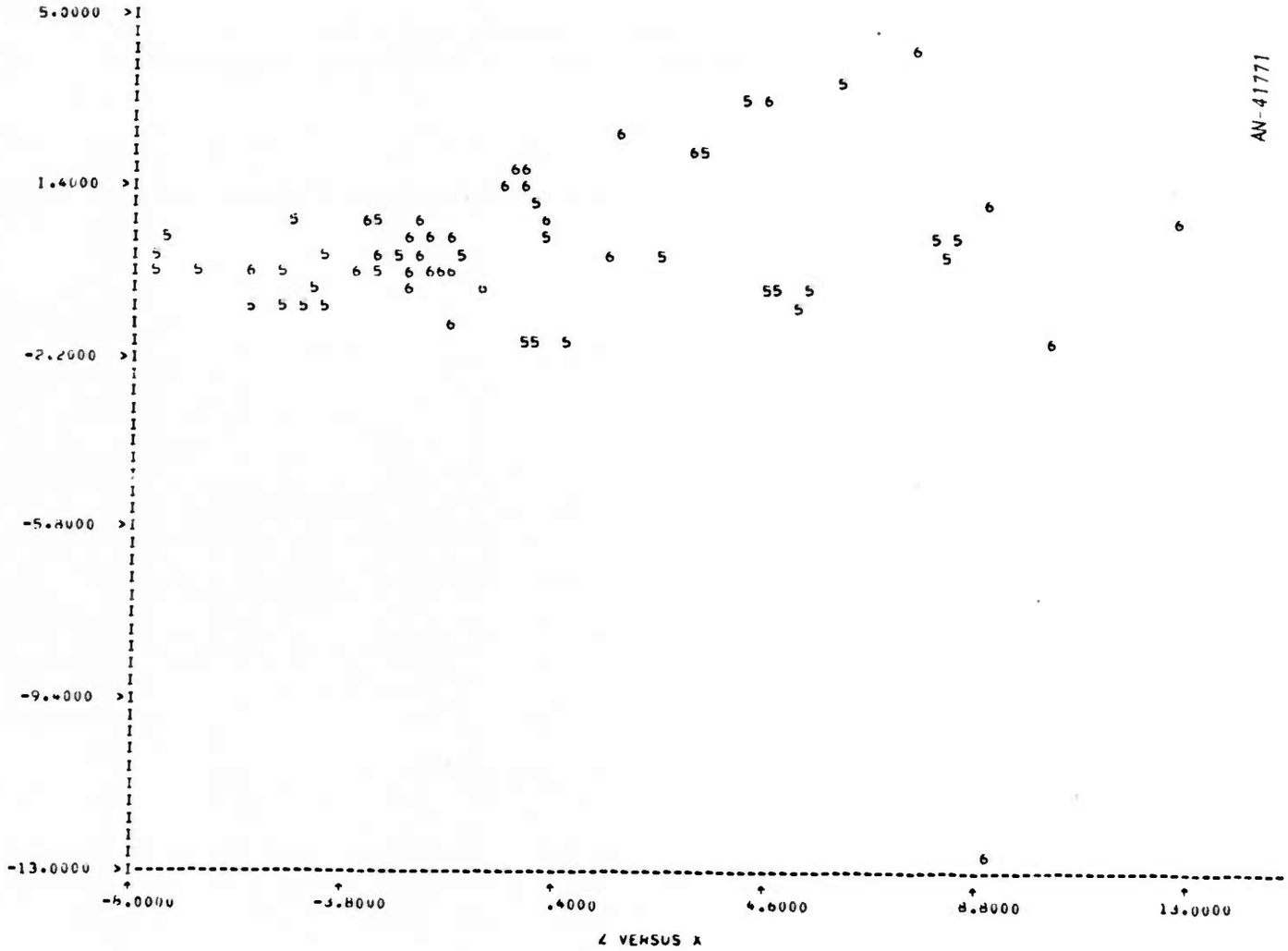


Figure 7.14. Projections of Preprocessed Samples in the XZ-Plane: Code 5 and 6 Data (Frequency Domain)

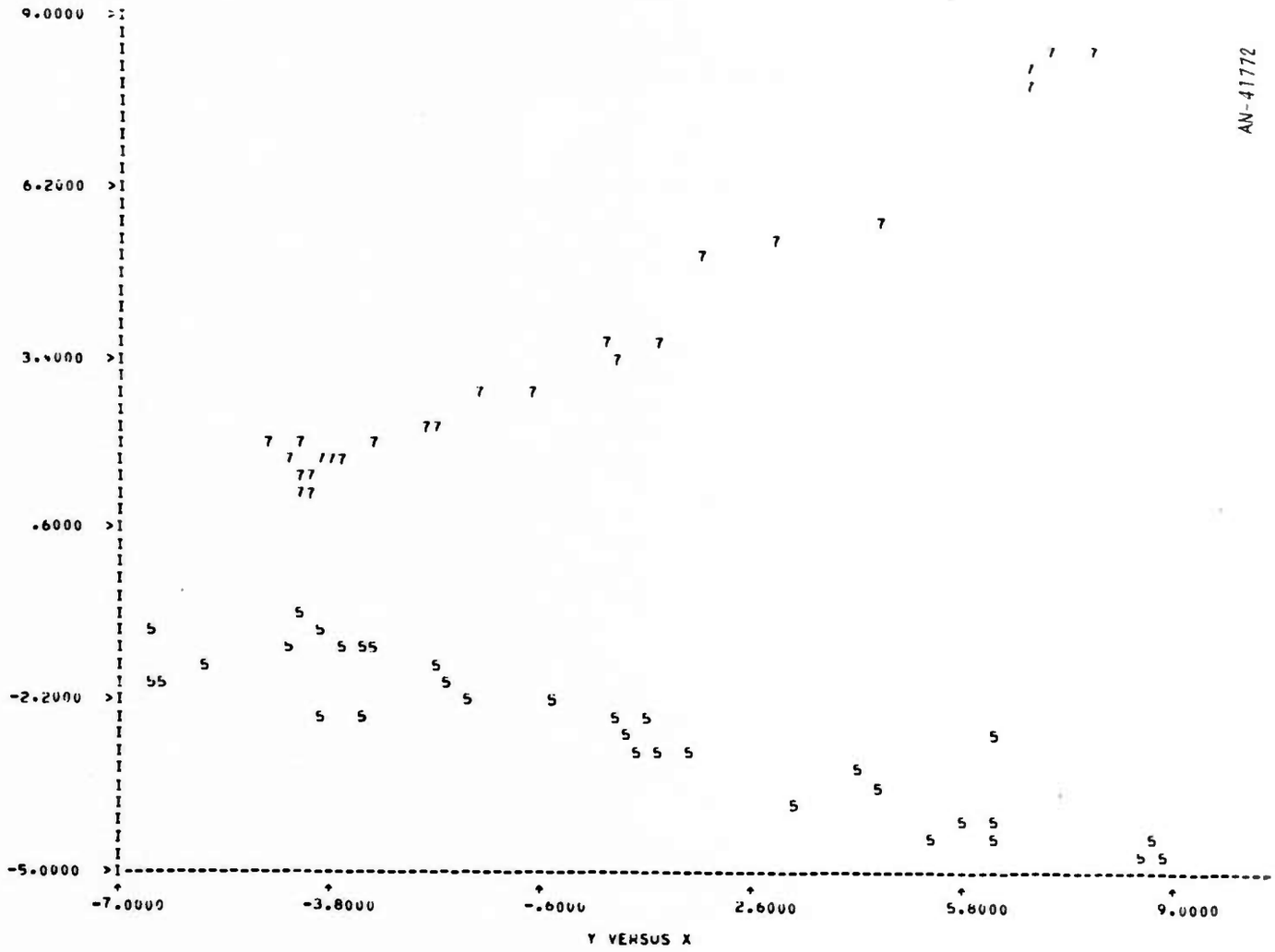


Figure 7.15. Projections of Preprocessed Samples in the XY-Plane: Code 5 and 7 Data (Frequency Domain)

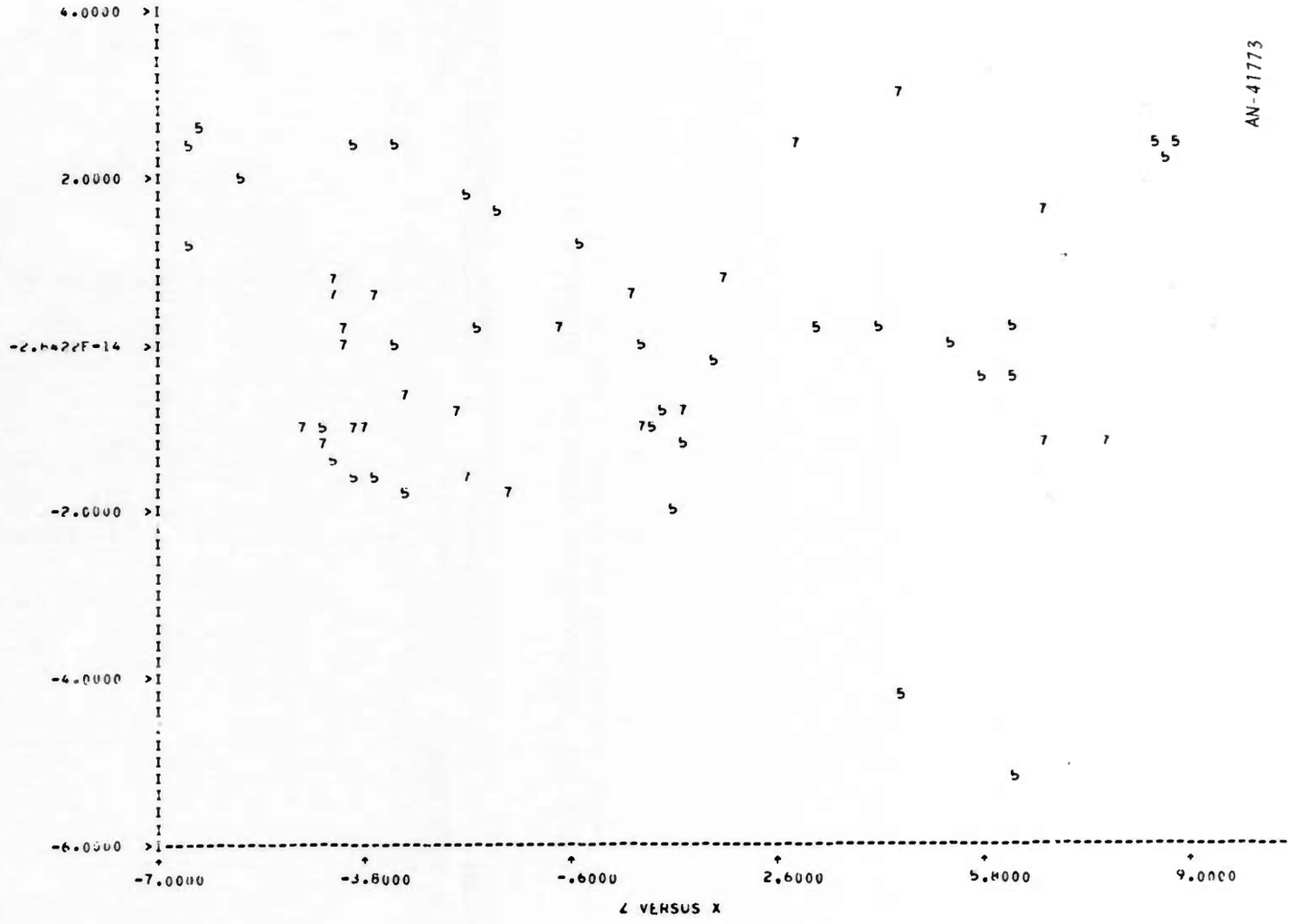


Figure 7.16. Projections of Preprocessed Samples in the XZ-Plane: Code 5 and 7 Data (Frequency Domain)

7.15 is so clear that it should not be necessary to go to a higher dimensional subspace or invoke LERNMOD in order to see that the objects seem to be dissimilar; a linear discriminant would probably work quite well.

However, for confirmation we did apply LERNMOD in two, three, and four dimensions. The groupings found are given in Table 7.9. The grouping in two and three dimensions is almost the same, but that in four dimensions is substantially different. This may be due to noise, as discussed in Sec. 7.3.2.

The value of the similarity index in three dimensions is

$$S_3(5,7) = 0.22$$

TABLE 7.9

NUMBER AND COMPOSITION OF SAMPLE GROUPS; TWO TO FOUR DIMENSIONS RETAINED AFTER PRINCIPAL COMPONENTS ANALYSIS

Number of Dimensions Retained	Group Number	Number of Samples in Group	
		Code 5	Code 7
2	1	17	18
	2	16	0
	3	0	7
3	1	16	18
	2	17	0
	3	0	7
4	1	10	25
	2	8	0
	3	12	0
	4	3	0

and in four dimensions it is

$$S_4(5,7) = 0.076$$

Thus the similarity is much less than between code 5 and code 6 data.

7.3.8 Comparison of Unnormalized Samples in the Frequency Domain for Object Codes 5, 6, and 7

Finally, to confirm that there are no surprises if all the samples for 4630 were put together, we input all the code 5, 6, and 7 samples simultaneously into our programs.

Figures 7.17 and 7.18 show sample projections in the XY- and XZ-planes. They appear about as anticipated. In the XY-plane, the 7's are separated from the 5's as before, while the 5's and 6's are strongly intermingled. In the XZ-plane, all are intermingled. Thus the code 5 and code 7 samples appear to have the strongest influence on the orientation of the error ellipsoid. Figure 7.17 again shows clear separation between code 7 and the other objects at least in the two-dimensional subspace; in this plane a linear discriminant would say that object 7 is definitely dissimilar from the other two. In three, four, or five dimensions, the same groupings were found, but these were different from those seen in the XY-plane. Group 1 contained 22 code 5 samples, 24 code 6 samples, and all 25 code 7 samples; group 2 contained the remaining 11 code 5 samples and 6 code 6 samples.

The corresponding similarity indexes are as follows:

$$S(5,6) = 0.72$$

$$S(5,7) = 0.33$$

$$S(6,7) = 0.40$$

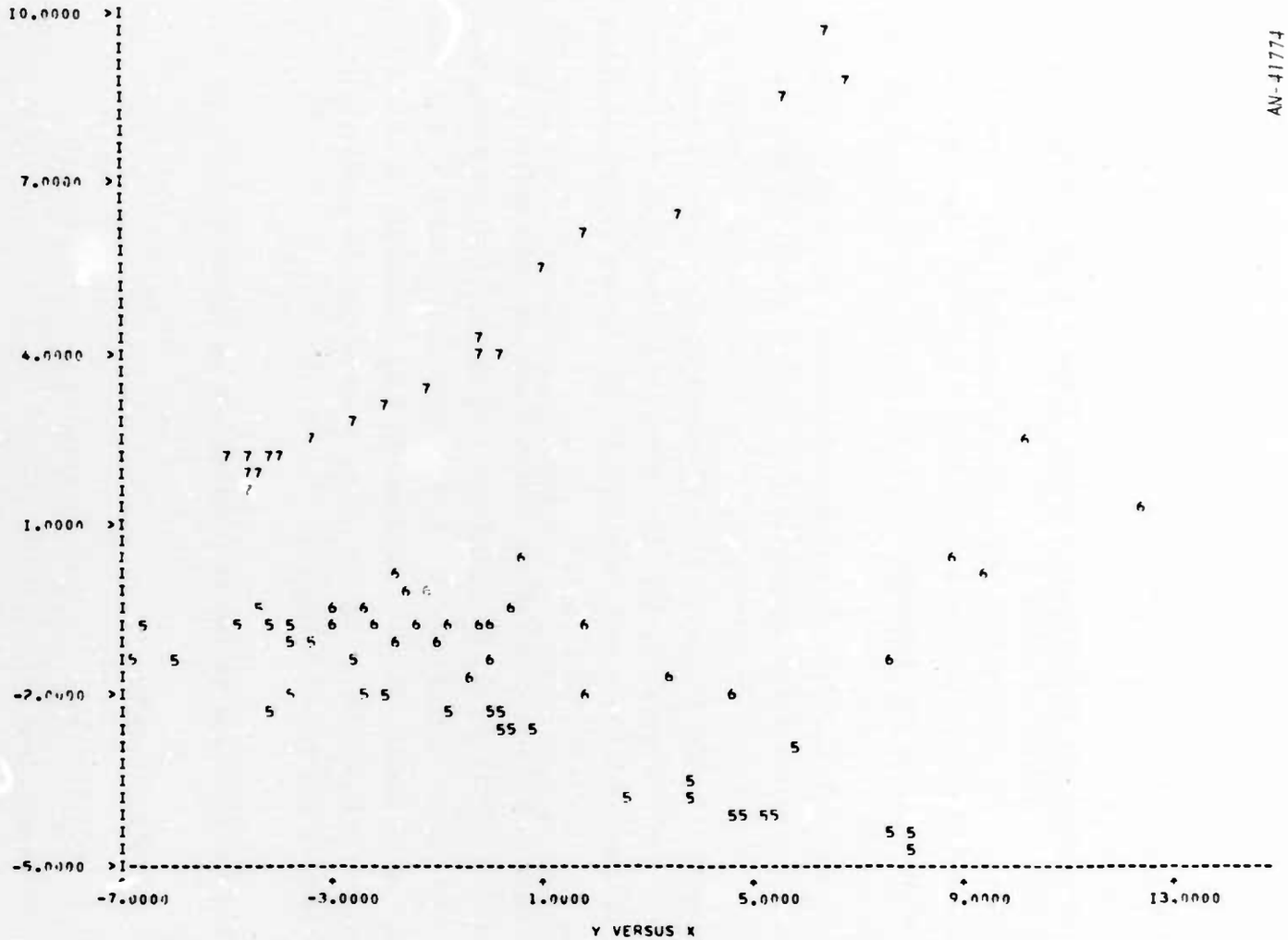


Figure 7.17. Projections of Preprocessed Samples in the XY-Plane: Code 5, 6, and 7 Data (Frequency Domain)

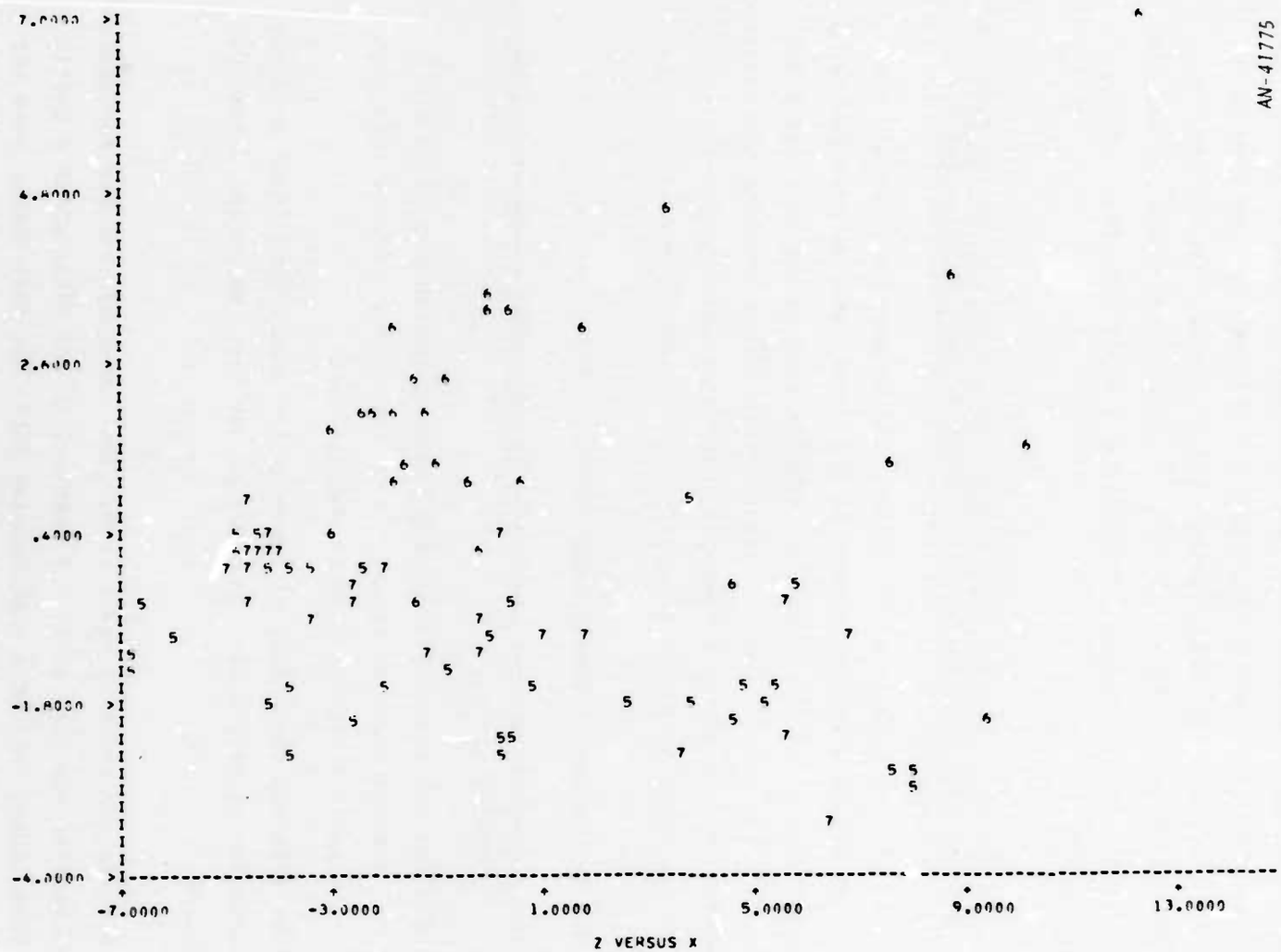


Figure 7.18. Projections of Preprocessed Samples in the XZ-Plane: Code 5, 6, and 7 Data (Frequency Domain)

The similarities S(5,6) and S(5,7) have increased somewhat over the values previously found when the data sets were examined in pairs. This is in accordance with what usually happens with our methods when more data sets are added for simultaneous classification. In the present case the effect may be due partly to the rather small number of samples of each type. But in general, it reflects a minor deficiency of the method itself.

7.3.9 The Effect of Two Forms of Sample "Normalization" in the Frequency Domain, and of Using Factor Analysis Instead of Principal Components Analysis

Primarily to obtain some results where we were more certain that calibration errors were not present in the input data, we used two forms of "normalization" of the F;0,98,2 samples used in the work described above. Then for one of these we examined the effect of using the covariance matrix instead of the correlation matrix of the samples--i.e., of using Factor Analysis rather than Principal Components Analysis.

The two forms of normalization were:

N1: Divide each sample component by the sample component corresponding to $f = 0$.

N2: For all the samples having the same code number, find the average value of their $f = 0$ components. Then divide each sample component by this average value.

The first form has the effect of removing the mean signal level entirely as a basis for classification, and may go too far. Hence the second was also stored.

Results are shown in Table 7.10. The comparable results for unnormalized samples are also given. A dash (-) in the table under a particular code number indicates that samples with that code number were not included. The table also shows the effect of changing from the correlation matrix to the covariance matrix when the second form of normalization (N2) is used.

TABLE 7.10

EFFECT ON SAMPLE GROUPINGS OF TWO FORMS OF NORMALIZATION AND USE OF THE COVARIANCE MATRIX

Case Number	Form of Normalization	Number of Dimensions Retained	Group Number	Number of Samples in Group				
				Code 4	Code 5	Code 6	Code 7	Code 8
1	None	4	1	1	22	21	25	-
			2	0	11	9	0	-
			3	23	0	0	0	-
2	N1	4	1	0	4	20	25	-
			2	0	29	6	0	-
			3	24	0	4	0	-
3	N2	4	1	0	0	0	25	-
			2	24	33	30	0	-
4	None	4	1	0	15	2	25	29
			2	24	18	28	0	1
5	N1		1	0	0	0	0	25
			2	21	14	23	25	0
			3	3	19	7	0	1
6	N2	4	1	0	4	1	0	29
			2	24	29	29	25	1
7	N2 with covariance matrix	4	1	0	9	5	25	30
			2	24	24	25	0	0

The first three cases shown in the table are for codes 4-7 samples. In all three cases the similarity index S(4,7) is at most very small. In the first two cases S(4,5) and S(4,6) are also very small, but this is not so in the third case. Also in case 3, S(5,7) and S(6,7) are all zero, while S(5,6) is unity.

The next three cases include the code 8 samples. In all these cases S(4,8) is either small or zero. S(4,7) is zero for the unnormalized samples, but is high for both types of normalization. S(5,6) is sizable in all cases.

The last two cases are for N2 normalization samples; case 6 processing used the sample correlation matrix, while case 7 used the covariance matrix. The covariance matrix reflects the significance of the actual magnitude of sample-to-sample variations in each component of the sample vectors and has a tendency to be dominated by swings at the frequencies where the average amplitude is high. On the other hand, the correlation matrix we have adopted in our standard procedure reflects the relative variance, and hence does not share this tendency; its potential problem is with noise at frequencies where the signal level is low. In both these cases, S(4,8) is very low. S(4,7) is high in case 6 and zero in case 7, whereas S(7,8) is just the opposite. S(5,6) is high in both cases. Thus the main effect of changing from the correlation to the covariance matrix before entering LERNMOD is to move the code 7 samples from one group to another.

Thus, data normalization can cause pronounced shifts in groupings. At present, the significance of this finding is not clear, since we have no yardstick to determine whether or not classification has been improved.

7.4 SENSITIVITY OF RESULTS TO THE INITIAL DIMENSIONALITY OF SAMPLES AND THE CHOICE OF SAMPLE FREQUENCIES

All the work described earlier in this section was carried out with 50-dimensional samples--the maximum dimensionality that our programs

can currently handle. The question arises as to whether a subset of these 50 dimensions contains the bulk of the information on which grouping is ultimately based.

Since we found that F;0,98,2 sampling led to better grouping than F;0,49,1 , we proceeded by progressively removing more and more of the higher frequency information from the F;0,98,2 samples.

The results obtained when only the first N dimensions of the F;0,98,2 samples were input to Principal Components Analysis (PCA) and LERNMOD are given in Table 7.11.

The table shows that grouping remains stable down to N = 45 , but changes drastically by N = 40 . This shows that it is essential to retain the higher frequency information.

Though the above result strongly suggests that a minimum of about 40 dimensions must be retained, for satellites (4,5) we tried the effect of using fewer dimensions, but taking each to be a harmonic of the satellite quarter-period. That is, we used the values of the first 16 spectral peaks as the sample data. In this case a subspace of two dimensions was retained after PCA. In this subspace LERNMOD found two groups: group 1 contained three code 4 samples and 26 code 5 samples, while group 2 contained 21 code 4 samples and seven code 5 samples. Thus S(4,5) = 0.20 , and some similarity is indicated. This contrasts with the results of Table 7.4 (Sec. 7.3.2), which gives S(4,5) = 0 . Thus the 16 spectral peaks do not appear to contain enough of the available information to serve as a good basis for classification. This suggests that intervening data points, though much smaller in magnitude than the peaks, contain much significant information. One might say that the diffuse, as well as the specular, scattering data are required in order to properly characterize an object.

TABLE 7.11

EFFECT OF INITIAL SAMPLE DIMENSIONALITY ON GROUPING

N	Number of Dimensions Retained after PCA	Group Number	Number of Samples in Group			
			Code 4	Code 5	Code 6	Code 7
50	4	1	1	22	21	25
		2	0	11	9	0
		3	23	0	0	0
49	4	1	1	22	23	25
		2	0	11	7	0
		3	23	0	0	0
45	4	1	0	22	23	25
		2	1	11	7	0
		3	23	0	0	0
40	4	1	1	17	21	0
		2	23	16	9	0
		3	0	0	0	25
30	4	1	0	9	4	14
		2	24	24	26	0
		3	0	0	0	11
20	4	1	0	10	6	25
		2	24	23	24	0

APPENDIX

COMMENTS ON SAMPLE SIZE AND DIMENSIONALITY

In classification problems where the underlying probability distributions are unknown, classification is frequently based on class boundaries determined by a subset of the available samples. As pointed out in papers by Foley* and others, the classification error rate for these samples can then be significantly lower than the rate for a much larger population drawn from the same distributions. Foley's results, for the Fisher linear discriminant, indicate that the ratio of the number of samples of each class to sample dimensionality should be at least 3 if this problem is to be avoided.

The extent to which LERNMOD is affected by the sample size-to-dimensionality ratio is not known. In previous experience, we have not encountered the low error rate phenomenon at a ratio as low as 2,** but we have no theoretical basis for asserting that this would always be true. Therefore, we have cautioned the reader that our results may be biased if Foley's criterion is not satisfied.

With regard to sample dimensionality we feel that it is always safe, but sometimes extremely conservative, to apply the criterion with the full number of sample dimensions. For example, if we have seven sample points per class in a space of two dimensions, Foley's criterion is satisfied. Suppose now that these points are perturbed slightly into a third dimension. Intuitively, we feel the sample size would still be adequate, though nominally the criterion calls for at least nine. Of

---

*D.H. Foley, "Considerations of Sample and Feature Size," IEEE Trans. on Information Theory, Vol. IT-18, No. 5, September 1972.

**Twenty samples in 10 dimensions.

course, the criterion is formulated to cover any displacements in the third dimension, not merely small ones. With this in mind, we have pointed out in the text how many principal axes of the error ellipsoid (of the correlation matrix of the pooled samples) are within a factor of one-tenth of the length of the major axis. We feel that three times this number is likely to be much closer to the actual minimum sample size requirement for each data set than the nominal requirement for 150 samples.

We should point out also that if a dimension which contains significant discrimination information is unwittingly discarded following Principal Components Analysis, our results will, in contrast to the above, be biased in the direction of a high classification error rate.

REFERENCES

1. M. Bair, et al., Optical Properties of Satellite Materials, Infrared and Optics Division, Environmental Research Institute of Michigan, Report No. 194100-6-F, July 1973 (rough draft).

2. J.A. Pilkington, "The Visual Appearance of Artificial Earth Satellites," Planet Space Science, Vol. 14, 1966, pp. 1281-1289; and Vol. 15, 1967, pp. 1535-1548.

3. C. Pitts, Photometric Data Package for General Research Corporation: Cloudcroft Data--Satellite 9450 (17 May 1972), KMS Technology Center, 26 July 1973.

4. C. Pitts, Photometric Data Package for General Research Corporation: Cloudcroft Data--Satellite 4630 (10 November 1972), KMS Technology Center, 10 October 1973.

5. C. Pitts, Photometric Data Package for General Research Corporation: AMOS Data--Satellite 4630 (10 November 1972), KMS Technology Center, 10 October 1973.

6. C. Pitts, Photometric Data Package for General Research Corporation: RML Data--Satellite 4630 (10 November 1972), KMS Technology Center, 11 October 1973.

7. C. Pitts, Photometric Data Package for General Research Corporation: Cloudcroft Data--Satellite 5587 (9 May 1972), Psi-Tran Corporation, 10 January 1974.

8. J. Dufay, Introduction to Astrophysics: The Stars, Dover Publications (New York), 1964.